Operačný program
Efektívna
verejná správa

Európska únia
Európsky sociálny fond

Formulácia odporúčaní pre zlepšenie dátovej kvality

Ako zlepšiť kvalitu údajov?

Projekt

Zlepšenie využívania údajov vo verejnej správe

Zmluva o dielo č. 321/2018

Výstup č.3

KPMG

ÚRAD PODPREDSEDU VLÁDY SR
PRE INVESTÍCIE
A INFORMATIZÁCIU

Zoznam skratiek

AS-IS	Súčasný východiskový stav
DK	Dátová kvalita
ERP	Enterprise resource planning (systémy plánovania zdrojov)
GDPR	Global data protection regulation (Všeobecné nariadenie na ochranu osobných údajov)
HW	Hardvér
IČO	Identifikačné číslo organizácie
ISVS	Informačný systém verejnej správy
KPI	Key Performance Indicator – kľúčový ukazovateľ výkonnosti
NA	Neaplikovateľné
OVM	Orgán verejnej moci
RA	Register adries
RACI	Matica zodpovednosti (R – zodpovedný vykonávateľ, A – zodpovedný vlastník, C – konzultuje, I – je informovaný)
RFO	Register fyzických osôb
RPO	Register právnických osôb
RÚ	Register úpadcov
SPOC	Single point of contact (centrálny komunikačný uzol)
SQL	Structured query language (štruktúrovaný dotazovací jazyk)
SW	Softvér
TO-BE	Cieľový stav
UPVII	Úrad podpredsedu vlády pre investície a informatizáciu Slovenskej republiky

Obsah

1 Manažérske zhrnutie

Dôveryhodné kvalitné údaje v rámci verejnej správy umožňujú spoľahlivé rozhodovanie, informované nastavovanie politík štátu, podporujú opätovné využívanie údajov a predovšetkým sú nevyhnutné pre kvalitné poskytovanie služieb občanom a podnikateľom.

Jednotlivé organizácie verejnej správy si zároveň čoraz viac uvedomujú výhody zdieľania údajov vo verejnom sektore, ako aj s inými externými organizáciami a verejnosťou. Zdieľanie kvalitných údajov umožňuje:

— včasné a informované rozhodovanie;
— reálne využitie dát;
— spoluprácu medzi verejnou správou a verejnosťou na základe transparentnosti a dôveryhodnosti;
— predchádzať duplikácii zhromažďovania údajov.

Dáta predstavujú aktíva štátu s obrovskou hodnotou. Aby táto ich hodnota mohla byť využitá, verejná správa by mala zdieľať **konzistentný prístup k správe údajov a riadeniu ich kvality**. Podmienkou je nastavenie jasných štandardov kvality dát vo verejnej správe a ich pravidelné meranie, vyhodnocovanie a prijímanie relevantných nápravných opatrení.

V súčasnej dobe v rámci verejnej správy v podstate neexistuje systematické riadenie kvality dát. V rámci základných registrov je síce identifikovaná množina problémov, ale reálny stav dátovej kvality nie je meraný štruktúrovanými testami, na základe ktorých by bolo možné kvalifikovane povedať a zhodnotiť, v akej reálnej kvalite je stav jednotlivých registrov a aká je teda ich reálna vypovedacia schopnosť a praktická použiteľnosť.

Zavedenie účinných postupov v oblasti kvality údajov zvýši hodnotu údajov v rámci štátu ako strategického aktíva. Definovanie jasného súboru požiadaviek, ktoré sa majú uplatňovať na všetky kritické a zdieľané údaje, poskytne pevný základ pre konzistentný prístup k meraniu, komunikácii a k zlepšeniu kvality údajov v rámci verejnej správy.

Obsahom dokumentu **“Ako zlepšiť kvalitu údajov?**” je poskytnúť inštitúciám verejnej správy (i) prehľad najlepšej praxe a skúseností v oblasti riadenia kvality dát a (ii) návrh odporúčaných opatrení pre zlepšenie kvality dát vo verejnej správe.

Navrhnuté odporúčania vychádzajú ako z najlepšej praxe v oblasti riadenia kvality dát tak aj z konkrétnych problémov identifikovaných v rámci uskutočneného merania kvality dát na vybraných registroch verejnej správy (Register právnických osôb, Register adries a Register úpadcov).

2 Úvod do zlepšovania kvality údajov

2.1 Čo to je kvalita dát

Dátová kvalita je súhrnom viacerých parametrov:

— presnosť;
— správnosť;
— kompletnosť;
— unikátnosť;
— aktuálnosť;
— strojová spracovateľnosť;
— referenčná integrita;
— konzistentnosť.

Jednotlivé parametre dátovej kvality vo verejnej správe sú bližšie popísané v dokumente „Metodika merania dátovej kvality vo verejnej správe“, ktorý zároveň obsahuje pre jednotlivé parametre definíciu konkrétnych merateľných ukazovateľov, návrh ich prahových hodnôt a návod na ich výpočet.

Za kvalitné dáta sú považované dáta, ktoré dosahujú požadované prahové hodnoty pre všetky parametre dátovej kvality.

Cieľom zlepšovania dátovej kvality nie je dátová kvalita sama o sebe, ale vyššia kvalita života a lepšie kvalitnejšie služby verejnosti. Tie sú silno prepojené a závislé práve na primeranej dátovej kvalite. Dátová kvalita nie je abstraktný izolovaný cieľ, ale cesta a základňa pre zvyšovanie kvality poskytovaných služieb.

2.2 Definovanie cieľov a prostriedkov pre zlepšovanie kvality dát

Implementácia systematického a udržateľného manažmentu kvality dát vyžaduje od organizácie adresovanie viacerých predpokladov efektívneho riadenia dátovej kvality:

Organizačné zabezpečenie

Procesné zabezpečenie

Technologické zabezpečenie

Dátové štandardy, pravidlá a artefakty

Na základe stupňa adresovania týchto štyroch dimenzií, je možné rozlíšiť niekoľko štádií, v ktorých sa organizácia v oblasti riadenia kvality dát nachádza:

„Nevedomá“	Zdroje nealokované	Bez porozumenia dopadov kvality dát na organizáciu		
„Reaktívna“	Alokované zdroje bez priradenia jasných rolí	Identifikovanie kvality dát ako procesu	Bez špecializovaných nástrojov	Reaktívne opravy na úrovni jednotlivých organizácií
„Proaktívna“	Nastavené metriky úspechu s jasne definovanými rolami	Nastavené procesy s jasne definovanými zodpovednosťami	Aplikované podporné nástroje	Zameranie sa na analýzu príčin a proaktívne objavovanie problémov
„Optimalizovaná“	Vytvorená rola „Chief Data Officera“ a stratégia pre správu dát ako jedného z hlavných aktív spoločnosti	Kvalita dát monitorovaná a pravidelne vyhodnocovaná ako súčasť bežnej prevádzky	Dedikovaná platforma pre riadenie dát	Nastavené biznis pravidlá a dátové i technologické štandardy

Tabuľka 1: Prehľad stavov organizácie pre oblasť dátovej kvality

Pre realizáciu zlepšenia kvality dát organizácie je dôležité nastavenie stratégie pre riadenie dátovej kvality a budovanie relevantnej kompetencie vo všetkých štyroch oblastiach.

Nie je žiadnym tajomstvom, že kvalita údajov je dôležitá, pretože rozhodnutia a špecificky rozhodnutia vo verejnej správe by mali byť založené na reálnych faktoch - bez úplných a presných údajov, na ktoré sa môžeme spoľahnúť, môže byť ohrozená správnosť jednotlivých rozhodnutí.

Zlepšovanie kvality dát je proces, ktorý môže každá organizácia začať realizovať ešte dnes.

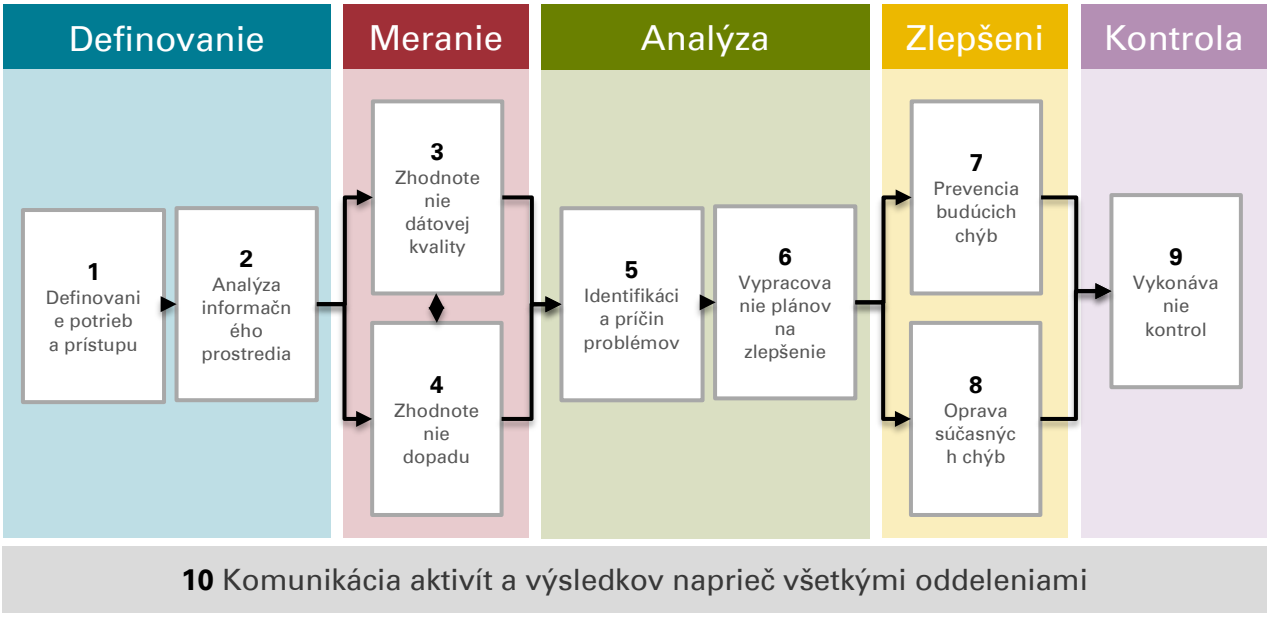
Pred implementáciou akéhokoľvek plánu na zlepšenie kvality údajov musí organizácia pochopiť, kde dnes stojí. Organizácia by mala identifikovať, v akom štádiu riadenia kvality dát sa dnes nachádza a podľa toho primerane zamerať svoje úsilie a zvoliť ďalšie kroky.

Ciele a stratégiu riadenia dátovej kvality je potrebné zadefinovať na začiatku iniciatívy pre zlepšenie kvality dát a jasne v nej určiť ambície, ciele, ako aj prostriedky a nástroje pre oblasť kvality dát.

Tento dokument ma ambíciu poslúžiť organizáciám verejnej správy ako inšpirácia a návod pre jednotlivé aktivity v oblasti zlepšovania kvality dát, ktoré organizácia môže aplikovať v závislosti od jej konkrétnych potrieb.

