Evropská unie
Evropský sociální fond
Operační program Zaměstnanost

Metodika zajišťování datové kvality v rámci Technologické agentury ČR

– se zaměřením zejména na činnost Oddělení strategií a analýz a Evaluačního úseku a dále na vývoj a provozování informačních systémů TA ČR

Část I

Zpracováno v rámci projektu ProEval

Úvod

Evidence based policy nelze praktikovat bez adekvátní datové podpory. Je tedy zřejmé, že kvalita evidence based policy je výrazně ovlivněna kvalitou použitých dat. Téma kvality dat však bylo v rámci Technologické agentury ČR z nejrůznějších důvodů poněkud opomíjeno. Tato metodika by proto měla především nastolit nejrůznější otázky týkající se kvality dat v různých kolekcích TA ČR a nabídnout relevantním skupinám zaměstnanců postupy, které povedou ke zvyšování datové kvality a v neposlední řadě také pomoci budovat povědomí o důležitosti pojmu datové kvality.

Cílem materiálu není poskytnout teoretický vhled do problematiky datové kvality, ale spíše poskytnout stručný souhrn informací o konceptech datové kvality prakticky využitelných v rámci Technologické agentury ČR spolu s postupy, které lze v oblasti datové kvality aplikovat.

Základní pojmy a činnosti v oblasti datové kvality

Teoreticky zaměřené práce začínají výkladem hierarchie *"data-informace-znalosti"* (s případným doplněním *moudrosti*). Z naší perspektivy se však nejedná o nezbytnou kapitolu, v rámci dalšího textu postačí intuitivní představa o rozdílech mezi daty, informacemi a znalostmi. Podíváme se zejména na pojem datové kvality a její (případné) měření.

Definice datové kvality

Datová kvalita, stejně jako řada pojmů na obdobné úrovni obecnosti v oblasti informatiky nemá jednotnou, obecně sdílenou definici – existuje řada definic různých autorů, nicméně z pohledu reálné praxe jsou však rozdíly mezi nimi zanedbatelné.

Pro ilustraci uvádíme dvě vybrané definice:

> *„Data mají vysokou kvalitu, pokud tato odpovídá jejich zamýšlenému užití v provozu, rozhodování a plánování"*
>
> Joseph M. Juran, konzultant v oblasti managementu kvality

> *„Datová kvalita je stav úplnosti, konzistentnosti, včasnosti a přesnosti, který činí data vhodné pro konkrétní využití."*
>
> Will Helmer, konzultant v oblasti IT

V obou těchto definicích je shodný důraz na **využití** v praxi. Jinými slovy, nelze oddělit kvalitu dat od jejího použití v organizaci, čili datová kvalita je relativní vůči kontextu organizace.

Měření kvality dat

I když není radno zveličovat význam nejrůznějších metrik v procesu řízení čehokoliv, v procesu řízení datové kvality mohou vybrané metriky kvality dat poskytnout podnětný vhled do problematiky a nasměrovat nás k identifikaci klíčových oblastí.

Metriky, o které nám jde, lze rozdělit do dvou základních tříd:

- **objektivní** – typicky kvantitativní – metriky. Nezáleží u nich na tom, kdo určuje jejich hodnoty, dají se typicky určit zpětně a jejich měření je bez problémů opakovatelné.
  Příkladem objektivní metriky může být např. procento syntakticky nesprávných hodnot v konkrétním sloupci/atributu – např. u rodných čísel
- **subjektivní** – často ne-kvantitativní – údaje, které jsou spojeny s určováním, popř. hodnocením vybraných aspektů dat jednotlivými uživateli/hodnotiteli. Výsledky jsou na nich tudíž závislé. Nemusí být možné určit tyto hodnoty zpětně a nemusí být opakovatelné.

Příkladem může být např. otázka důvěryhodnost, srozumitelnost, užitečnost aj.

