Acta Informatica Pragensia
3(1), 2014, 44–56, ISSN 1805-4951
Online: aip.vse.cz

Sekce / Section:
Recenzované stati / Peer-reviewed papers

CADAQUES: Metodika pro komplexní řízení kvality dat a informací

CADAQUES: The Methodology for Complex Data and Information Management

David Pejčoch[1]

[1] Katedra informačního a znalostního inženýrství,
Fakulta informatiky a statistiky, Vysoká škola ekonomická v Praze
nám. W. Churchilla 4, 130 67 Praha 3

david.pejcoch@vse.cz

Abstrakt: Dnešní doba je charakteristická stále se zvětšujícím množstvím pořizovaných a zpracovávaných dat. Cílem tohoto článku je poukázat na různorodost současně používaných datových zdrojů, ukázat jejich specifika z pohledu řízení kvality a představit vlastní metodiku, která umožňuje řízení kvality dat a informací napříč těmito zdroji. Hlavní součástí této metodiky je sada základních principů a činností, které je možné univerzálně aplikovat. Jedním z klíčových doporučení této metodiky je zaměření se na relativně malou sadu vlastností dat, kterou lze efektivně řídit. Součástí metodiky je rovněž model zralosti datového zdroje, který slouží pro zhodnocení míry rizika spojené s použitím konkrétního zdroje.

Klíčová slova: Datová kvalita, Big Data, zdroje dat, Data Governance, Linked Data, CADAQUES.

Abstract: The present time is characterized by an ever-increasing amount of acquired and processed data. The aim of this article is to highlight the diversity of currently used data sources, to show their specifics in terms of quality control and introduce own methodology that allows data and information quality management across these sources. The main component of this methodology is a set of basic principles and actions that can be universally applied. One of the key recommendations of this methodology is to focus on a relatively small set of data characteristics, which is relatively easy to manage. Part of the methodology is also a Data Source Maturity Model which could be used to assess the risk associated with the use of a particular data source.

Keywords: Data Quality, Big Data, Data sources, Data Governance, Linked Data, CADAQUES.

1 Úvod

Žijeme ve světě produkujícím enormní, stále se zvětšující množství dat. V současné situaci je diskutabilní, zda jsme vůbec toto množství dat schopni zpracovat a zda toto zpracování přinese adekvátní dodatečnou hodnotu. Zdomácněly již pojmy jako je řízení kvality dat a informací či Data Governance. S rozrůstajícím se univerzem dat se objevily též různé klony těchto pojmů jako je např. Linked Data Quality (řešící kvalitu původních dat převedených do Resource Description Framework formátu a vzájemně prolinkovaných), Metadata Quality (reflektující specifika řízení kvality metadat, tedy velmi zjednodušeně dat o datech) a konečně Big Data Quality, soustředící svou pozornost na data, která již není možné (nebo není efektivní) zpracovávat pomocí konvenčních RDBMS (Relation DataBase Management System) prostředků.
Nepouštějme se nyní do polemik, do jaké míry jsou tyto pojmy tzv. buzzword. Produkce buzzwordů je fenoménem, se kterým v dnešní době stěží něco svedeme. V každém případě bychom zde mohli identifikovat celou řadu dalších takových pojmů, relevantních pro jiné specifické formy dat, o nichž tu zatím nepadlo ani slovo (např. Multimedia Data Quality, Stream Data Quality, Geospatial Data Quality, apod.). Odhlédněme též od nedokonalosti definice velkých dat ve smyslu takových dat, která činí potíže zpracovat pomocí „konvenčních" prostředků. Pojem „konvenční" prostředek je v tomto kontextu velice relativní, neboť v online byznysu začínají být konvenčními již i takové technologické prostředky, na kterých současná Big Data fungují. Podstatným problémem je snaha roztříštit celou oblast řízení kvality dat a informací, potažmo Data Governance, do zvláštních disciplín. Dle mého názoru je nutné uvažovat jednu komplexní sadu pravidel, politik a principů univerzální pro všechny datové zdroje, se kterými příslušný subjekt přichází do styku. Tyto principy by měly být jednoduché, srozumitelné a snadno implementovatelné.
Cílem článku je poukázat na různorodost současně používaných datových zdrojů a představit konkrétní metodiku, která umožňuje řízení kvality dat a informací napříč těmito zdroji. Tato nová metodika je rovněž autorovým přínosem k dané problematice.

2 Směrem k jednotnému konceptu

2.1 Zdroje dat

Před vlastním popisem jednotlivých zdrojů dat považuji za vhodné definovat pojem univerzum dat. Teorie množin definuje univerzum (též základní množinu) jako množinu všech prvků, které jsou relevantní v rámci daného kontextu (domény, problému). V kontextu dat chápu univerzum jako množinu všech datových zdrojů (na různé úrovni granularity), se kterými konkrétní subjekt přichází do styku. Na univerzum dat, která jsou v dnešní době zpracovávaná, lze pohlížet z několika úhlů pohledu. Podle míry jejich strukturovanosti se jedná o data strukturovaná, semi-strukturovaná a nestrukturovaná. Z pohledu jejich původu je lze dělit na data interní a externí. Mezi interní data lze zařadit např. data z OLTP (Online Transaction Processing) systémů, datových skladů, datových tržišť, OLAP (Online Analytical Processing) kostek, různé systémové číselníky a seznamy hodnot, systémové logy, archivovanou elektronickou poštu, znalostní báze, různé registry metadat a v neposlední řadě data produkovaná různými senzory. Jako příklad externích dat lze uvést data získaná ze sociálních sítí, cloud řešení (např. Salesforce), data vyměňovaná a sdílená v rámci B2B (Business To Business) vztahů, veřejně dostupné registry a číselníky.
Z pohledu konkrétní použité technologie a formátu se jedná o relační databáze, celou řadu tzv. NoSQL databází (sloupcové, grafové či dokumentové databáze), různé klony XML

(eXtensible Markup Language), textové soubory s oddělovačem či pevnou šířkou, audio, video soubory, obrázky s různou mírou komprese a barevné hloubky, apod. Specifickou formou jsou data mající podobu toku (streamu). Z pohledu fyzického uložení dat lze rozlišovat data uložená v rámci souborového systému, dedikovaného úložiště či v paměti. Z pohledu dostupnosti lze rozlišovat mezi online daty (např. Data as a Service) či offline daty. Na základě popsaných faktů lze současné univerzum zpracovávaných dat označit bez nadsázky za velmi rozmanité. Nyní přikročím k definici základních pojmů, které souvisejí s řízením kvality takto definovaného univerza dat.