2.3 Proces zlepšovania kvality dát

Zlepšovanie kvality je kontinuálny proces, ktorý by mala implementovať každá organizácia. Návrhy vzorového desať krokového procesu pre riadenie dátovej kvality vo verejnej správe sú bližšie popísané v dokumente „Metodika merania dátovej kvality vo verejnej správe“, kapitola 2.3. Návrh metodiky pre riadenie, správu a meranie dátovej kvality:



Obrázok 1: 10-krokový proces pre zhodnotenie, zlepšenie a vytvorenie dátovej kvality

Dáta sú obrazom reality. Čím kvalitnejšie dáta máme, tým hodnovernejšie sú informácie a znalosti z nich odvodené. To, čo chceme systematicky zlepšovať, musíme aj nejakým spôsobom merať a vyhodnocovať. Definovanie kvality dát (resp. požadovaných parametrov kvality dát) a objektívne meranie kvality dát je úvodnou fázou v rámci procesu efektívneho riadenia kvality dát.

Zistite, čo „chcete“ od svojich údajov a ktoré parametre sú pre Vás kľúčové. Kvalita údajov znamená niečo iné v rôznych organizáciách. Kvalita údajov je v podstate o tom, že sú vhodné a pripravené na požadovaný účel.

Prvým krokom je tak určenie účelu používania Vašich údajov a zosúladenie Vašej definície kvality s týmto účelom. Odtiaľ môžete odvodiť špecifické KPI, ktoré sú relevantné pre vašu organizáciu, takže budete môcť sledovať výsledky úsilia o zlepšenie kvality údajov. Mať zavedené jasné metriky Vám pomôže nielen posúdiť Vašu efektívnosť, ale tiež Vám pomôže definovať jasnú návratnosť investícií do zlepšovania kvality údajov.

3 Dôvody a príčiny nedostatočnej kvality údajov

Vysokokvalitné údaje sú absolútne najväčšou hnacou silou pre vytváranie prínosov v moderne fungujúcich organizáciách, bez ohľadu na to, či sa jedná o biznis alebo verejnú správu. Dobré údaje môžu spôsobiť zvyšovanie kvality poskytovaných služieb a dobrého postavenia organizácie v očiach verejnosti a klientov. Na druhej strane nízka kvalita údajov spôsobuje pokles výkonnosti organizácie, zvyšovanie kapacitných nárokov na zabezpečovanie procesov ako aj znižovanie transparentnosti rozhodovania a dôvery v organizáciu.

Efektivita, presnosť práce, kvalita rozhodovaní v organizácii je v súčasnosti založená na údajoch. Je veľmi dôležité prijímať správne rozhodnutia o budúcnosti a smerovaní organizácie a uľahčiť tak klientom využívanie služieb ako aj pochopenie ich požiadaviek a širší kontext zasadenia organizácie do prostredia, v ktorom funguje. Ak však existuje vážny problém s kvalitou údajov, s najväčšou pravdepodobnosťou budú prijímané rozhodnutia založené na chybných predpokladoch.

Príklad:

Spoločnosť Gartner odhaduje, že obrovské množstvo podnikových údajov je buď nepresných, neúplných alebo nedostupných, pričom táto zlá kvalita údajov stojí mnoho firiem milióny USD ročne.[1]

Preto je veľmi dôležité vedieť, aké sú hlavné problémy, dôvody a príčiny nízkej dátovej kvality. Pred samotným rozhodovaním a riadením, založeným na dátach a faktoch, je potrebné preskúmanie a vyčistenie údajov a nastavenie procesov na zabezpečenie merania a vyhodnocovania kvality dát v kontinuálnom režime.

3.1 Čo teda znamenajú zlé údaje vo všeobecnosti?

Podľa rôznych organizácií, ktoré sa venujú kvalite údajov, existuje viacero problémových oblastí, ktoré práve spôsobujú nízku alebo nedostatočnú kvalitu dát. V nasledujúcej časti sú definované základné problémy s kvalitou údajov na to, aby sme pochopili, čo sa nemá robiť ak chceme efektívne pracovať v data-driven prostredí.

Duplicitné údaje

S duplicitnými údajmi sa borí každé odvetvie a každá organizácia. Často je to spôsobené ako výsledok nekonzistentných procesov a viacerých systémov, v ktorých sa zaznamenávajú tie isté informácie. V okamihu spojenia rôznych zdrojov pre potreby napr. analýz sa objavia viacnásobné kópie toho istého záznamu a výsledky analýz môžu byť preto zmätočné. Výsledkom môže byť oslovovanie klientov s tou istou otázkou alebo oblasťou duplicitne, čo vedie k zaťažovaniu a nespokojnosti klientov, k strate času a iných zdrojov.

[1] Zdroj: https://www.dataversity.net/bad-data-crippling-data-analytics/

Ak sa narába s viacnásobnými zdrojmi údajov, je práve nekonzistencia ukazovateľom potenciálneho problému s kvalitou údajov. V mnohých prípadoch môžu v databáze existovať viacnásobné záznamy. Duplicitné údaje sú jedným z najväčších problémov, ktoré existujú pre organizácie, ktoré chcú byť data-driven a môžu ich výkonnosť znížiť, podobne aj iné problémy s údajmi.

Nekonzistentné formáty

Nekonzistentnosť údajov je spôsobená zadávaním informácií, ktoré pokrývajú rovnaké oblasti, akurát sú uložené v rôznych formátoch. Systémy, ktoré tieto údaje využívajú, majú problém použiť tieto údaje a tak dochádza k nepresným výsledkom. Príkladom je napr. zadávanie dátumu v rôznych formátoch. Stále je to dátum, ale napr. formát v US a EU je rôzny - DD/MM/YYYY vs. MM/DD/YYYY, čo v prípade analýzy a spájania údajov môže viesť k nekorektným údajom. Rovnako to môže by v prípade telefónnych čísel, kedy sa raz zadajú aj z informáciou o oblasti resp. štátu a v inom prípade sa zadajú bez tejto informácie.

Často sú údaje zlé definované (napr. na úrovni metadát), čo spôsobuje zmätok v ich riadení. Napríklad sú niekedy údaje zaradené do nesprávnej kategórie, čo spôsobí ozajstný „neporiadok“ vo Vašej databáze a skomplikuje prácu s ňou.

Nekompletné informácie

Polia, ktoré nie sú kompletne vyplnené, alebo ostanú úplne prázdne, môžu byť hlavným problémom pre systémy ako CRM alebo automatizované systémy, ako aj algoritmy využívajúce BIG DATA princípy. Príkladom môžu byť nevyplnené položky s PSČ, čo spôsobí nielen problémy pri priamom oslovovaní klientov, ale môže tiež spôsobiť, že kľúčové analytické procesy organizácie budú neefektívne, pretože nebudú mať k dispozícii základné geografické informácie, ktoré pomáhajú zisťovať trendy alebo správne rozhodovať.

Nepresné údaje

Nemá žiadny význam začínať s analýzami alebo zabezpečovať kontakt s klientami na základe zle zadaných (nesprávnych) údajov. Je veľa dôvodov, prečo sa to deje. Od nadiktovania nesprávnych informácií klientom, až po preklepy pri zadávaní údajov alebo vyplnenie nesprávnych polí. Toto môže byť jedným z najväčších problémov, na ktoré je potrebné myslieť. Napr. zadanie zlého čísla poistenia, ktoré však spĺňa pravidlá formátu, nie je ľahko verifikovateľné, ak systém môže kontrolovať len jeho formálnu správnosť.

Nedostatočné riadenie údajov

Ak neexistujú jasne nastavené pravidlá pre všetky údaje, ktoré sa nachádzajú v organizácií, dochádza k rôznym interpretáciám údajov pre rôzne procesy, v ktorých sa údaje spracovávajú. Samotná konzistencia údajov je tak narušená a nie sú definované pravidlá zaznamenávania údajov.

Príliš veľa údajov

Podľa prieskumov až 40% pracovníkov uvádza, že pre správnu prácu disponujú často až priveľa údajmi. Napriek tvrdeniu, že príliš veľa údajov nemôže byť zlá vec, opak býva pravdou. Príliš veľa údajov môže spôsobovať, že sa spotrebuje omnoho viac času „vyhodnocovaním“ nesprávnych a zlých dát, pričom k tým dobrým sa pracovník ani nedostane alebo ich ľahko prehliadne.[2]

Nízka ochrana údajov

Až okolo 20 % ľudí tvrdí, že si nevie predstaviť opätovne spolupracovať s organizáciami, ktoré nedokázali pracovať bezpečne a profesionálne so svojimi údajmi. Obzvlášť v prípade verejnej správy, kde je výskyt osobných citlivých údajov obrovský, je potrebné zaistiť vysoké bezpečnostné štandardy, aby dáta neboli zneužité a bola zabezpečená potrebná ochrana údajov.[3]

Nesprávne údaje

Údaje sa „entropicky“ priemerne „rozpadajú“ rýchlosťou 2,2% za mesiac. Preto sa často stáva, že niektoré z Vašich údajov sú zastarané. Je to obrovský problém, pretože mnoho existujúcich údajov vo Vašom systéme má v sebe tieto chyby.[4]

Nízka obnoviteľnosť údajov

Ľudia všeobecne strávia v priemere okolo 30% svojho času hľadaním správnych údajov, ktoré potrebujú. Čo je však ešte horšie, tak až v priemere 40% prípadoch tieto údaje nenašli na prvom mieste, kde začali hľadať.[5]

3.2 Čo spôsobuje nekvalitu údajov v kontexte verejnej správy?

V kontexte vyššie uvedeného boli identifikované základné problémové oblasti spôsobujúce dátovú nekvalitu v údajoch verejnej správy. Každú z definovaných oblastí je možné zaradiť resp. odvodiť od všeobecných problémov, ktoré sa v oblasti dátovej kvality vyskytujú. Každá identifikovaná oblasť je stručne popísaná v nižšie uvedenej tabuľke. Vyriešením resp. elimináciou definovaných oblastí sa výrazne prispeje k zvýšeniu kvality údajov.