Přestože objektivní metriky jsou z řady důvodů IT či analytické veřejnosti „sympatičtější“, budeme se věnovat jak objektivním, tak subjektivním metrikám (neboť i data splňující vysoký standard kvality na základě hodnocení objektivními metrikami nemusejí být v praxi užitečná, naproti tomu i ne zcela kvalitní data podle objektivních metrik mohou mít vysoký přínos).

Následující seznam shrnuje vybrané čtyři vlastnosti spolu s jejich způsoby měření s návaznostmi na příslušné metriky.

Vlastnost dat / možná metrika	Způsob měření
Důvěryhodnost Míra všeobecné akceptovanosti dat uživateli, míra přesvědčení uživatelů o jejich správnosti	Dotazníkové šetření, řízené rozhovory s uživateli
Unikátnost Míra duplicitních záznamů	Analýza shlukování/slučování podle daných kritérií
Syntaktická správnost Podíl hodnot, které neodpovídají formálním pravidlům	Kontrola dat z hlediska splňování pravidel, vyjádření podílu počtu hodnot, které neprošly těmito testy
Sémantická správnost Podíl chybných hodnot daného atributu	Porovnání s validními zdroji (existence jména v databázi jmen), analýza outlierů („věk osoby nad 150 let“) a vyjádření příslušného podílu.

Těmito čtyřmi vlastnostmi se budeme v dalším textu zabývat, další z nich uvádíme bez vysvětlení či ilustrace spíše pro dokreslení komplexnosti tématu.

- Aktuálnost
- Množství zjevně neaktuálních údajů
- Včasnost
- Volatilita
- Konzistentnost
- Úplnost
- Pokrytí
- Dostupnost
- Srozumitelnost
- Interoperatibilita
- Bezpečnost
- Náklady na pořízení a aktualizaci
- Náklady na uložení, sdílení, distribuci, zálohování a archivaci

- Náklady na ochranu
- Přesnost

Uvedené vlastnosti (a k nim příslušné metriky) nejsou samozřejmě nezávislé, jsou mezi nimi nejrůznější korelace – ovlivnění jedné může znamenat ovlivnění jiné atp. Vlastnosti lze sdružovat do skupin (endogenní aspekty, časové aspekty, kontextuální aspekty, …).

Poslední vlastností v uvedeném seznamu je **přesnost**, která se poněkud vymyká ostatním vlastnostem/metrikám – jde v podstatě o soulad s realitou, přičemž je zřejmé, že ověřování této vlastnosti je kvalitativně na jiné úrovni než u ostatních vlastností.

Příčiny nedostatečné kvality dat

Příčin, proč mohou být v databázi/datovém souboru/kolekci nekvalitní data, je celá řada. Lze je však rozdělit do kategorií a příčiny z jednotlivých kategorií řešit samostatně.

Přenos dat z vnějších zdrojů

Tato skupina příčin se týká jak manuálního, tak automatizovaného zpracování dat. Jde v podstatě o propagaci chyb z jednoho zdroje dále.

Datové konverze

Naplnění databáze se typicky začíná tím, že se do ní konvertují data z jiných databází. Není-li zajištěna kvalita ve výchozím zdroji, je situace v novém zdroji stejná jako v předchozím. Při konverzi však může navíc dojít ke ztrátě (meta)dat.

V rámci TA ČR se může jednat např. o konverzi dat z IS Patriot do IS TA.

Systémová konsolidace

Situace částečně podobná jako v předchozím případě, zkomplikovaná ovšem tím, že nejde o přenesení dat z původního zdroje do nového, ale o existenci více datových zdrojů, které mají být pokryty jednou databází. Struktura databází, které se konsolidují, bývá typicky odlišná, a tudíž snadno vznikají značné problémy, “podpořené” třeba ještě nedostatečnou dokumentací. V určitě databázi mohou být např. tituly psány do jednoho atributu, v jiné mohou být rozděleny na „tituly před“ a „tituly za“. Pokud se nepodaří zvládnout způsob rozdělování, vede tento proces ke snížení kvality dat.