2.2 Kvalita dat a informací, Data Governance

Podle (SDM, 2005) je datová kvalita „*vnímání nebo posouzení vhodnosti dat sloužit svému účelu v daném kontextu*“. Za aspekty datové kvality tento zdroj považuje správnost, úplnost, aktuálnost, relevantnost, konzistentnost napříč datovými zdroji, důvěryhodnost, přiměřenou prezentaci a dostupnost. Podobnou definici poskytuje (Redman, 2001), který za data mající vysokou kvalitu považuje taková, jež „*odpovídají jejich zamýšlenému užití v operativních činnostech, rozhodování a plánování*“. Přičemž data označuje za odpovídající svému užití, pokud jsou „*prosta defektů a uchovávají si požadované vlastnosti*“. Tato definice vychází z přístupu M. Jurana, publikovaného v (Juran, Godfrey, 2010). S adekvátnosti dat jejich zamýšlenému užití se můžeme v souvislosti s datovou kvalitou setkat též v (Strong, Lee, Wang, 1997; Wang, Strong, Guarascio, 1996; Olson, 2003). Dle spíše procesně orientované definice, publikované v (Hyland, Elliott, 2008), “*datová kvalita nastavuje množinu opakovatelných procesů pro monitoring dat a zlepšování jejich přesnosti, úplnosti, aktuálnosti a relevantnosti*.“ Správná, aktuální, relevantní, úplná, srozumitelná a důvěryhodná jsou kvalitní data rovněž podle (Olson, 2003).

Jak je patrné, datová kvalita je často definována jako míra určitých požadovaných vlastností. Téma vlastností dat samotných je zpracováno řadou autorů. Král a Žemlička (2006) rozlišují vlastnosti objektivní a subjektivní z pohledu jejich měřitelnosti. Za objektivní považují takové (zpravidla numerické) vlastnosti, které lze vždy znovu vypočítat z dat, kterých se týkají. Nejobsáhlejší přehled alternativních klasifikací (přesto však ne úplný) poskytuje (Zaveri et al., 2012). Ve své práci se odkazuje na celkem 21 zdrojů různých autorů. Výsledkem je klasifikace celkem 109 vlastností do 6 dimenzí. Řízení takto enormního množství vlastností dat považuji za téměř nemožné. Jako podpůrný argument pro toto tvrzení mohu použít analogii z oblasti řízení rizik, kde je obecně považována hranice několika desítek rizik za únosnou míru, kterou je firma ještě schopna efektivně uřídit – viz např. (Doucek et al., 2011).

Na základě (Batini, Scannapieco, 2006; Král, Žemlička, 2006; Lee et al., 2006; McGilvray, 2008; Pipino, Lee, Wang, 2002; Redman, 2001; Voříšek et al., 2008; Zaveri et al., 2012) a vlastních zkušeností z oblasti pojišťovnictví, bankovnictví a online byznysu jsem proto sestavil redukovanou sadu vlastností, které považují za klíčové. Tato klasifikace uvažuje pět dimenzí vlastností: (1) endogenní, (2) časovou, (3) kontextuální, (4) užití a (5) ekonomickou. Zatímco s prvními čtyřmi dimenzemi se můžeme setkat napříč citovanými zdroji, pátá je zmiňována pouze v (Voříšek et al., 2008). Nicméně bez měření nákladové stránky vlastnictví dat si lze řízení dat jen těžko představit. Její absenci lze proto přičíst spíše neúplnosti ostatních přístupů nebo jejich přílišnému zaměření na konkrétní technická řešení kalkulace vlastností.

V rámci endogenní dimenze uvažuji Důvěryhodnost (míru všeobecné akceptovatelnosti dat jejich uživateli), Unikátnost (míru výskytu nechtěných duplicit), Syntaktickou správnost (míru, v níž hodnoty atributu odpovídají přípustné syntaxi), Sémantickou správnost (míru, v níž hodnoty atributu odpovídají oboru přípustných hodnot) a Přesnost (míru, v níž data přesně popisují konkrétní entitu).