[2] Zdroj: https://www.ringlead.com/blog/7-common-data-quality-issues/
[3] Zdroj: https://www.ringlead.com/blog/7-common-data-quality-issues/
[4] Zdroj: https://www.ringlead.com/blog/7-common-data-quality-issues/
[5] Zdroj: https://www.ringlead.com/blog/7-common-data-quality-issues/

Oblasť	Popis problému	Duplicitné údaje	Nekonzistentné formáty	Nekompletné informácie	Nepresné údaje	Nedostatočné riadenie	Nízka ochrana údajov	Nesprávne údaje	Nízka obnoviteľnosť údajov
Neexistuje Centrálny model údajov	Vzhľadom na fakt, že neexistuje centrálny model údajov, nevedia jednotlivé organizácie s akými údajmi a o kom rozsahu už štát alebo verejná správa prišli do styku a kde sú uložené. Jednotný centrálny model by vyriešil problematiku duplicitne vedených údajov ako aj nekonzistentností v zadávaných údajoch (sú často zadávané v rôznych formátoch a hodnotách).	X	X		X	X			
Objekty evidencie nie sú prepojené	Neprepojením údajov na úrovni objektov evidencie dochádza k potrebe manuálnych zásahov do procesov, čo spôsobuje chyby najmä pri rozhodovaniach alebo pri vyhodnocovaní analýz.		X	X	X	X		X	
Nízky počet referenčných údajov	Vzhľadom na vyššie uvedené, nie je možné vyhlasovať údaje za referenčné, čím by sa zamedzilo duplicitnému zadávaniu údajov, ako aj nesprávnemu zadávaniu údajov. Pri dostatočnom referencovaní by boli referenčné hodnoty vedené len v jednej databáze (registri), z ktorého by sa využívali napr. pre potrebné procesy vo verejnej správe.	X	X	X	X	X	X	X	X
Biznis pravidlá nie sú definované pre objekty evidencie	Vo verejnej správe vo väčšine prípadoch neexistuje formálna, validovaná a schválená dokumentácia biznisových pravidiel, ktorá by určovala kritéria merania dátovej kvality. „Správne“ biznis pravidlá by mali obsahovať aj informácie o vzťahoch medzi jednotlivými atribútmi a na získanie tejto informácie je potrebné poznať architektúry informačných systémov nielen referenčných registrov, ale aj samotných zdrojových systémov.			X	X	X		X	
Pri zbere údajov a zadávaní údajov nie sú kontrolované pravidlá	Jedná sa o základné nastavenia databáz a procesov, ktoré zabezpečujú verifikáciu údajov na vstupe. Problémy to spôsobuje najmä vytváraním nekonzistentností alebo nepresnosti údajov.		X	X	X	X		X	
Historické nepresné údaje	Historické údaje boli zadávané do systémov (ak boli zadávané) bez akýchkoľvek pravidiel a kontrol. Veľa z týchto údajov sa často ani v systéme nenachádzajú. Je preto potrebné tieto údaje jednorazovo naplniť v požadovanej kvalite využívajúc všetky dostupné nástroje na riadenie kvality údajov.				X			X	X
Duplicity v evidencii	Duplicity v evidencii sú väčšinou následkom nedostatočného riadenia údajov alebo absencie nástrojov v oblasti riadenia kvality údajov. Samotné duplicity spôsobujú jednak problémy pri analýzach a rozhodovaní ako aj v prípade vyhodnocovania, či daná databáza môže byť referenčnou alebo nie.	X							

Tabuľka 2: Prehľad problémov a oblastí spôsobujúcich dátovú nekvalitu

4 Najlepšia prax a možnosti pre zlepšenie kvality dát

Analýza postupov a skúsenosti najlepšej praxe je rozdelená do štyroch základných identifikovaných oblasti:


Organizačné zabezpečenie


Procesné zabezpečenie


Technologické zabezpečenie


Dátové štandardy, pravidlá a artefakty

Riadenie kvality dát zahŕňa množinu činností organizačného, procesného, technologického a štandardizačného charakteru.

Hoci hmatateľné prínosy týchto činností nie sú vždy okamžite zrejmé, tieto činnosti sú dôležité na to, aby sa zlepšila hodnota dátového aktíva. Prostredníctvom zlepšenia kvality údajov sa aktíva údajov stávajú viac využiteľnými interne a ak sú zdieľané, tak aj inými organizáciami, čím sa zvyšuje ich hodnota. Zlepšenie kvality dát, ktoré majú nízku kvalitu, opravou problémov alebo automatizovaným spracovaním, zníži náklady na následné spracovanie dát a umožní zamestnancom sústrediť sa skôr na používanie dátového aktíva namiesto strácania času na opravu. Riziko sa možné znížiť aj poskytnutím kvalitnejších údajov, ktoré umožňujú lepšie rozhodovanie.

Pri plánovaní zlepšovacích aktivít je dôležité byť realistický. Ak je základná kvalita údajov extrémne zlá, je možné, že bude potrebné naplánovať a dokončiť základné činnosti (ako napríklad tie, ktoré sú uvedené nižšie) pred implementáciou komplexnejších aktivít. Ak sa zistia konkrétne problémy, týkajúce sa jednotlivých dátových aktív, bude potrebné vyvinúť osobitné aktivity špecificky zamerané na tieto dátové prvky.

4.1 Organizačné zabezpečenie

V súčasnosti má len málo organizácií centralizovanú stratégiu kvality údajov pod jedným vlastníkom. To znamená, že väčšina spoločností vidí veľa rôznych, oddelených a neprepojených stratégií, čo vedie k menej efektívnemu úsiliu o kvalitu údajov.

Centralizáciou vlastníka údajov sa dosahuje stav, kde jedna osoba preberá zodpovednosť/znalosť za kvalitu a normy, týkajúce sa dát.

Role a zodpovednosti

Rôzne role v rámci štruktúry riadenia majú rôzne zodpovednosti. Všetky by mali prispieť ku konsolidovanému a konzistentnému prístupu ku kvalite údajov v rámci organizácie. Je dôležité poznamenať, že role a zodpovednosti rolí úloh môžu byť delegované.

— **Vlastník**

Konečnú zodpovednosť za riadenie dátových aktív ma vedenie organizácie. V praxi môže zodpovedný pracovník (vlastník) delegovať zodpovednosť za informačné aktíva na delegovaného vlastníka, ktorý zasa môže ďalej poveriť výkonom správcu informácií.

— **Delegovaný vlastník**

Delegovaný vlastník je výkonná rola zodpovedná za konkrétne dátové aktíva, pričom zabezpečuje, aby dáta boli presné, aktuálne, chránené, prístupné a podľa možnosti aj zdieľané. Vlastník poskytuje usmernenie, aby sa zabezpečila vhodnosť údajov pre ich daný účel použitia.

Vlastník by mal zabezpečiť, aby sa kvalita údajov zlepšovala a udržiavala na stanovenej úrovni kvality. Vlastníci by mali podporovať vypracovanie plánov riadenia, monitorovanie kvality údajov a zaviazať sa k podpore vykonávania procesov a činností na zlepšenie.

Hoci delegovaný vlastník sa môže rozhodnúť delegovať úlohy na inú výkonnú úroveň alebo na depozitára, celková zodpovednosť im zostáva.

V prostredí verejnej správy bola veľká časť kompetencií delegovaná na rolu Dátového kurátora (pre viac informácií pozrite výstup č.1 Metodika merania dátovej kvality vo verejnej správe, kapitola 2.3.2.4 Rola dátového kurátora v procesoch merania dátovej kvality).

— **Konzumenti informácií**

Kľúčoví aktéri v oblasti kvality údajov sú často aj konzumenti informácií. Konzumenti informácií sú interní zákazníci, ktorí potrebujú transformovať obchodné údaje na operačné, taktické alebo strategické rozhodovacie účely. Ich požiadavky sa musia zohľadniť pri vývoji systému a jeho kontrolných mechanizmov.

Konzumenti informácií by sa mali zúčastňovať na procese zlepšovania kvality údajov, pretože sú často prví, ktorí zistia rozdiely v údajoch, ktoré nie sú pre operatívne orientovaného zamestnanca vždy zrejmé.

4.1.1 Potrebné vzdelanie a kurzy

Cieľovou skupinou vzdelávania v oblasti efektívnej správy údajov a riadenia kvality údajov by mali byť nielen vlastníci a dátový kurátori, ale optimálne všetci zamestnanci, ktorí sa podieľajú na životnom cykle údajov organizácie.

Vzdelávanie zamerané na budovanie kompetencií: Od zamestnancov sa vyžaduje, aby absolvovali odbornú prípravu, ktorá je špecifická pre schopnosti správy údajov konkrétneho dátového aktíva alebo typu údajov. Školenie môže byť vo forme on-line vzdelávania, workshopov alebo mentoringu.

V prípade vlastníkov a dátových kurátorov je najlepšou praxou dedikované vzdelávanie a vzájomné zdieľanie skúseností medzi jednotlivými pracovníkmi v rámci pracovných skupín.

Školenie môže zahŕňať:

— príslušné právne predpisy a normy týkajúce sa údajov;
— dostupné zdroje na pomoc pri pochopení údajov, napr. dátové slovníky;
— obchodné procesy súvisiace s riadením „dátového majetku“, napr. zber, analýza alebo interpretácia údajov;
— postupy pre meranie a vyhodnocovanie kvality údajov.

Spätná väzba: Vlastníci údajov a dátový kurátori by mali dostávať spätnú väzbu, pretože môžu poskytnúť informácie o tom, či sú údaje vhodné na daný účel, či sú užitočné, presné a včasné.

Dôležitým komponentom vzdelávacích aktivít je **vzdelávanie zamerané na zvyšovanie povedomia**: Zamestnanci môžu absolvovať odbornú prípravu v oblasti informovanosti o dôležitosti kvality údajov.

Témy môžu zahŕňať:

— prečo je dôležitá kvalita údajov;
— aké sú použitia údajov vo vašej organizácii;
— dôsledky nízkej kvality údajov.

Tento typ vzdelávania (typicky vo forme on-line vzdelávania alebo prezentácií) by mal byť poskytnutý čo najširšie možnej cieľovej skupine zamestnancov, ktorých činnosti súvisia alebo podporujú hlavné aktivity danej organizácie.

4.2 Zmeny procesov

Implementovať proaktívne procesy

Po získaní správnych ľudí, zlepšenie kvality údajov sa opiera predovšetkým o proaktívne procesy. Práve tieto procesy sú výsledkom úsilia o kontrolu kvality, ktoré sa stane súčasťou každodenných aktivít v rámci organizácie. V opačnom prípade sú problémy, ktoré negatívne ovplyvnia prevádzku alebo rozhodovací proces organizácie, odhalené reaktívnym spôsobom.

Na začiatku správneho nastavenia proaktívnych procesov je analýza účelov využitia údajov, ktorá zabezpečí primeranosť kontrolných procesov vzhľadom k účelu využitia dát. Odtiaľ môžu byť vyvinuté najlepšie postupy riadenia a zlepšovania kvality údajov tak, aby vyhovovali týmto potrebám.

Jednotlivé kľúčové elementy procesov pre zlepšenie kvality údajov sú nasledujúce:

4.2.1 Ako merať dátovú kvalitu

To, čo chceme systematicky zlepšovať, musíme aj primeraným spôsobom vyhodnocovať merať a kvantifikovať. Aby bolo možné objektivizovať súčasný aj plánovaný stav a zároveň čo najobjektívnejšie sledovať trendy a dynamiku vývoja. Meranie kvality je kontinuálny proces.

Viac informácií ako merať dátovú kvalitu je k nahliadnutiu vo výstupe č.1 Metodika merania dátovej kvality vo verejnej správe, kapitola 2.3.1 Komplexný popis procesov pre potreby riadenia a správy dátovej kvality, ktorá vychádza z najlepšej praxe pre riadenie a správu dátovej kvality.

4.2.2 Ako vytvoriť správne biznis pravidlá

Biznis pravidlá predstavujú reprezentáciu toho, čo sú to správne dáta. Biznis pravidlá teda definujú správne dáta na základe existujúcich procesov. Bez dobrej znalosti biznis pravidiel nie je možné efektívne merať ukazovatele dátovej kvality.

Jasne definované biznis pravidlá sú jedným zo základných stavebných kameňov efektívneho nastavenie proaktívnych procesov pre zlepšovanie dátovej kvality.

V dátovej kvalite sa biznis pravidlá uplatňujú na jednotlivých atribútoch dátových entít datasetu.