Významným problémem v této situaci je též riziko vzniku duplicitních záznamů.

V rámci TA ČR se může jednat např. o konsolidaci dat z více zdrojů do DAFOS.

Manuální zadávání dat

Manuální zadávání bylo, je i bude významným způsobem zadávání dat do systémů mnoha systémů. I přes maximální snahy tvůrců uživatelského rozhraní a opatrnost uživatelů (na které se však reálně nelze spoléhat) dochází k nejrůznějším chybám. Jejich kategorizace

a analýza jejich výskytu je důležitá v procesu zlepšování uživatelského rozhraní i edukace/upozorňování uživatelů a tvorbu nápověd, stejně jako při samotném čištění dat.

Následující seznam poskytuje typologii problémů, s nimiž se setkáváme:

- Překlepy a přepsání (přesmyčky, záměna „y“ a „z“)
- Gramatické chyby
- Typografické chyby (prostrkávání, psaní zbytečných kapitálek/verzálek, zaměňování pomlček, spojovníků – problém např. ve jménech firem)
- Zadávání dat bez diakritiky
- Špatné výběry např. z roletových menu, zaškrtnutí vedlejších check-boxů
- Nevyplnění pole
- Nedodržení stanovených standardů (např. na formát tel. čísla, kalendářních dat)
- Vložení dat do špatného pole (mailová adresa na místo tel. čísla – reálný případ)
- Nepřípustné kombinace dat
- Neznalost (např. zapsat ampersand, zavináč, … – různé náhrady)
- Nepochopení obsahu příslušného pole formuláře (snaha vydedukovat „co by tam tak mohlo být“ a následné chybné vyplnění)
- Úmyslné chyby (vyžadování informací, které uživatel nechce vyplnit – uživatel pak zkouší, které smyšlené údaje „projdou“)
- Nemožnost zadat správnou hodnotu (např. vyžadování rodného čísla od osoby ze zahraničí)

V případě TA ČR se jedná o kapitolu velmi významnou, neboť řada klíčových dat pochází z manuálního zadávání uživateli – ať už se jedná o registraci nových oponentů či podávání projektové žádosti.

Dávkové zpracování

Při dávkovém zpracování může docházet ke vzniku dalších chyb, není-li proces dávkového zpracování náležitě otestován/ošetřen, situace je podobná jako v prvním případě.

V případě TA ČR se může týkat zejména systému DAFOS.

Realtimové zacházení

Při požadavku na rychlé zpracování nelze často provádět potřebné kontroly datové kvality a tak při nejrůznějších procesech dochází k propagaci chyb, opět podobné prvními případu.

V TA ČR se může týkat perspektivně propojování DAFOSu a ISTA.

Interní změny dat

Zpracování dat

Informační systémy se typicky používají ke zpracování dat. Data z databáze jsou načtena, uživatel prostřednictvím uživatelského rozhraní spouští funkcionalitu aplikační logiky, dochází ke změnám dat, která jsou následně ukládána zpět do databáze.

Není-li funkcionalita patřičně otestována, mohou vznikat nejrůznější chyby v datech.

Problém při zpracování dat nemusí ovšem nutně pocházet z „programátorského směru“ – problémem může být např. i spuštění nějakého procesu dříve, než měl být proveden (např. rozeslání informace o výsledku nějaké akce, která ještě nebyla skončena).

Jedná se o záležitost, která se týká v podstatě všech informačních systémů, TA ČR nevyjímaje.

Čištění dat

Ač je to na první pohled paradoxní, samotný proces čištění dat (který je původně zamýšlen jako akce ke zvýšení datové kvality) může vést k jejímu snížení.

Rizikem jsou zejména hromadné úpravy dat, jejichž výběr může být proveden na základě pravidel, která zahrnou i ta data, která ve skutečnosti neměla být měněna. To může vést k tomu, že u části dat se kvalita po čištění zvýší, u jiné ovšem sníží. Riziko roste s komplexitou pravidel a míra dopadu s množstvím dat, která se čistí.