V rámci časové dimenze uvažuji především Aktuálnost dat a jejich Volatilitu (proměnlivost v reálném prostředí). Jak jsem ukázal v (Pejčoch, 2011), ostatní vlastnosti této dimenze jsou buď z těchto vlastností odvozené (Včasnost), anebo představují naopak jejich vstup (Časová synchronizace).
V rámci kontextuální dimenze rozlišuji Interní a Externí konzistentnost, ve smyslu konzistentnosti hodnot atributů v rámci jednoho datového zdroje, vs. konzistentnosti napříč datovými zdroji. Dále míru Úplnosti (podíl nechtěných chybějících hodnot) a Pokrytí všech hodnot, které se pro danou entitu vyskytují v reálném světě (např. všech čísel mobilního telefonu daného subjektu).
V rámci dimenze užití uvažuji Dostupnost dat uživatelům, Srozumitelnost (ve smyslu formátu), Interoperabilitu (míru existence metadat) a Bezpečnost přístupu. V rámci ekonomické dimenze uvažuji náklady na pořízení, aktualizaci, uložení, sdílení, archivaci a ochranu dat.
Zatímco (English, 1999) hovoří striktně o kvalitě informací, jíní autoři, jako např. (Redman, 2001), zmiňují kvalitu dat. Je skutečně nutné striktně rozlišovat tyto dva pojmy? Data lze definovat jako základní diskrétní stavební prvky informací, reprezentované podle různé míry strukturovanosti fakty uloženými v datových atributech, textových dokumentech, zvukových záznamech či formou obrazu/videozáznamů. Informacemi bychom potom rozuměli data zasazená do kontextu a zřetězená ve vyšší smysluplné celky, mající určitý význam. Dává smysl uvažovat kvalitu dat, když pro koncového uživatele má význam kvalita informace, hovoříme o životním cyklu informace a název celého oboru je Informatika? Dle mého názoru je vhodné uvažovat oba pojmy současně. Mnoho vlastností uváděných jako vlastnost dat (např. srozumitelnost) jsou spíše vlastnosti informace, tedy dat vztažených v kontextu a prezentovaných např. formou reportu. Nicméně např. správnost informace do značné míry závisí na správnosti elementárních datových prvků, z nichž je složena. Např. informace, že pan Novák se narodil 12. října 1977, může být správná pouze za předpokladu, že je správně vyplněno příjmení příslušné osoby a datum narození.
Ladley (2012) chápe Data Governance jako „*organizaci a implementaci politik, procedur, struktur, rolí a odpovědností, které vymezují a prosazují pravidla účasti, rozhodovací práva a odpovědnosti pro efektivní řízení informačních aktiv*“. Zdůrazňuje oddělení Governance od řízení informací ve smyslu kontroly, monitoringu a dohlížení na aplikaci standardů a pravidel definovaných v rámci Governance. Výsledkem Data Governance iniciativy jsou podle Ladley (2012) principy (ve smyslu základních norem, doktrín a předpokladů) a politiky (ve smyslu pravidel nebo kodexu jednání). Konkrétní principy mimo jiné zmiňují (1) komplexní řízení veškerých dat a informačního obsahu organizace jako korporátních aktiv, (2) nutnost specifikace standardů pro všechny datové struktury / informační obsah a (3) řízení rizik, nabádající k nutné obezřetnosti (due dilligence).
Co je podstatné, Ladley (2012) uvažuje při zavádění principů Data Governance postupnou evoluci namísto masivní změny. Dle mého názoru podloženého dosavadní praxí jsou právě implementace založené na masivní změně jedním z důvodů, proč je datová kvalita v současné době chápána převážně jako dodatečně vynaložené náklady, nikoliv z pohledu budoucí hodnoty. Řady neúspěšných nebo dlouho trvajících projektů daly vzniknout mýtu, že Data Governance je pouze pro velké firmy, protože ty malé na to nemají potřebné prostředky. Tato představa je však mylná.
V souvislosti s Data Governance je často zmiňován tzv. Data Stewardship, založený na více či méně formálně vymezené roli správce dat z příslušné předmětné oblasti (domény) nebo konkrétního datového zdroje.

2.3 Specifika jednotlivých typů dat z pohledu řízení kvality

Pro řízení kvality dat v rámci takových zdrojů, které mají charakter strukturované tabulky, textového souboru s oddělovačem nebo pevnou šířkou jednotlivých sloupců, existuje dostatečná opora jak v odborné literatuře, tak i ve funkcionalitě nástrojů dostupných na trhu. Řízení kvality těchto zdrojů je založeno na řízení sady jejich vlastností, uvedených v části 2.2. Jak je tomu však u méně obvyklých typů dat? Co je činí natolik výjimečnými, že si v některých případech zasloužily své vlastní vývojové větve v genezi datové kvality?

Metadata mohou být uložena buď separátně od vlastního datového zdroje (např. u Linked Data s použitím Protocol of Web Description Resources nebo Gleaning Resource Description for Dialects of Language), anebo přímo obsažena ve zdroji, který popisují. Příkladem druhého uvedeného případu mohou být obrázky či video, pro které jsou podle Metadata Working Group (2010) typické standardy XMP (Extensible Metadata Platform), EXIF (Exchengable Image File Format) a IPTC Information Interchange Model, jež ukládají metadata přímo jako součást souborů. Bez ohledu na formu uložení lze v případě metadat zkoumat jejich úplnost, správnost, konzistentnost, srozumitelnost a stáří, stejně jako je tomu u „klasických" dat uložených v tabulkách relačních databází či souborech o různé míře strukturovanosti. I v tomto případě nás bude zajímat původ těchto dat (a z ní vyplývající důvěryhodnost). Specifika jednotlivých typů zdrojů lze tudíž z tohoto podhledu uvažovat pouze na bázi aplikace konkrétního technologického postupu pro extrakci metadat a jejich analýzu.

V případě sekundárních (tedy nepůvodních), resp. terciárních dat (chápejme jako data odvozená ze sekundárních dat), se můžeme setkat s dalšími specifickými rysy. Kvalita těchto zdrojů je zatížená kvalitou zdrojů původních a kvalitou transformací, které jsou na původní data aplikovány. Příkladem takových dat jsou již zmiňovaná Linked Data a Big Data. V případě Linked Data je nutné čelit problémům spojeným s nedostatečnou úrovní metadat původních zdrojů (včetně těch popisujících aktuální úroveň kvality těchto zdrojů), chybami vzniklými převodem původních datových zdrojů do RDF (Resource Description Framework) formátu a automatickým generováním vazeb mezi RDF trojicemi. V případě Big Data se navíc setkáváme s problémy spojenými s integrací zdrojů, které mají často různou míru strukturovanosti.

MediaBistro (2010) uvádí v souvislosti s Linked Data některé vlastnosti, jež nejsou zpravidla zmiňovány v souvislosti s „klasickými" daty. Např. ty, které jsou spojené se syntaxí zápisu konkrétního formátu a mírou údržby. Syntaxi zápisu lze dle mého názoru chápat jednak ve smyslu již definované syntaktické správnosti a jednak ve smyslu srozumitelnosti formátu. Míru údržby chápu jako kombinaci aktuálnosti zdroje a jeho interoperability (tj. míry dostupné dokumentace / metadat). Neshledávám proto důvod tuto vlastnost explicitně vyčleňovat. V případě takových vlastností dat, které jsou zaměřené ryze na konkrétní technologickou realizaci (např. granularita modelování), je diskutabilní, zda se jedná o skutečné vlastnosti dat, anebo faktory mající na tyto vlastnosti vliv (tedy příčiny nekvalitních dat).