Biznis pravidlá definuje a spravuje vlastník registra, kde efektívny postup vytvárania biznis pravidiel zahŕňa:

— Východisko v existujúcej legislatíve, interných smerniciach a procesnej dokumentácie.

— Zahrnutie do procesu definovania biznis pravidiel vecných expertov na procesy využívajúce dané dáta.
— Využitie profilácie dát pre identifikovanie biznis pravidiel a ich úvodnú validáciu.
— Sústredenie sa na kľúčové dátové prvky a pri definícii biznis pravidiel umožňuje postupovať inkrementálne.
— Využitie agilného prístupu.

Centrálne riadenie biznis pravidiel

V dynamickom prostredí neustálych legislatívnych zmien a komplexného využívania dát rôznymi organizáciami verejného aj súkromného sektoru, nie je jednoduché efektívne riadiť biznis pravidlá. Centrálne riadenie túto úlohu podstatne zjednodušuje. Jednotlivé pravidlá je potrebné párovať na konkrétne atribúty dátových entít centrálneho dátového modelu verejnej správy. Narastá aj potreba aplikovania rovnakých biznis pravidiel naprieč všetkými informačnými systémami, ktoré pracujú s identickými atribútmi dátových entít v celej verejnej správe. Biznis pravidlá, ktoré striktne vychádzajú z legislatívy, je potrebné pripojiť k relevantným zákonom. Napríklad pravidlá často potrebujú zdôrazniť existenciu verejných referenčných číselníkov, ktoré zjednocujú povolené hodnoty pre zvolené atribúty dátových entít. Tvorenie týchto číselníkov je vedľajší efekt centrálneho riadenia biznis pravidiel a priamo podporuje schopnosť merať dátovú kvalitu. Biznis pravidlá sa neustále vyvíjajú. Vznikajú, zanikajú a menia sa na rôznych miestach v rôznych inštitúciách. Vývoj týchto pravidiel je potrebné monitorovať a zdieľať medzi inštitúciami, ktoré si ich potrebujú osvojiť. Jedna z úloh centrálneho riadenia biznis pravidiel je zabezpečiť túto koordináciu.

4.2.3 Ako využiť referenčné údaje

Referenčné údaje predstavujú dôležitý nástroj pre riadenie dátovej kvality v rámci celej verejnej správy. Za každý objekt evidencie, ktorý je vyhlásený ako referenčný, zodpovedá jedna inštitúcia, ktorá by mala garantovať jeho kvalitu a dostupnosť v informačnom prostredí verejnej správy – cez Centrálnu integračnú platformu (IS CSRÚ). Zoznam referenčných údajov postupne rastie a jeho základ tvorí Register fyzických osôb (RFO) a Register právnických osôb (RPO). Každá osoba tak má jasný kmeňový záznam, s jedinečným identifikátorom, ktorý môžu používať ostatné informačné systémy. Ďalším dôležitým referenčným registrom bude Register adries (RA).

Referenčné údaje je potrebné zaviesť do správy kmeňových záznamov v rámci inštitúcie:

— Prvým krokom je párovanie referenčných údajov a následné vyčistenie údajov.
— Druhým krokom je nastavenie synchronizácie medzi referenčným registrom a informačným systémom, aby sa referenčné údaje nemenili v systéme, ale boli aktualizované pri zmene v zdrojovom systéme (sú možné aj iné módy správy kmeňových záznamov, tento je však odporúčaný). Informačný systém je potrebné upraviť spôsobom, aby dátové objekty boli jasne určené identifikátorom, čo zjednoduší prácu a umožní mať presnejšie údaje. Odporúča sa tiež využívať „push“

notifikácie o zmene referenčných údajov. Napríklad, ak sa zmení trvalé bydlisko občana, inštitúcie môžu automaticky aktualizovať svoje adresy.

— Tretím krokom je úprava elektronických služieb a formulárov takým spôsobom, aby nevyžadovali zadávanie údajov, ktoré sú už referenčné. Zníži sa tým administratívna záťaž (formulár bude jednoduchší) a eliminuje sa riziko chýb.

— Mnohé inštitúcie, ktoré sú v priamom kontakte s klientami verejnej správy (napríklad Sociálna poisťovňa) majú aktuálnejšie údaje ako referenčné registre. Je preto potrebné nastaviť proces, ako informovať o chybných resp. neaktuálnych údajoch a umožniť tak referenčnému registru využiť tento vstup pre zvýšenie svojej kvality.

Pre podporu týchto krokov sú už teraz k dispozícii služby v rámci Centrálnej integračnej platformy (IS CSRÚ).

4.2.4 Ako nastaviť proces zberu, aby nevznikali chybné dáta

Podľa odporúčania, proces zberu údajov by mal obsahovať nasledované prvky:

— Jasne určiť biznis pravidlá pre dátové objekty, ktoré sú predmetom zberu dát.

— Používať štandardné dátové prvky (v Centrálnom modeli verejnej správy). Ak štandardný dátový prvok neexistuje, odporúča sa požiadať o jeho štandardizáciu a evidenciu v Centrálnom dátovom modeli.

— Nezbierať údaje, ktoré sú už definované ako referenčné resp. s nimi pracuje iná inštitúcia verejnej správy (zjednodušenie formulárov).

— Používať jedinečné identifikátory na identifikáciu dátových objektov.

— Zaviesť automatizované kontroly voči konzistencii dát a nastaveným biznis pravidlám.

4.2.5 Ako identifikovať chybné údaje

Prvým krokom pri zlepšovaní kvality údajov je odhalenie chýb v údajoch prostredníctvom profilovania údajov, čo je proces analýzy množiny údajov na ich správnosť, úplnosť, jedinečnosť, konzistentnosť a primeranosť.

Profilovanie údajov

Profilovanie dát je špecifický druh analýzy dát, ktorý sa používa na objavovanie a charakterizáciu dôležitých vlastností súborov údajov. Profilovanie dát je zároveň využiteľné pre odhaľovania chýb a anomálií v údajoch. Umožňuje organizáciám získať základný prehľad o údajoch, ako aj správne identifikovať problémy.

Profilovanie zahŕňa aj kontrolu obsahu údajov prostredníctvom profilu dát alebo percentuálneho rozdelenia hodnôt. Výsledky profilovania sa dajú následne porovnať s očakávaniami, alebo môžu poskytnúť základ, na ktorom sa budú ďalej stavať poznatky o údajoch.

4.2.6 Ako vyčistiť chybné údaje

Čistenie údajov je proces opravy údajov po identifikovaní záznamov obsahujúcich nesprávne údaje. Jedným z príkladov je oprava údajov, ktoré nespĺňajú požiadavky formátu alebo štandardu, ako napríklad adresa alebo dátum v nesprávnom formáte.

Čistenie dát je proces zahrňujúci:

— **Špecifikáciu pracovného postupu**: Detekcia a odstránenie anomálií sa vykonáva sledom operácií. Aby sa dosiahol správny pracovný postup, musia sa starostlivo zvážiť predovšetkým príčiny anomálií a chýb v údajoch a správne identifikovať zdrojové systémy pre realizáciu vlastného čistenia údajov.

— **Realizáciu pracovného postupu**: V tejto fáze sa pracovný postup vykoná po dokončení jeho špecifikácie a overení jeho správnosti a efektívnosti.

— **Následnú kontrolu a „post-processing“**: Po ukončení pracovného postupu sa skontrolujú výsledky, aby sa overila správnosť dát. Ak je to možné, údaje, ktoré nebolo možné opraviť automaticky, sa manuálne opravia. Výsledkom je nový cyklus v procese čistenia dát, kde sú údaje opäť auditované, je aktualizovaný pracovný postup, tak aby sa ďalšie čistenie údajov umožnilo už automatickým spracovaním.

Dôležitým aspektom pri korekcii a čistení údajov je zabezpečiť, aby sa problém neopakoval prostredníctvom chýb alebo postupov pri zadávaní údajov (prevencia) a aby sa opravené údaje opravili aj pri zdroji, teda v pôvodnej databáze. Ak údaje nie sú opravené v pôvodnom vstupnom bode alebo zdroji, existuje vysoké riziko, že sa budú musieť znova opraviť.

Čistenie údajov je časovo náročný a nákladný proces a čistenie všetkých údajov zvyčajne nie je ani nákladovo odôvodnené ani praktické. Na druhej strane, neočistenie údajov je rovnako neprijateľné. Preto je dôležité starostlivo analyzovať, ktoré dátové prvky sú kritické, dôležité alebo naopak nepodstatné a tomuto prispôsobiť plán čistenia údajov. Nie je vždy nevyhnutné vyčistiť všetky údaje a zároveň nie je nevyhnutné to robiť naraz.

4.2.7 Publikovanie otvorených údajov

Publikovanie otvorených údajov môže mať výrazne pozitívny charakter na kvalitu údajov, vďaka samotnému procesu pre publikovanie otvorených údajov a vďaka systému spätnej väzby.

Proces správnej publikácie otvorených údajov si vyžaduje aplikovanie postupov, ktoré vo svojom dôsledku vedú k odhaľovaniu chýb a zvýšeniu konzistentnosti dátového zdroja. Publikácií otvorených údajov vo verejnej správe sa venuje napríklad príručka: Metodická príručka pre povinné osoby – Ako zverejňovať otvorené dáta[6], ktorú pripravil NASES ako prevádzkovateľ portálu data.gov.sk. Proces publikácie otvorených údajov je možné chápať v dvoch fázach:

[6] https://www.slovensko.sk/_img/CMS4/Navody/Metodicka_prirucka_pre_povinne_osoby.pdf

— **Fáza 1**: Spustenie pravidelného publikovania otvorených údajov.
— **Fáza 2**: Pravidelné publikovanie otvorených údajov.

Fáza 1: Spustenie pravidelného publikovania otvorených údajov

1. Oboznámenie sa s konceptom otvorených údajov a základnými princípmi: víziou NKIVS je, aby všetky údaje, s výnimkou citlivých údajov a utajovaných skutočností boli zverejňované vo formáte otvorených údajov, pričom pre osobné údaje sa použije vhodná metóda anonymizácie.
2. Oboznámenie sa s legislatívnymi povinnosťami vo vzťahu k publikovaniu otvorených údajov.
3. Organizačné zabezpečenie – malo by byť v súlade s celkovým organizačným zabezpečením manažmentu údajov v organizácii. Pre malé organizácie sa odporúča vymenovať Dátového kurátora (zodpovednú osobu) a pre väčšie organizácie vytvoriť rezortnú Dátovú kanceláriu.
4. Príprava úvodného zoznamu datasetov.