V rámci TA ČR jde o jakékoliv akce čištění dat.

Mazání dat

Odebírání nepotřebných/neaktuálních záznamů opět přináší rizika. V případě odebírání uživatelem do hry vstupuje riziko „obyčejné chyby“ (uklepnutí, …). V případě automatizovaného/hromadného procesu odebírání nastávají problematické situace zejm. v případě změn struktury databáze – datových formátů aj., přičemž se neupraví skript/program/aplikace, která má na starosti mazání těchto nepotřebných záznamů.

Vede to k chybám – řečeno slovy statistiků – prvního i druhého druhu: nejsou odebrána data, která odebrána být mají, naproti tomu mohou být odebrána data, která odebrána býti nemají.

Tyto procesy samozřejmě mohou interagovat i s dalšími datovými „nekvalitami“ – manuálně vyplněná data mohou „náhodou“ vyhovět podmínce pro automatické mazání, a tak dojde k výmazu reálně potřebných dat.

V rámci TA ČR jde o akce jakéhokoliv mazání dat.

Manipulace s daty

Nezachycení změn v datech

Data by měla korespondovat s realitou, která je však přirozeně mění. Firma změní sídlo, akademický pracovník získá další titul – o těchto změnách se může a nemusí provozovatel/správce databáze dozvědět.

V rámci TA ČR je toto téma relevantní v mnoha datových zdrojích, databáze obsahují řadu údajů „rizikového charakteru“.

Upgrady systému

Upgrady jsou jedním z dalších zdrojů datové nekvality. I přes snahu o důkladné otestování, nebývají testy prováděny často na nekvalitních datech – naopak se provádějí na datech, která jsou umělá a kvalitní. Nasazení do reálného provozu pak může vést ke snížení datové kvality.

V rámci TA ČR se jedná o téma svázané s vývojem jednotlivých IS a procesem vývoje, testování a releasování.

Nové využití dat

Měnící se potřeby a reálný stav informačního systémů vede uživatele ke „kreativitě“ při využívání nejrůznějších polí – pole pro poznámky se najednou stane místem, kde se ukládají nejrůznější data, která by reálně měla fungovat jako samostatné atributy. I to je příčinou datové nekvality.

V rámci TA ČR by mohla být provedena analýza využívání nejrůznějších volných textových polí aj.

Ztráta odborníků

Mnozí klíčoví odborníci, kteří pracují s daty zajišťují udržování datové kvality nejrůznějšími ad hoc zásahy podle pravidel, která nejsou explicitně zaznamenána. Dojde-li k odchodu takovéhoto pracovníka, vede to automaticky k poklesu kvality dat.

V rámci TA ČR je téma rovněž relevantní. Souvisí s mapováním znalostí jednolivých zainteresovaných zaměstnanců (zejm. IT)

Automatizace procesů

Při manuálním zpracování dat člověk často odhalí (syntakticky nebo sémanticky) chybná data a upraví je. Při nasazování automatických procesů se ne vždy podaří zachytit možnou variabilitu dat a některá zjevně chybná data se mohou začít dostávat do systémů.

V rámci TA ČR téma relevantní v souvislosti s rozvojem informačních systémů a

rozšiřování jejich funkcionality.

Dopady nízké datové kvality

Je zjevné, že nízká datová kvalita se negativně projeví v řadě oblastí, v tomto kontextu lze uvažovat o následujících aspektech:

- Ekonomické
- Časové
- Analytické
- Pokles důvěryhodnosti
- Legislativní

Tyto aspekty se navzájem prolínají, nekvalitní data vyžadují typicky větší zdroje (např. osobní náklady) na úpravy či korekce dat, což je přímo spjato s časem, kvalita analýz je odvislá od kvality vstupních dat, nekvalitní analýzy mohou přispívat k případnému poklesu důvěryhodnosti atp. (V případě, že analýzy jsou využity v rámci přípravy podkladů pro legislativní proces, lze hovořit následně i o legislativních dopadech.)