Obsah pojmu Big Data Quality / Governance je do určité míry zastřen tajemstvím. Soares (2012) chápe Big Data Governance jako součást širšího konceptu globálního Governance informací. Pozornost věnuje jak proaktivnímu řešení datové kvality na úrovni zdrojů velkých dat, tak i prostředkům pro její ex-post řešení na úrovni samotného Hadoopu. Výklad pojmu Big Data Quality však může být i širší. Informace extrahované pomocí Hadoop mohou v praxi sloužit pro validaci (např. odhadované pohlaví klienta v případě online byznysu) nebo obohacení ostatních datových zdrojů (např. o identifikované vztahy na základě analýzy nestrukturovaných dat ze sociálních sítí). Případně je možné využít výpočetní síly samotného

Hadoopu pro realizaci některých časově náročných úloh typických pro řízení datové kvality, jako je deduplikace (viz např. existující řešení Dedoop).
V každém případě se u velkých dat opět setkáváme s podobnou sadou vlastností jako u dat „klasických“ (správnost, konzistentnost, aktuálnost, úplnost, …). Obdobně jako u Linked Data je možné identifikovat některé vlastnosti specifické pro dané technologické řešení (např. odolnost architektury řešení vůči výpadkům jednotlivých komponent clusteru, na jehož aplikaci jsou velká data založena). Zde je však snad ještě mnohem více patrné, že se spíše jedná o příčiny nekvalitních dat, než o samotné vlastnosti dat.
Obecně tedy lze říci, že specifika typů dat, které si v minulosti vysloužily svou vlastní vývojovou větev v rámci řízení kvality, spočívají spíše v konkrétní použité technologii a okolnostech, zda se jedná o data původní či odvozená. Jak již naznačuje Soares, (2012), je vhodné uvažovat globální úroveň Governance. Já tuto globální úroveň chápu jako řízení a správu všech datových zdrojů podle jednotné sady politik a pravidel napříč univerzem.

3 Metodika CADAQUES pro komplexní řízení kvality dat a informací

Příkladem návodu pro komplexní řízení kvality dat a informací je vlastní navržená metodika CADAQUES (Complex Approach to Data and Information Quality Management within Enterprise Systems). Skládá se s následujících stavebních prvků: (1) základních principů, (2) redukované sady vlastností dat, kterou je vhodné řídit napříč všemi typy datových zdrojů, (3) modelu hodnocení zralosti datového zdroje, (4) základních činností, realizovaných v rámci různých úrovní řízení kvality dat a informací napříč všemi typy datových zdrojů, (5) simulačního modelu pro měření dopadu vlastností dat do metrik výkonnosti IT / podniku a (6) šablon dokumentů. Většina výstupů je pro registrované uživatele dostupná na autorsky vyvíjeném portálu http://www.dataquality.cz.

3.1 Základní principy metodiky CADAQUES

Metodika se zaměřuje na audit a řízení kvality komplexního univerza dat. Reflektuje heterogenní charakter současných datových a informačních zdrojů. Doporučuje jednotný koncept řízení kvality napříč všemi datovými zdroji. Vychází přitom z přesvědčení, že jen tak lze efektivně vyvážit přínosy a rizika spojená s užitím dat. V případě sekundárních a terciárních datových zdrojů preferuje řízení kvality na úrovni původních zdrojů. Považuje za efektivnější vyřešení prapůvodní příčiny nekvalitních dat před opakovanou retrospektivní nápravou již vzniklých defektů na úrovni zpracování dat.
Metodika klade důraz na prolinkování zaváděných principů Data Governance s existujícími postupy pro implementaci principů IT Governance, jmenovitě COBIT (Control Objectives for Information and Related Technology). Data Governance chápe pouze jako součást IT Governance. V této souvislosti doporučuje rovněž realizovat audit kvality dat v kontextu komplexního auditu informačního systému, s využitím postupů, které jsou v souladu s metodikou použitou pro implementaci IT Governance. Cílem je zamezit zmatkům plynoucím z koexistence více různých (často složitých) metodik. Konkrétní postup, jakým realizovat audit kvality dat v souladu s návodem IT Assurance Guide: Using COBIT jsem teoreticky popsal a kriticky zhodnotil v (Pejčoch, 2012) a jeho praktickou realizaci demonstroval na příkladu z oblasti pojišťovnictví v [22].
Metodika reflektuje skutečnost, že některé datové zdroje nejsou pod přímou kontrolou jejich konzumentů. Doporučuje udržovat informaci o míře kontroly přístupnou uživatelům dat a zohlednit ji při definici očekávané míry vlastností příslušných atributů. Doporučuje též