Fáza 2: Pravidelné publikovanie otvorených údajov

5. Aktualizácia plánu pre publikovanie otvorených údajov.
6. Spracovanie údajov na publikovanie: znamená najmä prípravu prístupu k dátovým zdrojom, vyhotovenie transformačných procedúr, popis datasetov (pridanie metadát) a priradenie licencií. Pri návrhu transformačných procedúr a definícií metadát je kľúčové správne generovanie referencovateľných identifikátorov (ideálne v súlade s Centrálnym dátovým modelom). Spracovanie údajov na publikovanie je kľúčovým krokom procesu, ktorý má vplyv na výslednú dátovú kvalitu.
7. Publikovanie údajov na portály data.gov.sk. Ak sú údaje publikované ako otvorené údaje, verejnosť s nimi môže pracovať a zadávať spätnú väzbu v prípade, že identifikuje nezrovnalosti alebo nedostatky. Dôležité je, aby organizácia nastavila riadny proces pre spracovanie takejto spätnej väzby a mala vypracované procedúry, ako sťažnosti na dátovú kvalitu riešiť. Práve transparentnosť dát je kľúčová pre čistenie dát na základe takzvaného „crowdsourcingu“.
8. Aktualizácia publikovaných údajov, ktorá sa deje z dôvodu nových údajov, alebo pri oprave a doplňaní katalógových záznamov pri zistených nezrovnalostiach a oprava identifikovaných nedostatkov alebo chýb v samotnom datasete. Údaje sa tak aktualizujú na dataset, ktorý má vyššiu kvalitu, ako ten predchádzajúci.
9. Archivovanie publikovaných údajov.
10. Riadenie kvality otvorených údajov prebieha počas celej fázy 2.

4.2.8 Služba Moje dáta a kvalita údajov

Ďalším spôsobom, ako je možné získať spätnú väzbu na kvalitu údajov je využiť službu Moje dáta. Táto služba (postupne sa zavádza), umožní občanom a podnikateľom získať cez API prístup k údajom, ktoré o nich verejná správa eviduje. Občan tak môže prezerať svoje údaje a keďže sa ho týkajú, je veľká pravdepodobnosť, že identifikuje prípadné

nezrovnalosti. Služba bude nastavená tak, že umožní občanovi reklamovať chybný objekt evidencie a tiež požiadať o aktualizáciu a upresnenie údajov.

Z pohľadu inštitúcie verejnej správy je dôležité (potom, ako sa zapojí do služby Moje dáta) nastaviť proces pre riešenie týchto sťažností i žiadostí a využiť tiež autorizáciu občana. V princípe je potrebné nastaviť, ktoré atribúty môže meniť priamo občan a kde je potrebné, aby prebehlo rozhodovanie o zmene dát na strane inštitúcie.

Službu Moje dáta môže využiť inštitúcia aj ako nástroj na zber nových dát (formou ankety) alebo pri potrebnej autorizácie zmeny (opravy údajov) od občana.

4.2.9 Transparentné zverejňovanie KPI dátovej kvality

Najlepšou praxou pri kontinuálnom monitorovaní je implementovať „dashboard“ dátovej kvality poskytujúci prehľad do akej miery jednotlivé dátové objekty jednotlivých organizačných jednotiek naplňujú jednotlivé parametre kvality:

Parametre dátovej kvality
Presnosť
Kompletnosť
Aktuálnosť
Unikátnosť
Referenčná integrita
Strojová spracovateľnosť
Konzistentnosť
Správnosť

Tabuľka 3: Zoznam parametrov dátovej kvality

Pre viac informácií o parametroch dátovej kvality pozrite výstup č.1 Metodika merania dátovej kvality vo verejnej správe, kapitola 3. Špecifikácie parametrov dátovej kvality vo verejnej správe.

Dôležitým aspektom zverejňovanie stavu dátovej kvality je zobrazenie vývoja v čase, s cieľom ukázať dynamiku vývoja parametrov dátovej kvality.

Riadenie dát a zlepšovanie ich kvality by sa mali stať natívnou súčasťou operatívnych činností každej organizácie a je dobrou praxou zverejňovať a pozitívne hodnotiť tie organizácie, resp. organizačné zložky, ktoré dosahujú najlepšie výsledky, či už v absolútnom vyjadrení alebo dosahujú najlepšie relatívne zlepšenie stavu v čase.

4.3 Technológie pre dátovú kvalitu

Vzhľadom na objem dát, ktoré dnes organizácie zhromažďujú, kľúčovú rolu v efektívnom riadení kvality dát majú technológie a to aj v prípade, ak časť čistenia dát je potrebné adresovať manuálne.

Využitie jednotných technológií

Implementácia jednotných technologických riešení v oblasti kvality údajov v rámci celej spoločnosti môže zefektívniť a zjednotiť úsilie o zlepšenie údajov. Výsledkom bude presadzovanie konzistentných dátových štandardov v celej organizácii. To znamená, že údaje sa budú dať lepšie použiť pre rozhodovanie.

4.3.1 Aké nástroje máme k dispozícii

Celá koncepcia a metodika merania monitorovania a vyhodnocovania dátovej kvality je technologicky nezávislá a platformovo agnostická. Dá sa aplikovať pre všetky dostupné štandardné riešenia pre riadenie a meranie dátovej kvality.

Pre konkrétne dáta a meranie ich kvality je potrebné použiť niektoré konkrétne technologické riešenie. Tieto riešenia sú štandardne poskytované a dostupné aj v móde Platforma ako služba (PaaS).

Riešenie pre dátovú kvalitu umožňuje používateľom profilovať, čistiť, maskovať a pripravovať dáta a zároveň monitorovať dátovú kvalitu v čase, bez ohľadu na ich formát alebo veľkosť. Umožňuje opätovné použitie pravidiel dátovej kvality v rôznych integráciách. Pravidlá zahŕňajú de-duplikáciu, validáciu, štandardizáciu a obohatenie na základe strojového učenia.

V prípade verejnej správy si rezorty, ktoré nevlastnia technologické riešenie pre podporu procesov zlepšovania dátovej kvality, môžu pre tento účel využiť dostupné cloudové služby poskytované Úradom podpredsedu vlády pre investície a informatizáciu.

Poskytované služby Master data manažment a Cloudové služby pre stotožnenie údajov s referenčnými údajmi, sú určené predovšetkým pre dátových kurátorov a dátových analytikov v oblasti riadenia dátovej kvality a bližšie sú popísané v katalógu cloudových služieb (http://www.informatizacia.sk/poskytovanie-sluzieb-vladneho-cloudu/22858s)

Poskytované služby využívajú najmä platformu Talend. „Magic quadrant“ pre nástroj dátovej kvality 2019 od Gartner Group potvrdzuje výbornú pozíciu tohto riešenia medzi porovnateľnými produktami .

Využívanie spoločnej centrálnej platformy v rámci celej organizácie je dobru praxou, ktorá podporuje zdieľanie a budovanie interných skúsenosti a zručnosti.

4.3.2 Využitie AI pre čistenie údajov

Čistenie údajov je časovo náročný a nákladný proces, ktorý využíva manuálne prostriedky aj štandardné počítačové programy. Efektivitu týchto postupov by mohli, vďaka svojmu vzdelávaciemu potenciálu, zvýšiť aj prostriedky umelej inteligencie (AI).

Rýchlejší proces

Využitie AI by mohlo urýchliť celkový proces určovania a extrahovania problematických dát. Lokalizácia dátových chýb sa môže ukázať ako značne časovo náročná. Práve AI však môže potenciálne pracovať rýchlejšie vďaka svojmu "samoučeniu" - môže sa vzdelávať v tom, ako byť efektívnejší.

Využívanie vzorov dát

Dáta vytvárajú vzory. Štatistická analýza skúma tieto vzorce. Čím zložitejšie sú však vzory, tým ťažšie sa štandardnými postupmi lokalizujú. Schopnosť čítať a analyzovať údaje v spojení so samoučením je ďalším spôsobom pridanej hodnoty AI pri čistení údajov - bez schopnosti strojového učenia, program nemusí byť schopný objaviť všetky anomálie vo vzoroch.

Návrh na vyčistenie údajov

Identifikácia chýb je prvým krokom. Ďalším krokom je zadanie správnych údajov, ktoré nahradia nepresnosti. Objavovanie chyby nie je rovnaká úloha ako zadávanie správnych hodnôt. V tomto prípade nie je možné predpokladať, že AI bude riešením pre každý typ dátového problému. AI môže postupne skutočne poskytnúť riešenie, ktoré „vyčistí“ dáta s minimálnym zásahom človeka. Takými prípadmi sú napríklad:

— Deduplikácia dát (odstrániť duplicitné kópie opakovaných údajov).
— Prepojenie záznamov (záznamy, ktoré odkazujú na tú istú entitu naprieč rôznymi zdrojmi).
— Normalizácia dát (konvertovanie dát s viac ako jednou reprezentáciou do štandardného formátu).

Je malo pravdepodobné, že komplexný problém s dátovou kvalitou, je možné plne adresovať len riešeniami strojového učenia. Nástroje AI však už dnes môžu zásadne zefektívniť celý proces čistenia dát.

4.4 Dátové štandardy, pravidlá a artefakty

Nasadenie udržateľných procesov zlepšovania kvality údajov vyžaduje implementáciu viacerých kľúčových predpokladov:

— definícia biznis pravidiel;
— definícia dátových štandardov;
— definícia metadát;

— definícia dátových modelov.

Najlepšou praxou je zadefinovanie týchto základných dátových artefaktov (pričom ich určenie môže byť inkrementálny a agilný proces) už v úvodnej fáze aktivít pre zlepšovanie údajov. Tieto dátové artefakty samozrejme majú priamy dopad na požadované zvýšenie kvality dát, ich jednoznačná definícia hneď v úvode môže rádovo zefektívniť a skvalitniť následný proces zlepšovania kvality dát.

Biznis pravidlá

Biznis pravidlá predstavujú reprezentáciu toho, čo sú to správne dáta. Biznis pravidlá teda definujú správne dáta na základe existujúcich procesov - bez dobrej znalosti biznis pravidiel nie je možné efektívne merať ukazovatele dátovej kvality.

V dátovej kvalite sa biznis pravidlá uplatňujú na jednotlivých atribútoch dátových entít datasetu.

Dátové štandardy

Dátové štandardy sú dokumentované definície reprezentácie, formátu, štruktúry, označovania, postupov manipulácie, používania a správy údajov.

Dátové štandardy by mali obsahovať aj zásady a postupy, ako sú prevádzkové postupy, postupy kontroly zmien, postupy riešenia sporov údajov a štandardy formátov dokumentácie.

Normy pre pomenovanie údajov a skratky zabezpečujú konzistentnosť dát naprieč organizáciou. Doporučené je použitie overených medzinárodných štandardov, pre všetky definované prípady použitia.

Dobrou praxou je publikovanie štandardného zoznamu skratiek pre celú spoločnosť, ktorý obsahuje odvetvovo špecifické a organizačne špecifické skratky.

Metadáta

Metadáta sú popisné kontextové informácie. Sú to dáta o dátach. Pre účel podpory procesov v oblasti zlepšovania kvality dát, sú relevantné predovšetkým metadáta zahrňujúce:

— účel údajov;
— spôsoby vytvárania údajov;
— čas vytvorenia/zmeny;
— autor údajov;
— použité normy;
— zdroj údajov;
— proces použitý na vytvorenie údajov.

Dobrou praxou je vytvoriť štandardy alebo usmernenia, ktoré určujú, kto zachytáva vybraté komponenty metadát a ako, kedy a kde ich zachytiť. Úložisko metadát by malo byť nastavené tak, aby podporovalo štandardy pre správu a používanie metadát.