Činnosti v procesu řízení datové kvality

Následující sekce obsahuje přehled různorodých činností, se kterými se setkáváme v nejrůznějších fázích procesu řízení datové kvality. Mnohé z nich však lze aplikovat i mimo kontext procesů – v případě tvorby ad-hoc analýz nemusí být samozřejmě procesy k dispozici, přesto lze mnohé z těchto činností provádět.

Data profiling

Jedná se o soubor technik, které slouží k prvotnímu prozkoumání struktury dat, k detekování anomálií, provádění analýz (hodnot) atributů. Může jít o nejrůznější ***frekvenční analýzy*** atp. Cílem je ***získat základní představu o datech,*** aby mohly být následně prováděny další činnosti, které se bez základní představy o charakteru dat neobejdou.

Standardizace

Standardizace je proces, při kterém dochází k převodu dat do jednotné podoby odpovídající jejich charakteru. Může se jednat např. o parsování (“rozpad”) adres a jejich převod do podoby „jméno ulice–číslo popisné–číslo orientační), ujednocení konvence psaní jmen vzhledem k velkým a malým písmenům aj.

Porovnávání a slučování

Jde o proces eliminace duplicitních záznamů – jejich identifikace prostřednictvím více či méně sofistikovaného porovnávání a v případě, že záznamy vykážou určitou hraniční míru podobnosti, je konkrétním způsobem vytvořen jeden záznam obsahující informace z jednoho či obou záznamů.

Doplňování chybějících záznamů

Vlivem nejrůznějších okolností mohou mít záznamy chybějící atributy. Jejich doplnění je další činností při zvyšování datové kvality.

Validace

Validace je činnosti, při kterých je ověřováno, zda data jsou validní (správná) – ať už syntakticky či sémanticky. Protože se jedná o klíčovou činnost, rozebereme ji následně poněkud podrobněji.

Obohacování o externí zdroje

Jde o rozšiřování databází o další zdroje, typicky oficiální číselníky, registry aj. s cílem napomoci verifikaci a validaci dat.

Validace dat

Validace dat bývá definována jako proces kontrolování, že data vyhovují specifikaci. Je to po profilingu první prováděný proces nad nezpracovanými daty. Podle toho, jakou metodu validace zvolíme, budeme kontrolovat konkrétní vlastnost dat – jde typicky o kontrolu syntaktické a sémantické správnosti, řídčeji pak volatility a konzistentnosti.

Kontrola validity může být prováděna v různých fázích procesu práce s daty, ideálně hned při jejich zadávání do systému – do sféry spadají nejrůznější kontroly ve webových formulářích, které „nevpustí“ do systému nevalidní data. Jak už bylo zmíněno, vyžadování vyplnění dat všude však může mít na kvalitu dat negativní vliv. Proto má smysl provádět i ex-post validaci, tj. v momentě, kdy již data v systému jsou. Validace v této fázi je výchozím bodem pro další čištění dat.

Syntaktická správnost

Jde vlastně o formální kontrolu správnosti. Typicky se při ní používají regulární výrazy business pravidla (IF-THEN) pravidla.

Regulární výrazy

Regulární výrazy jsou řetězce speciálních znaků, které reprezentují jakýsi „mustr“ – vzor či masku, jemuž musí porovnávaný řetězec znaků vyhovovat: např. mohou popisovat, že řetězec může začínat pouze velkým písmenem, následovat může několik číslic, pak právě jedna tečka, … atp.)

Příkladem využití regulárních výrazů může být např. kontrola syntaktické správnosti emailových adres, které vyhoví řetězce typu klm@bfm.cz, ale nikoliv kpc%seynam.cy.

Business pravidla

Business pravidla, jsou pravidla, která mají charakter podmínek: jestliže..., pak...

(IF-THEN), která můžeme následně přepsat do podoby nějakého skriptu, který bude procházet a kontrolovat data.

Příkladem business pravidla může být sada podmínek pro rodné číslo.