zhodnotit míru rizika spojenou s použitím takových datových zdrojů. Datové zdroje, s jejichž použitím je spojena příliš vysoká míra rizika, doporučuje nepoužívat. Jako podpůrný prostředek pro posouzení míry rizika metodika poskytuje model zralosti datového zdroje.
Metodika doporučuje dodržování principu Data Lineage (rodokmenu dat), tedy sledování původu dat, způsobu pořízení, stáří dat, frekvenci aktualizace a historii všech provedených transformací. Dodržování tohoto principu usnadní dohledávání původních příčin vzniku defektů v datech.
S ohledem na stále se zvyšující množství zpracovávaných dat metodika doporučuje důsledně řídit životní cyklus dat/informací. Pokud jsou nějaká data nepotřebná či duplicitní, je třeba je archivovat nebo smazat. Důvodem jsou náklady na správu, uložení, dostupnost a chaos. V tomto bodě se opírám o vlastní zkušenosti z oblasti online byznysu, v rámci něhož jsou běžným jevem situace, kdy i samotná strukturovaná data uložená v datových skladech nabývají velikosti desítek PB (petabajt, 10^{15} bajtů). Takové objemy dat přerůstají možnosti běžných RDBMS a zajištění dostupnosti dat je zejména u velkých subjektů spojeno se značnými náklady.
Metodika klade důraz na využití dodatečných znalostí při řízení kvality dat a informací. Tyto znalosti doporučuje spravovat ve znalostní bázi. Konkrétní znalosti spojené s řízením kvality dat mohou mít např. podobu schémat použitých pro standardizaci hodnot, masek pro validaci syntaktické správnosti či validačních byznys pravidel. Metodika doporučuje jejich centrální uložení a správu. Cílem je efektivní opakované využití již existujících znalostí, jednotně řízená kontinuita zvyšování kvality těchto znalostí a jejich soulad s požadavky regulací příslušeného odvětví (např. Solvency II/III v oblasti pojišťovnictví nebo Basel II/III v oblasti bankovnictví).
Metodika doporučuje použití kanonického (obecného) datového modelu při budování znalostní báze orientované na řízení kvality dat a informací (QKB). Pojem Kanonický datový model (též Společný datový model, CDM) pochází z oblasti datové integrace, kde jej např. (Štumpf, Džmuráň, 2008) zmiňuje jako model nezávislý na konkrétní aplikaci. Howard (2008) hovoří o „*datovém modelu, který překlenuje podnikové aplikace a různé datové zdroje*“. Chappell (2004) uvažuje Kanonický datový model za soubor kanonických XML formátů jako prostředek pro vyjádření dat putujících skrze podnik napříč architekturou tvořící ESB (Enterprise Service Bus). Existuje řada dostupných kanonických modelů pro různé vertikály. Jako příklad lze zmínit ACORD (Association for Cooperative Operations Research and Development) pro oblast pojišťovnictví či SID pro oblast telekomunikací. Za hlavní argument použití kanonického datového modelu jako základu znalostní báze považuji usnadnění její integrace do stávajícího ESB, založeného na kanonickém modelu, a snadné sdílení (a rozvoj) znalostí obsažených v QKB ve vztazích B2B (Business To Business) a B2G (Business To Governance).
Metodika doporučuje udržování takové úrovně vlastností dat a informací, která vede k optimálnímu dopadu do metrik výkonnosti IT a podniku při současném efektivním využití podnikových zdrojů. Pokud reálné použití některých atributů nevyžaduje vysokou úroveň některých měřených vlastností nebo je tato úroveň spojena s příliš vysokými náklady, metodika doporučuje slevit z požadavků na tyto vlastnosti. Pro měření dopadu do metrik výkonnosti používá kauzální simulační model, který jsem navrhl a popsal v (Pejčoch, 2011). Tento model je založen na vazbách mezi jednotlivými atributy, jejich užitím, naměřenými vlastnostmi a důsledky úrovně těchto vlastností. Součástí publikovaného článku byl rovněž praktický příklad demonstrující použití tohoto modelu v oblasti pojišťovnictví. Tento model je univerzálně aplikovatelný napříč univerzem dat.

Metodika klade důraz na multidimenzionální přístup, charakterizovaný různými pohledy na vlastnosti dat a informací (stávající / potenciální / optimální úroveň), uvažováním současného i potenciálního užití dat a informací a uvažováním různých kategorií dopadu nekvalitních dat a informací.

3.2 Základní činnosti podle metodiky CADAQUES

Metodika CADAQUES předpokládá, že základní činnosti realizované v rámci řízení datové kvality a Data Governance lze členit podle tří úrovní řízení (na strategickou, taktickou a operativní). V rámci strategické úrovně řízení probíhá definice základních principů a politik Data Governance. Na taktické úrovni probíhá přiřazování odpovědnosti za jednotlivé datové zdroje a rozšiřování stávajících rolí o tyto odpovědnosti. Dále sem patří návrh a rozvoj jednotlivých částí QKB a navazujících znalostí, jednorázové audity kvality dat a informací. V neposlední řadě je na taktické úrovni realizována sada typických činností, jako je standardizace, unifikace, porovnávání a slučování, doplňování chybějících pozorování a návrh validačních pravidel a kontrol.

Tyto typické činnosti jsem identifikoval na základě komparativní analýzy funkcionality nástrojů pro podporu řízení kvality, zařazených podle Talend (2013) v roce 2013 do Gartner Magic Quadrants pro oblast Data Quality (SAS Data Flux, Talend Open Studio MDM a Ataccama), vybraných nástrojů používaných pro přípravu dat v rámci procesu získávání znalostí z databází (SAS/BASE, SAS Enterprise Miner, Rapid Miner), specifických nástrojů pro řešení problematiky kvality dat v rámci Big Data (Pig, Hive, Impala, Talend Open Studio for Big Data, Dedoop) a analýzy níže citovaných zdrojů. Jednotlivé pojmy uvedené v souvislosti s typickými činnostmi budu definovat v následujících odstavcích.

Konečně na operativní úrovni probíhá implementace jednotlivých mechanismů, kontrol a jejich kontinuální monitoring. Při členění základních činností se metodika CADAQUES opírá o model ITGPM popisovaný v (Voříšek J. et al., 2008), který rovněž uvažuje těžiště řízení dat jako jednoho z klíčových zdrojů informatiky na taktické úrovni řízení, a tzv. Hierarchii řízení dat publikovanou v (Dyché, Levy, 2006), řadící Data Governance na úroveň strategického řízení dat.