Dátové modelovanie

Existuje rozdiel medzi logickým modelovaním a fyzikálnym modelovaním dát. Kvalita údajov sa musí riešiť v oboch modeloch. Okrem toho samotné dátové modely musia spĺňať štandardy kvality modelovania údajov, pokiaľ ide o dátové politiky a pravidlá modelovania, ako je napríklad dodržiavanie konvencií pomenovania, konzistentné používanie typov údajov, dátových domén a podobne.

Za účelom nájdenia nadbytočných a nekonzistentných údajov, je logické modelovanie entít a vzťahov s úplnou normalizáciou údajov stále najúčinnejšou technikou, pretože ide o techniku analýzy, ktorá zahŕňa identifikáciu, racionalizáciu a štandardizáciu údajov prostredníctvom metadát.

5 Riešenie problémov (príručka pre zlepšenie dátovej kvality)

Najzásadnejšie problémy, ktoré boli z pohľadu dátovej kvality identifikované, sú uvedené v nasledujúcej tabuľke:

ID	Problémy v oblasti zlepšenia dátovej kvality
1	Chýbajúce hodnoty
2	Duplicitné záznamy
3	Chybné hodnoty (preklepy)
4	Neaktuálne hodnoty
5	Nekonzistentné formáty

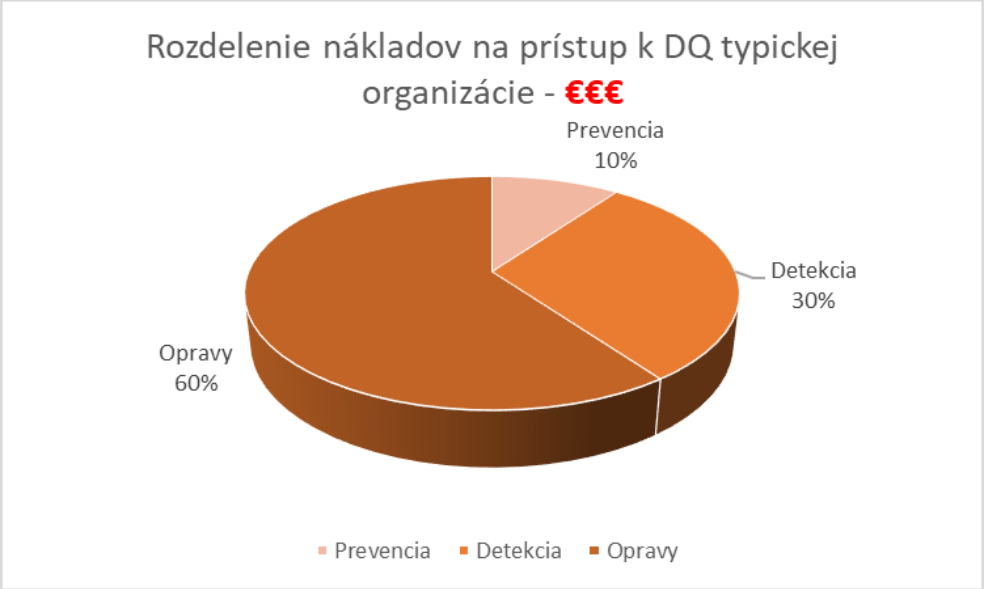
Tabuľka 4: Prehľad problémov v oblasti zlepšenia dátovej kvality

Riešenie problémov v dátovej kvalite, týkajúce sa vyššie uvedeného zoznamu, je možné vykonať prostredníctvom prevencie alebo terapie.

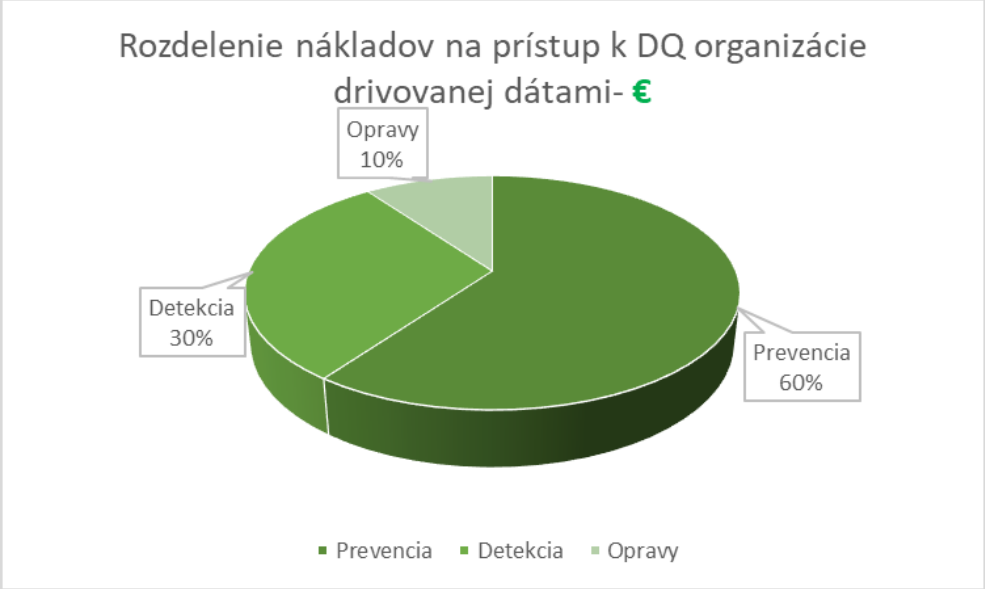
Existuje postupnosť udalostí pri určovaní a optimalizácií dátovej kvality – prevencia, detekcia, opravy. V skratke možno popísať túto sekvenciu nasledovne:

— **Prevencia** – znamená nevpustenie zlých údajov do databáz organizácie.
— **Detekcia** – jedná sa o zavádzanie proaktívneho prístupu k vyhľadávaniu už zlých údajov v systémoch.
— **Opravy** – predstavujú odstraňovanie alebo opravy zlých údajov v systéme.

Nedostatok pozornosti venovanej kvalite a presnosti údajov môže mať závažné následky pre prácu nadväzujúcich systémov. Je to predovšetkým preto, že hľadanie chýb v predradených systémoch (generujúcich údaje) je veľmi finančne aj časovo náročné. Na nasledujúcom grafe je rozdelenie nákladov pre dva typy organizácií v oblasti zabezpečenia dátovej kvality.



Obrázok 2: Štruktúry nákladov v riadení dátovej kvality sústredená na opravu



Obrázok 3: Štruktúra nákladov v riadení dátovej kvality sústredená na prevenciu

Je potrebné zdôrazniť, že práve náklady na opravy údajov, sú spojené s potrebou ľudského zásahu. Je to práve ta najnákladnejšia alternatíva.

5.1 Prevencia

Prevencia patrí medzi hlavné prístupy v riadení kvality údajov, pričom jej základným posolstvom je zabezpečiť kombináciou opatrení, aby sa do databázy nedostali chybné údaje.

Online validácia dát pri vstupe

Často sa stáva, že práve pri vstupe údajov do databáz organizácií sú zadané, alebo prebrané zlé údaje, ktoré následne spôsobujú problémy pri práci systémov alebo znehodnocujú kvalitu vytváraných databáz. Existuje množstvo preventívnych opatrení, ktorých aplikáciou je možné práve tieto nedostatky minimalizovať. V nasledujúcej časti kapitoly sú uvedené najbežnejšie opatrenia, ktoré sa využívajú v rámci prevencie údajov:

5.1.1 Opatrenie prevencie: Kontroly voči biznis pravidlám

Základným nástrojom na prevenciu nekvalitných údajov je nastavenie biznis pravidiel a overovanie ich dodržiavania ešte pred vstupom do databázy. Biznis pravidlá technického charakteru (ako napr. kontrola formátu zadanej hodnoty, alebo kontrola povinného záznamu) by mali byť podporované technologických riešením, ktoré je pre danú databázu optimálne. Zároveň je vhodné, aby bola podporovaná aj kontrola dodržiavania vecných biznis pravidiel (ako napr. každé konanie musí mať priradený súd a daný súd má mať pridelenú adresu súdu).

5.1.2 Opatrenie prevencie: Kontroly voči referenčným hodnotám

V tomto prípade sa jedná o zabezpečenie referencovania údajov na referenčné registre, ktoré je možné realizovať prostredníctvom už existujúcich nástrojov ako sú IS CRSU. Jedná sa o to, že prijímané údaje, ktorých hodnoty sa nachádzajú v referenčných registroch (ako napr. RFO, RPO a pod.) sa pred vstupom do databázy referencujú na tieto registre. V prípade, ak nastane chyba, bude táto označená a zapísaná do databázy s pôvodným, ako aj s referenčným údajom.

5.1.3 Opatrenie prevencie: Prevencia kontaminácie údajov

Zdroje kontaminácie dát v dôsledku chýb pri zadávaní údajov je možné eliminovať alebo výrazne znížiť pomocou techník kontroly kvality. Jednou z veľmi účinných stratégií je, aby údaje boli nezávisle od seba zadané dvomi pracovníkmi a potom na základe dohody overené počítačom. Táto prax je bežná v profesionálnych organizáciách zaoberajúcimi sa dátami, ako aj v iných odvetviach služieb.

5.1.4 Opatrenie prevencie: Nastavenie filtrov na odhalenie „nelegálnych“ dát

Jedná sa o nastavenie takých filtrov alebo kombinácie premenných, ktoré odhalia údaje, ktoré sú doslova nemožné vzhľadom na zadávanú skutočnosť. Ide často o počítačový program, ktorý jednoducho skontroluje zadávaný súbor a cez kompletný zoznam

obmedzení s premenlivými hodnotami. Tento program následne vytvorí zoznam s podrobnosťami každého porušenia.

5.1.5 Opatrenie prevencie: Detekcia údajov mimo systém štandardného modelu

Ide o detekcia údajov mimo systém alebo teda mimo rozsah štandardného modelu. Je súčasťou procesu kontroly predpokladov štatistických modelov (proces, ktorý by mal byť integrovaný a konzistentný s akoukoľvek formálnou analýzou údajov). Avšak samotná eliminácia extrémnych hodnôt nemusí byť súčasťou hodnotenia dátovej kvality aj vzhľadom na fakt, že extrémne hodnoty sú často súčasťou života.

5.2 Terapia

Terapia v oblasti kvality údajov znamená nasadzovanie takých prostriedkov, ktoré pomôžu odstrániť už vzniknutý problém. Ako bolo vyššie uvedené je vhodnejšie investovať do nástrojov prevencie ako následne do nástrojov, ktoré riešia už vzniknuté problémy. Je však potrebné zdôrazniť, že databázy údajov už vznikali často v nekonzistentnosti so zásadami dátovej kvality. A tak sa stalo, že údaje sú chybné a vykazujú problémy.

Preto je potrebné v rámci terapie realizovať sekvenciu z dátovej kvality, detekcie a opravy.

Základom detekcie a následných opráv je:

— nadefinovanie biznis pravidiel pre požadované údaje;
— spustenie profilácie údajov;
— identifikáciu problémových oblastí resp. chybných údajov;
— vykonanie nápravných opatrení.