Sémantická správnost

Při kontrole sémantické správnosti se kontroluje, jestli daný atribut obsahuje „správnou“ hodnotu. Příkladem hodnoty, která by v případě sémantické kontroly neměla projít, je mail vyplněný jako cosi@gmial.cz – tato adresa splňuje sice všechny podmínky kladené na tvar emailové adresy, nicméně neexistuje na této doméně příslušný mailový server, který by poskytoval příslušné služby.

V praxi se sémantická správnost kontroluje zejména srovnáváním s relevantními číselníky a seznamy, např. seznam křestních jmen, příjmení, názvů obcí, PSČ, …

Konzistence

Konzistence je brána jednak jako soulad záznamů napříč zdroji (příslušná data o jedné entitě ve více databázích by se měla shodovat, dojde-li k úpravě v jedné databázi, pak musí tato změna být propagována i v dalších), druhak může být brána jako soulad mezi hodnotami, mezi kterými je nějaký vztah, např. jméno města a PSČ (oba údaje mohou být sémanticky validní – oba údaje mohou být obsažena v oficiální databázi názvů měst a obcí, resp. PSČ, ale nemusí si odpovídat – nekonzistence). Podobná nekonzistence může vzniknout v případě rodného čísla a pohlaví.

Pro ověřování konzistentnosti se typicky používají business pravidla nebo hashovací funkce (ty půjdou mimo rozsah tohoto pojednání).

Evropská unie
Evropský sociální fond
Operační program Zaměstnanost

Část II

Zpracováno v rámci projektu ProEval

Evropská unie
Evropský sociální fond
Operační program Zaměstnanost

Stručná metodika zajišťování datové kvality v rámci aktivit Oddělení strategií a analýz (OSA)

V následujících odstavcích stručně popíšeme kroky při zajišťování datové kvality spolu s konkrétními příklady a komentáři.

Kroky při zajišťování datové kvality

Vytipování atributů

Prvním krokem při čištění dat je určení, které atributy daných entit jsou pro danou úlohu významné – to je odvislé zejména od užití výsledků. V praxi OSA a dalších oddělení první sekce není realistické provádět formální analýzy (ranking atributů na základě různých kritérií), zcela postačuje expertní odhad dotyčného analytika, případně konsensus několika expertů/analytiků.

Typicky by však měly být mezi vytipovanými atributy primární a cizí klíče a atributy, které jsou klíčové pro řešení dané analytické úlohy.

Profiling

Při každé úloze, při níž je relevantní zabývat se čištěním dat, je první z prováděných činností data profiling. V případě numerických hodnot atributů (nikoliv numerických identifikátorů) je vhodné vytvořit příslušný histogram. V případě identifikátorů, které mají podobu řetězců, pak histogram četností jednotlivých znaků (characters) a rovněž délek identifikátorů.

I přes jednoduchost těchto postupů mohou tyto techniky poskytnout cenné podněty. Tímto způsobem (získání distribuce délek řetězců) můžeme hned v úvodu odhalit např. nesprávné zápisy IČ, např.:

`00216208` vs. `216208`,

k nimž mohlo dojít např. kvůli nastavení tabulkového procesoru a následného exportu/importu a “joinování” dat. Distribuce jednotlivých znaků pomůže odhalit využití problematických nealfanumerických znaků (využívání spojovníků, pomlček, mínusů a jejich zaměňování např. v názvech institucí), výskyt “neočekávaných” znaků, např. nealfanumerických znaků či písmen v IČ aj.:

`ico:00216208` vs. `00216208`

Tímto způsobem také často odhalíme přítomnost českých slov v atributu, ve kterém by se měly nacházet jen anglické výrazy aj.

Standardizace

V případě, že existuje přirozený relevantní standardní tvar, převedeme všechny hodnoty daného atributu do podoby, standardního tvaru. Příkladem může být převod tel. čísla do podoby typu:

`+420 123 456 789`

Do tohoto bodu může patřit též expanze zkratek (“UK” se převede na “Univerzita Karlova”).