Podstatou standardizace je syntaktické a sémantické sladění hodnot jednotlivých atributů. Za tímto účelem jsou aplikována zejména pravidla pro sjednocení velikosti písmen v jednotlivých tokenech a tzv. standardizační schémata, která převádějí různá synonyma, přezdívky, četné překlepy / chybné zápisy tokenů na jejich standardní zápis. Z pohledu vlastností dat se zaměřuje na Srozumitelnost dat. Představuje též hlavní vstup pro některé další navazující činnosti jako je porovnávání a slučování. S rolí standardizace se můžeme setkat napříč všemi typy dat v rámci univerza. V případě Linked Data a metadat roli standardizačních schémat plní slovníky (soubor termínů z určité oblasti) a ontologie (soubor termínů a jejich vztahů z určité oblasti), coby doporučené standardy. Za benefity použití standardizovaných metadat považuje (Cox, 2013) mimo jiné (1) úsporu času při hledání informace, (2) vyhledání přesnějších výsledků, (3) usnadnění client-server komunikace. V případě multimédií lze považovat za standardy např. doporučené rozlišení obrázku, použitá komprese, či kódování barev.

Validace je zpravidla založena na aplikaci syntaktických masek, vytvořených pomocí regulárních výrazů, WHERE podmínek v rámci nějakého dotazovacího jazyka (zpravidla SQL), kontrolních součtů, na aplikaci IF THEN byznys pravidel, porovnávání skutečných hodnot s referenční bází / standardem či definovaným povoleným rozsahem hodnot. Z uvedených příkladů je zřejmá vazba na vlastnosti dat Syntaktická správnost, Sémantická správnost, Interní a Externí konzistentnost. V rámci jednotlivých typů zdrojů dat má validace pouze technologická specifika své realizace. V případě Linked Data např. pomocí jazyka

SPARQL (SPARQL Protocol and RDF Query Language). V případě obrázků, audia, videa je velkou část validačních procedur možné realizovat až po jejich převedení do strukturované podoby (např. obrázek znázorňující světle zelené auto na modrém pozadí převést na trojici atributů objekt = auto, barva objektu = světle zelená, barva pozadí = modrá).
Porovnávání a slučování je realizováno zejména v souvislosti s deduplikací záznamů a rozhodováním, zda je nově vstupující záznam již obsažen v deduplikované bázi, či se jedná o záznam zcela nový. Souvisí tedy s vlastností Unikátnost dat. Pro tyto účely lze použít jednak řadu přístupů založených na kalkulaci měr podobnosti řetězců – viz (Chaudhuri et al., 2003), anebo porovnávacích kódů – viz (SAS Institute, 2008), jejichž nespornou výhodou je možnost jejich trvalého uložení v rámci atributů tabulek při současném zohlednění určité míry volatility, reflektující možné překlepy. Komplexní logika stojící v pozadí algoritmů generování porovnávacích kódů vychází z podrobné doménové znalosti o atributu (resp. atributech), na základě nichž je porovnávací kód vytvářen. Pro optimalizaci těchto metod jsou dle (Chaudhuri et al., 2003) používány fonetické algoritmy, kódující shodným způsobem stejně znějící hlásky, a tzv. blokovací strategie, využívající hodnot dodatečných atributů k redukci počtu porovnávaných záznamů. Jak ukazuje nástroj Dedoop, tato logika je zachována i v případě deduplikace nestrukturovaných dat v prostředí Hadoop. V případě multimediálních dat je deduplikace v triviálním případě založena na porovnávání bitů jednotlivých souborů (např. pomocí nástroje AntiTwin).
Doplňování chybějících pozorování souvisí s vlastností dat Úplnost. V prvé řadě lze do této oblasti zařadit identifikaci mechanismu výskytu chybějících hodnot – viz např. (Marwala, 2009), reflektujícího náhodnost výskytu a možnosti odvození chybějících hodnot pomocí ostatních atributů. Na jeho základě je následně zvolena strategie doplnění. Na tuto oblast navazuje definice vlastních metod pro doplňování (imputaci) chybějících pozorování. Problematiku chybějících pozorování lze chápat též jako obohacování stávajících dat o dodatečné zdroje (např. externí číselníky, registry, geokoordináty) a znalosti (vztahy mezi entitami jako je např. identifikace domácnosti či obchodní spřízněnost jednotlivých subjektů). Logika imputace opět zůstává stejná napříč univerzem dat, pouze technologická realizace je odlišná. King, Kutyniok a Lim (2013) například popisují metodu Image Inpainting pro doplňování chybějících pixelů v rámci obrazů. Tato metoda je podobně jako mnoho metod pro imputaci v prostředí relačních databází popsaných v (Marwala, 2009) založená na ostatních hodnotách / atributech (v tomto případě pixelech obrazu).
V rámci auditu kvality dat metodika CADAQUES uvažuje v prvé řadě realizaci tzv. technického profillingu Ten je založen na zjišťování základních popisných statistik jednotlivých datových atributů, identifikaci syntaktických vzorů, porovnávání reálných dat s metadaty a ověřování referenční integrity (tedy činností směřujících ke stanovení objektivních vlastností dat). Na základě doporučení formulovaných v (Lee et al., 2006) rozšiřuje tento kvantitativní audit o kvalitativní šetření cílené na hlavní uživatele dat. Tento retrospektivní audit kvality rovněž doporučuje doplnit o analýzu stávajících procesů s cílem identifikovat potenciální budoucí hrozby. Doporučuje se při tom zaměřit zejména na ty procesy, u nichž byly zjištěny závažné nedostatky v rámci technického profillingu nebo kvalitativního šetření. V kostce se tedy jedná o měření vlastností dat, testování současných navržených kontrol a identifikace potenciálních defektů. Tento přístup lze opět aplikovat napříč celým univerzem dat.