Prvým krokom pri zlepšovaní kvality údajov je odhalenie chýb v údajoch prostredníctvom profilovania dát – dátová archeológia / detekcia chýb – čo je proces analýzy údajov pre správnosť, úplnosť, jedinečnosť, konzistentnosť a primeranosť.

Akonáhle je známy rozsah „špinavých údajov“, je potrebné začať proces zlepšovania kvality dát práve čistením - opravy. Samotné čistenie je však časovo a nákladovo náročný proces. Častokrát čistenie všetkých údajov nie je praktické ani nutné.

Preto je dôležité starostlivo analyzovať zdrojové dáta a klasifikovať jednotlivé dátové prvky ako kritické, dôležité alebo nepodstatné pre organizáciu a procesy, v ktorých sa údaje využívajú. Potom je dôležité zameranie na očistenie všetkých kritických dátových prvkov a ak je to časovo opodstatnené, aj vyčistenie čo najviac dôležitých dátových prvkov. Zanedbateľné dátové prvky zostanú nezmenené. Inými slovami, nie je nutné vyčistiť všetky údaje a hlavne nemusia byť vyčistené naraz.

Z pohľadu predchádzajúcich analýz a meraní boli identifikované nasledovné problémy:

5.2.1 Riešenie problému: Chýbajúce hodnoty

Chýbajúce hodnoty nemusia byť problém v prípade, ak sú to hodnoty, ktoré nie sú nevyhnutné pre kompletnosť záznamu. V takomto prípade je však otázne, či tieto hodnoty je potrebné v databázach uvádzať. V prípade, ak sú chýbajúce hodnoty podstatné, existujú rôzne techniky ako ich doplniť:

— Identifikovanie referenčného registra, v ktorom sa údaj nachádza a príprava procedúry na doplnenie chýbajúce údaju.
— Využitie imputačných techník ako napríklad:
  - imputácia využívajúca medián alebo priemer;
  - imputácia s použitím najčastejšej hodnoty alebo nulových / konštantných hodnôt;
  - využitie algoritmu „najbližšieho suseda“;
  - imputácia použitím multivariačnej imputácie pomocou reťazenej rovnice;
  - imputácia využívajúca „Deep Learning“.

5.2.2 Riešenie problému: Duplicitné záznamy

Duplicitné záznamy predstavujú problém najmä pre nastavenie jednoznačnej komunikácie ale aj pre vyhodnocovanie a analýzy, ktoré sa nad údajmi realizujú. Množstvo duplicitných záznamov, s ktorými sa stretávame, patria do dvoch odlišných typov:

Neunikátne kľúče – jedná sa napr. o dva záznamy v tej istej tabuľke, ktoré majú rovnaký kód alebo kľúč, ale môžu mať rôzne hodnoty a významy. Toto sa môže stať pri pomiešaní údajov, alebo v prípade, ak údaje pochádzajú z nedatabázových zdrojov, ako sú textové súbory, (csv súbory z importu csv) alebo súbory programu Excel a pod.

Duplicitný význam – ide o častejší problém a často aj náročnejší na vyriešenie. Tieto druhy duplikátov sú často najškodlivejšie pre dobrú analýzu. Aj keď existujú nástroje na definovanie duplicít (v oblasti mien a adries), často je potrebný práve zásah človeka (ideálne skúseného odborníka), ktorý dokáže odhaliť správne označenie duplicity. Pre tieto potreby je nevyhnutné aktualizovať mapu duplicít a zaznamenávať ich vyriešenie.

V kontexte vyššie uvedeného je v rámci odstraňovania daného problému možné:

— Zlúčenie záznamov – predstavuje komparáciu záznamov a následne zlúčenie údajov z oboch záznamov do jedného, ak sa jedná o doplňujúce informácie, ktoré jednotlivé záznamy obsahujú.
— Odstránenie duplicitného záznamu – v prípade, ak sa jedná o totožný duplicitný záznam, je potrebné jeden z nich odstrániť.

5.2.3 Riešenie problému: Chybné hodnoty

Chybné hodnoty v údajoch sú často dôsledkom nezavedeného manažmentu dát a neexistujúceho koncepčného riešenia na sledovanie dátovej kvality. Práve chybné hodnoty sú jedným z adeptov na odstraňovanie preventívnymi prístupmi, ako je dodržiavanie biznis pravidiel alebo referencovanie na zdrojové registre. Napriek tomu nie je ľahké zabezpečiť kvalitu údajov a správnych hodnôt v databázach v plnej miere a preto je potrebné vysporiadať sa aj s týmito problémami vo fáze opráv.

Použitie referenčných dát

Ak došlo k problémom v dátach napr. na základe preklepov a chybné údaje boli identifikované, vhodným riešením je využitie referenčných údajov (ak tieto existujú) na „prepísanie“ chybných dát.

Referenčné dáta sú horšie ako dáta v systéme

Môže nastať situácia, kde kvalita údajov v referenčnom registri je horšia ako kvalita dát v zdrojovom systéme. V tomto prípade je potrebné pri identifikácií nezrovnalostí kontaktovať vlastníka referenčného registra a definovať spôsob, ako sa daný údaj v referenčnom registri opraví.

Automatizované opravy nepovoľuje legislatíva

V súvislosti s vyššie uvedeným vzniká niekedy problém, kde automatické opravy v referenčných registroch nie je možné realizovať z titulu legislatívnych obmedzení a teda úprava údajov je často procesne aj časovo zdĺhavá. Opravu musíme vykonať v zdrojovom registri, ktorý údaj pre referenčný register poskytuje.

5.2.4 Riešenie problému: Neaktuálne hodnoty

Neaktuálne hodnoty v databázach sú často spôsobené buď:

— pasivitou majiteľov údajov (napr. v prípade preťahovania nenahlási obyvateľ zmenu trvalého bydliska, alebo neposkytne poisťovniam zmenené údaje), alebo

— zlým nastavením procesu zmeny a zápisu údajov do databázy (zmena statusu konania alebo zmena sudcu a pod.).

Základným nástrojom na eliminovanie týchto problémov je definovanie jasných postupov, ako majú byť konkrétne údaje aktualizované.

— Ak ide o údaje, ktoré sú automaticky aktualizované v zdrojových a referenčných systémoch, tieto dáta by sa mali automaticky preniesť do všetkých databáz, ktoré s týmito údajmi pracujú.

— Ak ide o zmenu údajov v procese, ktorý dataset popisuje alebo dokumentuje, musí byť jasne určené, kto a akým spôsobom aktualizuje údaj. Tento postup zároveň zabezpečuje prístup do príslušnej databázy.

Zavedením týchto opatrení sa však neodstránia všetky chybné a neaktuálne hodnoty. V tomto prípade by bolo optimálnym riešením len manuálna kontrola a stotožňovanie údajov s realitou v momente, keď sa daný údaj má použiť.

5.2.5 Riešenie problému: Nekonzistentné formáty

Nekonzistentné formáty často spôsobujú problémy pri spájaní údajov alebo pri vyhodnocovaní rôznych aspektov, ktoré údaje prinášajú. Rovnako vznikajú problémy aj pri využívaní údajov v systémoch, kedy je tvar formátu aj jeho technické prevedenie v poriadku, ale obsahovo je formát rôzny. Napr. v prípade dátumu môže nastať problém – MM/DD/RR vs. DD/MM/RR. Prehodenie dňa a mesiaca môže v systémoch generovať zásadné chyby – napr. ak systém validuje počet mesiacov a zrazu je v hodnote číslo 15. V prípade konaní, pre ktoré sú dátumy podstatné, to môže mať značné negatívne dopady.

Nekonzistentnosť formátov je rovnako kandidát na posilnenie prevencie a to najmä nastavením exaktného formátu pre danú položku. Ide o vstupnú validačnú kontrolu jednotlivých položiek.

V prípade, ak sa identifikuje nesprávny formát vo fáze detekcie, je možné ho opraviť nasledovnými spôsobmi:

— nastavením automatizovaného algoritmu zmeny formátov (najčastejšie sa vyskytujúcich) na správny formát a následnú validáciu formátov prostredníctvom validačných nástrojov resp. algoritmov;
— „ručným“ prepísaním formátu do správneho tvaru a nastavením parametrov pre daný formát.

6 Odporúčania a návrh plánu realizácie opatrení

V tejto časti sú definované odporúčania a opatrenia, ako prispieť k zvýšeniu dátovej kvality ako na lokálnej (rezortnej), tak aj na centrálnej úrovni (napr. UPVII). Plán opatrení je rozdelený na 4 základné oblasti:

— dátové štandardy, pravidlá a artefakty;
— organizačné zabezpečenie;
— procesné zabezpečenie;
— technológie pre dátovú kvalitu.

Pre každé opatrenie bola definovaná priorita opatrenia zo škály 1 až 3, pričom interpretácia priorít je nasledovná:

— **Priorita 1** – nevyhnutné opatrenie, ktorého zavedenie by nemalo presiahnuť 1 rok.
— **Priorita 2** – dôležité opatrenie, nadväzujúce na opatrenia priority 1, ktorého realizáciou sa výrazne zvýši dátová kvalita a malo by byť zavedené do 2 rokov.
— **Priorita 3** – „Nice to Have“ opatrenie, ktorého zavedenie nie je nevyhnutné, ale pomôže k celkovému kontextu riadenia kvality údajov a bez jeho implementácie sa stav v budúcnosti môže zhoršiť.

Zároveň bola pre každé opatrenie stanovená úroveň zodpovednosti za jeho prípadnú implementáciu:

— **Lokálne** – reprezentuje rezort alebo organizáciu.
— **Centrálne** – reprezentuje opatrenia na úrovni štátu, ktoré by malo pripraviť UPVII prostredníctvom Dátovej kancelárie.