Otázky duplicity

Přítomnost duplicitních záznamů je zásadním problémem pro naprostou většinu analýz. Základem odstraňování duplicitních záznamů je existence podobnostní funkce, která vrací míru podobnosti dvou hodnot atributů, příp. (částí) záznamů. Může se jednat např. o editační vzdálenost dvou textových řetězců, které byly určitým způsobem normalizovány/standardizovány.

Normalizací může být např. provedení následující posloupnosti řetězcových operací:

1. odstranění bílých znaků (mezer, tabulátorů, konců řádku, …)
2. převod na malá písmena
3. odstranění diakritiky
4. odstranění nealfanumerických znaků (interpunkce, …, ale také mezer)

V takovém případě ja křetězec
`“Matematicko-fyzikální fakulta”`, tak `“Matematicko-fyzikalni fakulta”`
se převedou na stejný řetěze `“matematickofyzikalnifakulta”`, a přestože na začátku řetězce shodné nebyly, jejich otisky nyní shodné jsou. Při praktickém použití lze nastavit editační vzdálenost na nenulovou, avšak nízkou hodnotu (např. 1 až 2).

Zároveň je vhodné (v případě přijatelně velkého či spíše malého množství potenciálních kandidátů na sloučení) generovat upozornění a následné sloučení potvrzovat “manuálně”. Tento způsob byl s úpravami použit např. při čištění názvů institucí při převodu dat z IS Patriot do ISTA (při této příležitosti však byly názvy porovnávány ještě s kolekcí názvů z externí databáze).

Validace – syntaktická správnost

Při ověřování syntaktické správnosti zjišťujeme, zda daný výraz odpovídá předem daným pravidlům/podmínkám. Ty jsou k dispozici pro některé obecně významné atributy (např. rodná čísla, IČ, bankovní spojení, ...) Zde je dobré si uvědomit, že obecně známé poučky (dělitelnost rodného rodného čísla jedenácti) mohou mít výjimky, které by v takovýchto případech měly být korektně implementovány. Výhodou je, že řada skriptů či funkcí kontrolujících splnění daného pravidla je znovupoužitelná. V rámci projektu ProEval vzniklo během příslušné aktivity množství skriptů v jazyku R pro ověřování syntaktické správnosti nejrůznějších atributů.

Validace – sémantická správnost

V praxi TA ČR se jedná o srovnávání hodnot daného atributu s relevantními číselníky, seznamy aj., např s databázemi společnosti Bisnode: proti databázi právnických osob byly kontrolovány údaje pocházející z IS Patriot importované do systému ISTA.

Posloupnost kroků při čištění obecných dat v prostředí úloh OSA

1. Vytipování relevantních atributů
2. Profiling
3. Standardizace
4. Deduplikace
5. Syntaktická správnost
6. Sémantická správnost

Metodické podněty k zajišťování datové kvality v informačních systémech TA ČR

1. Před návrhem IS/zadáváním vývoje či návrhem doplňující funkcionality by si měli klíčoví aktéři přečíst tento text či další texty, které se věnují otázkám datové kvality. Cílem je vytvořit atmosféru, kdy otázky zajišťování kvality jsou brány jako přirozená součást vývoje, nikoliv “třešnička na dortu”.
2. Jak při návrhu, tak při provozování IS by měly být zmiňovány otázky kvality dat – návrh systému by měl v sobě již požadavky na příslušnou funkcionalitu, která umožní měřit vybrané aspekty datové kvality a sledovat jejich změny v čase.
3. Uživatelské rozhraní systémů by mělo v sobě zahrnovat práci s typologií problémů manuálního zadávání dat z předchozí části (formulářová pole kontrolovaná příslušnými skripty)
4. V případě, že lze u daného atributu provádět syntaktickou kontrolu, využít ji.
5. Umožnit získávání reportů pro relevantní atributy a v pravidelných intervalech provádět profiling.