3.3 Model zralosti datového zdroje

Jak bylo ukázáno na příkladu Linked Data, použití externích dat může s sebou přinášet rizika vyplývající z jejich původu a způsobu vytváření. Pro podporu rozhodování o použití konkrétního externího datového zdroje jsem proto jako součást metodiky CADAQUES

vyvinul model zralosti datového zdroje. Model vychází z mnou definované redukované sady vlastností dat popsané v části 2.2 tohoto článku. Důvodem je princip jednotného řízení těchto vlastností napříč celým univerzem, tedy i v rámci externích dat. Jednotlivým vlastnostem přiřazuje model konkrétní otázky, na které je dle mých zkušeností vhodné se zaměřit. Jedná se o vícekriteriální model, jehož jednotlivé části mohou mít pro různé zdroje a očekávané užití odlišné váhy. Výsledné rozhodnutí o použití je realizováno pomocí porovnání součtu vah s předem stanovenou kritickou hodnotou v rámci každé dimenze.
V případě endogenní dimenze obsahuje model tyto otázky (doplněné o komentáře):

- Patří tento datový zdroj mezi obecně používané v rámci daného odvětví? Pokud je zdroj obecně používán, je značná pravděpodobnost, že je důvěryhodný.
- Je znám původ dat a je akceptovatelný pro předpokládané užití? Pokud je znám původ, lze odhadnout jeho důvěryhodnost.
- Je k dispozici popis prováděných kontrol? Pokud je znám, lze odhadnout důvěryhodnost tohoto zdroje a též publikovaných hodnot jednotlivých vlastností.
- Je k dispozici informace o garantované míře unikátních záznamů, sémantické správnosti, syntaktické správnosti a je postačující pro předpokládané užití?
- Je k dispozici informace o garantované míře přesnosti reprezentace reálných entit a je postačující pro předpokládané užití? V případě chybějící informace ji lze alespoň částečně aproximovat jako *min* (syntaktická správnost, sémantická správnost, aktuálnost, úplnost).

V případě časové dimenze obsahuje tyto otázky:

- Je k dispozici informace o garantované míře aktuálnosti údajů a její úroveň je postačující pro předpokládané užití?
- Je k dispozici informace o frekvenci změn v reálném světě?
- Je k dispozici informace o frekvenci synchronizace jednotlivých částí zdroje publikovaných dat a je akceptovatelná z hlediska předpokládaného užití?

V případě kontextuální dimenze obsahuje tyto otázky:

- Je k dispozici informace o garantované míře konzistentnosti jednotlivých atributů / elementů a její míra je akceptovatelná z hlediska předpokládaného užití?
- Je k dispozici informace o garantované míře konzistentnosti zdroje s ostatními zdroji a tato míra je akceptovatelná z hlediska předpokládaného užití?
- Je k dispozici informace o garantované míře úplnosti zdroje a očekávané míře pokrytí všech výskytů hodnot a tato míra je akceptovatelná z hlediska předpokládaného užití?

V případě dimenze užití obsahuje tyto otázky:

- Jsou k dispozici informace o omezeních v dostupnosti datového zdroje a tato omezení jsou akceptovatelná z pohledu předpokládaného užití?
- Je na straně zdroje k dispozici kontaktní osoba byznys správce dat schopného podat dodatečné informace? Je k dispozici popis původního účelu, pro který byl zdroj vytvořen, a informace popisující omezení použitelnosti zdroje a jsou v souladu se záměrem použití dat?
- Je k dispozici byznys popis na úrovni jednotlivých atributů / elementů, je znám jejich význam a je postačující pro daný účel?

- Je známa osoba technického správce zdroje, který by mohl podat dodatečné informace? Byl při vytváření datového zdroje použit kanonický datový model nebo některá ze známých ontologií? Je znám technický popis jednotlivých atributů a elementů (datové typy a délky, integritní omezení, jsou k dispozici příklady hodnot)? Je dostupný popis integrační logiky pro dávkovou integraci (informace o použitém oddělovači nebo použité pevné šířce, informace o použitém formátu, použité znakové sadě)? Je dostupný popis integrační logiky pro online integraci na úrovni jednotlivých služeb?
- Je v rámci přenosu dat použita úroveň zabezpečení adekvátní charakteru dat?

V případě ekonomické dimenze obsahuje tyto otázky:

- Jsou náklady na pořízení datového zdroje a jeho aktualizaci adekvátní přínosům datového zdroje z pohledu předpokládaného užití?
- Jsou náklady na integraci, uložení či archivaci v datovém úložišti a jejich zpřístupnění uživatelům adekvátní přínosům datového zdroje z pohledu předpokládaného užití?
- Jsou náklady na zajištění bezpečného přístupu a zabránění neautorizovanému přístupu k datům adekvátní přínosům datového zdroje z pohledu předpokládaného užití?

3.4 Podpůrné nástroje metodiky CADAQUES

Metodika v současné době obsahuje několik podpůrných nástrojů pro její implementaci: (1) kanonický datový model pro vertikálu pojišťovnictví, (2) šablonu auditorské zprávy, (3) šablonu matice pro definici užití jednotlivých atributů, (4) šablonu matice měřených vlastností, (5) dokument mapující kroky IT Assurance Road Map (součást návodu IT Assurance Guide: Using COBIT) na jednotlivé konvenční kroky auditu datové kvality popsané v (McGilvray, 2008; English, 1999; Lee et al., 2006).

4 Závěr

Univerzum v současnosti zpracovávaných dat je rozmanité. Zahrnuje v sobě data o různé míře strukturovanosti, pro jejichž uchování jsou typické různé technologie. Přestože řízení kvality má v případě některých zástupců tohoto univerza určitá specifika, nestačí dle mého názoru tato skutečnost jako argument pro vznik samostatných disciplín (Linked Data Quality, Metadata Quality, Big Data Quality, apod.) a je vhodnější uvažovat jednotný koncept řízení kvality dat napříč těmito zdroji, zejména s ohledem na předpoklad, že současné univerzum dat není konečné. Dalším důvodem pro jednotný přístup jsou některé základní principy Data Governance, jako je řízení rizik spojených s používáním dat či princip komplexního řízení dat a informačního obsahu organizace. Dosud identifikovaná specifika se týkají buď konkrétní použité technologie (např. NoSQL a grafové databáze), formátu dat (např. audio, video, grafické soubory), anebo vyplývají ze skutečnosti, že se nejedná o zdroj původní (např. Linked Data).

Metodika CADAQUES, představená v rámci tohoto článku, je příkladem přístupu uvažujícího komplexní řízení kvality dat a informací napříč celým univerzem. Hlavní součástí této metodiky je sada základních principů a činností, které je možné aplikovat na všechny datové zdroje. Řada z nich vychází z výše zmíněných principů Data Governance. Jedním z klíčových doporučení této metodiky je zaměření se na relativně malou sadu vlastností dat, kterou lze efektivně řídit. Součástí metodiky je rovněž model zralosti datového zdroje, který slouží pro zhodnocení míry rizika spojené s použitím konkrétního zdroje. Většina stavebních prvků této metodiky je dostupná registrovaným uživatelům na portálu http://www.dataquality.cz.

Seznam použitých zdrojů

Batini, C., Scannapieco, M. (2006). *Data Quality: Concepts, Methodologies and Techniques*. Berlin: Springer-Verlag.

Cox, L. (2013) *Metadata: 96 Most Asked Questions - What You Need To Know*. Emereo Publishing.

Doucek, P., Novák, L., Nedomová L., Svatá, V. (2011). *Řízení bezpečnosti informací*. Příbram: Professional Publishing.

Dyché, J., Levy, E. (2006). *Customer data integration: Reaching a Single Version of the Truth*. New Jersey: Wiley & Sons.

English, L. P. (1999). *Improving Data Warehouse and Business Information Quality: Methods for Reducing Costs and Increasing Profits*. New Jersey: Wiley & Sons.

Howard, P. (2008). The importance of a common data model. *IT-Director.com.* Retrieved from http://www.it-director.com/technology/applications/content.php?cid=10292.

Hyland, H., Elliott, L. (2008). *No Data Left Behind: Federal Student Aid A Case History*. Retrieved from http://www.dama-ncr.org/Library/2008-03-11NoDataLeftBehind.ppt.

Chappell, D. A. (2004). *Enterprise Service Bus*. Sebastopol: O'Reilly.

Chaudhuri, S., Ganjam, K., Ganti, V., Motwani, R. (2003). Robust and Efficient Fuzzy Match for Online Data Cleaning. In: *SIGMOD 2003*. CA: San Diego.

Juran, J.M., Godfrey, A.B. (2010). *Juran's Quality Handbook: The Complete Guide to Performance Excellence*. New York: McGraw-Hill.

King, E.J., Kutyniok, G., Lim, W. (2013). Image inpainting: Theoretical analysis and comparison of algorithms. In *Wavelets and Sparsity XV: Proceedings of SPIE - The International Society for Optical Engineering 2013*.

Král, J., Žemlička, M. (2006) Kvalita dat a informací – základní omezení IT ve veřejné správě. In Pour, J., Voříšek, J (Eds.) *Systems Integration 2006* (pp. 215-222). Prague: University of Economics.

Ladley, J. (2012). *Data Governance: How to design, deploy and sustain an effective Data Governance program.* Waltham: Morgan Kaufmann.

Lee, Y. W., Pipino, L. L., Funk, J. D., Wang, R. Y. (2006). *Journey to Data Quality*. MA: MIT Press.

Marwala, T. (2009). *Computational Intelligence for Missing Data Imputation, Estimation, and Management: Knowledge Optimization Techniques*. Hershey: Information Science Reference.

McGilvray, D. (2008). *Executing Data Quality Projects: Ten Steps to Quality Data and Trusted Information*. Burlington: Morgan Kaufmann.

MediaBistro. (2010). Quality Indicators for Linked Data Datasets. *Semanticweb.com.* Retrieved from http://answers.semanticweb.com/questions/1072/quality-indicators-for-linked-data-datasets.

Metadata Working Group. (2010). *Guidelines for Handling Metadata*. Retrieved from http://www.metadataworkinggroup.org/pdf/mwg_guidance.pdf.

Olson, J. E. (2003). *Data Quality: The Accuracy Dimension*. Waltham: Morgan Kaufmann.

Pejčoch, D. (2011). Vztah řízení dat k ostatním oblastem řízení informatiky. In *Sborník prací účastníků vědeckého semináře doktorandského studia Fakulty informatiky a statistiky VŠE v Praze* (pp. 3-13). Praha: Oeconomica.

Pejčoch, D. (2012). Audit datové kvality podle IT Assurance Guide: Using COBIT - 3. díl. In *Data Quality CZ*. Retrieved from http://www.dataquality.cz/index.php?ID=5&ArtID=13&clanek=201203_DQA_IT_Assurance_Guide_3dil.

Pipino, L., Lee, Y. W., Wang, R.Y. (2002). Data quality assessment. *Communications of the ACM*. 45(4), 211-218.

Redman, T. (2001). *Data Quality: The Field Guide*. Boston: Butterworth-Heinemann.

SAS Institute. (2008). *SAS Data Quality Server 9.2: Reference*. Cary, NC: SAS Institute Inc.

SDM. (2005). Definice pojmu data quality. *Techtarget.com* Retrieved from http://searchdatamanagement.techtarget.com/definition/data-quality.

Soares, S. (2012). *Big Data Governance: An Emerging Imperative*. Boise: McPress.

Strong, D.M., Lee, Y.W., Wang, R.Y. (1997). Data quality in context. *Communications of the ACM*. 40(5), 103-110.

Štumpf, J., Džmuráň, M. (2008). Datová integrace prostřednictvím společného datového modelu. In *Proceedings of the 16th International Conference on Systems Integration*. Praha: CSSI.

Talend. (2013). Analyst Report: Magic Quadrant for Data Quality Tools. *Talend.com* Retrieved from https://info.talend.com/dataqualitytools.html.

Voříšek J. a kol. (2008). *Principy a modely řízení podnikové informatiky*. Praha: Oeconomica.

Wang, R.Y., Strong, D.M., Guarascio, L.M. (1996). Beyond Accuracy: What data quality means to data consumers. *Journal of Management Systems*. 12(4), 5-34.

Zaveri, A., Rula, A., Maurino, A. Pietrobon, R., Lehmann, J., Auer, S. (2012) Quality Assessment Methodologies for Linked Open Data. S*emantic-web-journal.net.* Retrieved from http://www.semantic-web-journal.net/content/quality-assessment-methodologies-linked-open-data.