6.1 Dátové štandardy, pravidlá a artefakty

Opatrenie	Priorita	Zodpovednosť	
		Lokálna	Centrálna
Zadefinovanie jasných biznis pravidiel pre riadenie dátovej kvality vrátane väzby biznis pravidiel na legislatívu upravujúcu dotknuté údaje.	1	☑	
Implementácia biznis pravidiel a pravidiel pre meranie dátovej kvality (vrátane nastavenia jednoznačných KPIs a ich prahových hodnôt ako aj stanovenie, pre ktoré atribúty / záznamy je potrebné merať dátovú kvalitu) v organizáciách a nastavenie ich riadenia.	1	☑	
Definovanie presného formátu pre každý zadávaný údaj, aby bola zabezpečená konzistencia všetkých zdrojov, ktoré sú v organizácií využívané.	1	☑	
Vypracovanie jednotného dátového slovníka pre verejnú správu ako komponent centrálneho dátového modelu.	1		☑
Vytvorenie centrálneho dátového modelu verejnej správy a definovanie kompetencií a organizačného zabezpečenia dátového modelu.	1		☑

Opatrenie	Priorita	Zodpovednosť	
		Lokálna	Centrálna
Vytvorenie lokálnych (rezortných dátových modelov) a zaviesť osobnú zodpovednosť pre túto oblasť centrálne aj na každom rezorte, plus jednotné pravidlá pre tvorbu a udržiavanie dátových modelov a dátovej architektúry.	1	☑	
Zadefinovanie jasných pravidiel pre stotožňovanie a referencovanie údajov medzi registrami a možnosti prepájať informácie medzi registrami tak, aby poskytovali potrebné informácie pre analýzy a riadenie kvality údajov.	2		☑
Zabezpečenie referencovania získavaných údajov organizáciou alebo údajov vedených v databázach organizácií na existujúce referenčné registre.	1	☑	
Zabezpečenie súladu súčasného stavu s existujúcimi štandardami dátovej kvality.	2	☑	
Definovanie oblastí dátovej kvality pre zabezpečenie líderstva v tejto oblasti na úrovni EU.	3		☑
Vypracovanie a udržiavanie dátovej dokumentácie každého rezortu a zabezpečenie jej správy – je potrebné definovať a zaviesť jednotné dátové štandardy, jednotné formy dátových modelov, jednotné dokumentácie k metadátam, jednotné štruktúry a formy biznis pravidiel.	3	☑	☑

Tabuľka 5: Prehľad opatrení pre dátovú kvalitu z oblasti dátových štandardov, pravidiel a artefaktov

6.2 Organizačné zabezpečenie

Opatrenie	Priorita	Zodpovednosť	
		Lokálna	Centrálna
Definovanie jasných kompetencií, právomocí a zodpovednosti pre pozície dátových kurátorov.	1	☑	☑
Posilnenie kapacít pre organizačné zabezpečenie riadenia dátovej kvality vo forme dátových kurátorov.	1	☑	
Zabezpečenie školení pre oblasť riadenia a správy dátovej kvality v rezorte a vybudovanie „call centra“ pre zabezpečenie poradenstva v oblasti dátovej kvality.	1		☑
Posilnenie kompetencie Dátovej kancelárie v oblasti dátovej kvality tak, aby v súčasnosti existujúce pravidlá vedela kontrolovať a vyvodzovať prípadné konsekvencie.	2		☑
Striktné zadefinovanie a prípadne preformulovanie pravidiel zadávania vstupných údajov do rezortných systémov tak, aby bol v čo najväčšej možnej miere eliminovaný ľudský faktor. Znamená to technické a kapacitné posilnenie na úrovni vstupov.	2	☑	
Nastavenie pravidiel pre opravy identifikovaných nezrovnalostí v záznamoch tak, aby bol proces čo najkratší a aby bol realizovateľný v momente vzniku alebo odhalenia konzistentnosti v údajoch (viď. opravy záznamov v obchodnom registri, ktoré sú zdrojom pre RPO – chyba sa zistí v RPO).	2		☑
Jasné nastavenie pravidiel hodnotenia a odmeňovania pre zodpovedné osoby za dátovú kvalitu reflektujúc aj proaktivitu, ktorá môže viesť k poukázaniu na neakceptovateľný stav v oblasti dátovej kvality tej ktorej organizácie – pravidlo, „ak si vedel o nekvalite a nepovedal si o nej budeš sankcionovaný“.	3		☑

Opatrenie	Priorita	Zodpovednosť	
		Lokálna	Centrálna
Implementovanie zmeny v oblasti dátovej kultúry a interpretácie údajov, ktoré má organizácia vo „vlastníctve“ vrátane zavedenia biznis metadát.	3	☑	

Tabuľka 6: Prehľad opatrení pre dátovú kvalitu z oblasti organizačného zabezpečenia

6.3 Procesné zabezpečenie

Opatrenie	Priorita	Zodpovednosť	
		Lokálna	Centrálna
Definovanie stratégie pre dátovú kvalitu (vrátane definovania jasných KPIs, ktoré budú následne adoptované organizáciami) a neustále monitorovať jej zavedenie v praxi a v pravidelných intervaloch vyhodnocovať a aktualizovať.	1	☑	☑
Zvedene povinnosti vykonávania pravidelného merania a vyhodnocovania dátovej kvality so zverejňovaním výsledkov vo forme open data.	2		☑
Prehodnotenie a nastavenie pravidiel opráv chybných záznamov tak, aby bol proces čo najkratší s jasne nastavenými KPIs vedúcimi k motivácii na jeho výkon (napr. v prípade súdov - prideliť túto zodpovednosť na jeden konkrétny súd, ktorý bude riadne vyškolený aj motivovaný na opravu takýchto chybných alebo neúplných záznamov).	2	☑	
Jasné definovanie povinných a nepovinných údajov pri jednotlivých záznamoch tak, aby bolo pri hodnotení údajov vidieť aká je štatistika pre povinné a nepovinné údaje.	1	☑	
Zabezpečenie lepšej osvety pre zadávanie nepovinných údajov a vysvetlenie pridanej hodnoty zadania nepovinného údaja do systému.	3	☑	
Zavedenie jasných pravidiel pre aktualizáciu záznamov a ich atribútov, ako aj implementácia procesov monitoringu a dodržiavania pravidiel.	2	☑	
Nastavenie jednoduchých procedúr pre transformovanie údajov do foriem využiteľných pre realizáciu procesov podporujúcich dátovú kvalitu alebo procesov tvorby reportov a analýz.	1	☑	
Nastavenie procesov prípravy údajov na analytické spracovanie (forma, štruktúra potrebná pre analýzy systémoch ako napr. R, SAS, Python, ...).	2	☑	
Zabezpečenie jednoduchšieho prístupu k datasetom potrebným pre meranie a riadenie dátovej kvality (napr. prístupy do všetkých informačných systémom by mali byť na úrovni READ kedykoľvek. Export dát z databázy by mal byť pre zodpovedné osoby ľahšie možný a najlepšie vo formáte SQL).	1	☑	

Tabuľka 7: Prehľad opatrení pre dátovú kvalitu z oblasti procesného zabezpečenia

6.4 Technológie pre dátovú kvalitu

Opatrenie	Priorita	Zodpovednosť	
		Lokálna	Centrálna
Zabezpečenie príslušnej technologickej podpory na riadenie kvality údajov pre všetky definované oblasti vychádzajúce zo stratégie riadenia kvality údajov.	1		☑
Implementácia a zabezpečenie nástroja na riadenie dátovej kvality (v súčasnosti existuje platforma Talend, ktorá je zabezpečená na úrovni PaaS služby v štátnom cloude).	1	☑	
Nastavenie technologických pravidiel pre prácu s databázami (jasne definované pravidlá správy údajov, ktoré sa dajú podporiť existujúcimi technológiami).	1	☑	
Aplikovanie systémov AI pre technologické riadenie kvality údajov.	3	☑	
Nastavenie procesov profilovania a analyzovania údajov priamo v technologickom riešení.	2	☑	
Nastavenie proaktívnych kontrol pre zdrojové systémy – kontrola bude prebiehať už pri vstupe údajov na základe definovaných technologických a procesných požiadaviek.	2	☑	
Zabezpečenie technologickej podpory pre definované biznis pravidlá pri vstupe údajov do systému.	2	☑	
Nastavenie pravidiel, aby záznamy nemohli byť skompletované, ak neobsahujú všetky potrebné náležitosti (povinné polia) a ak nebudú dodržané definované pravidlá zápisu údajov do príslušných polí.	1	☑	
Implementácia nástroja / algoritmov na definovanie duplicity záznamov pri vstupe údajov do systému alebo na podporu procesu deduplikácie záznamov.	2	☑	
Identifikovanie možností a následného zabezpečenia technologickej podpory pre elimináciu faktora „človek“ vo všetkých procesoch a fázach práce s údajmi.	3	☑	

Tabuľka 8: Prehľad opatrení pre dátovú kvalitu z oblasti technológie

6.5 Návrh harmonogramu zavedenia opatrení a výstupov pre centrálnu úroveň

Oblasť	Opatrenie	Priorita	Výstup	Termín
Dátové štandardy	Vypracovanie jednotného dátového slovníka pre verejnú správu ako komponent centrálneho dátového modelu.	1	Vypracovaný dátový slovník	31.10.2019
Dátové štandardy	Vytvorenie centrálneho dátového modelu verejnej správy a definovanie kompetencií a organizačného zabezpečenia dátového modelu.	1	Vypracovaný komplexný dátový model	31.1.2020
Dátové štandardy	Zadefinovanie jasných pravidiel pre stotožňovanie a referencovanie údajov medzi registrami a možnosti prepájať informácie medzi registrami tak, aby poskytovali potrebné informácie pre analýzy a riadenie kvality údajov.	2	Pripravená legislatíva	30.6.2020
Dátové štandardy	Definovanie oblasti dátovej kvality pre zabezpečenie líderstva v tejto oblasti na úrovni EU.	3	Aktualizovaná stratégia dátovej kvality	30.6.2020
Dátové štandardy	Vypracovanie a udržiavanie dátovej dokumentácie každého rezortu a zabezpečenie jej správy – je potrebné definovať a zaviesť jednotné dátové štandardy, jednotné formy dátových modelov, jednotné dokumentácie k metadátam, jednotné štruktúry a formy biznis pravidiel.	3	Vypracovaná dokumentácia	31.12.2020
Organizácia	Definovanie jasných kompetencií, právomocí a zodpovedností pre pozície dátových kurátorov.	1	Kompetenčný model	30.11.2019
Organizácia	Zabezpečenie školení pre oblasť riadenia a správy dátovej kvality v rezorte a vybudovanie „call centra“ pre zlepšenie poradenstva v oblasti dátovej kvality.	1	Zrealizované školenia pre dátových kurátorov	31.12.2019
Organizácia	Posilnenie kompetencií Dátovej kancelárie v oblasti dátovej kvality tak, aby v súčasnosti existujúce pravidlá vedela kontrolovať a vyvodzovať prípadné následky.	2	Pripravená legislatíva	31.3.2020
Organizácia	Nastavenie pravidiel pre opravy identifikovaných nezrovnalostí v záznamoch tak, aby bol proces čo najkratší a aby bol realizovateľný v momente vzniku alebo odhalenia konzistentnosti v údajoch (viď. Opravy záznamov v obchodnom registri, ktoré sú zdrojom pre RPO – chyba sa zistí v RPO).	2	Pripravené legislatíva	30.6.2020
Organizácia	Jasné nastavenie pravidiel hodnotenia a odmeňovania pre zodpovedné osoby za dátovú kvalitu reflektujúc aj pro-aktivitu, ktorá môže viesť k poukázaniu na neakceptovateľný stav v oblasti dátovej kvality tej ktorej organizácie – pravidlo, „ak si vedel o nekvalite a nepovedal si o nej, tak budeš sankcionovaný“.	3	Pripravené pravidlá hodnotenia	31.12.2020

Oblasť	Opatrenie	Priorita	Výstup	Termín
Proces	Definovanie stratégie pre dátovú kvalitu (vrátane definovania jasných KPIs, ktoré budú nasledovne adoptované organizáciami).	1	Stratégia dátovej kvality	30.11.2019
Proces	Neustále monitorovanie dátovej kvality, jej zavedenie v praxi a v pravidelných intervaloch vyhodnocovať a aktualizovať.	2	Report z monitorovania	**Kontinuálne**
Proces	Zavedenie povinností vykonávania pravidelného merania a vyhodnocovania dátovej kvality so zverejňovaním výsledkov vo forme open data.	2	Pripravená legislatíva	31.3.2020
Technológie	Zabezpečenie príslušnej technologickej podpory na riadenie kvality údajov pre všetky definované oblasti vychádzajúce zo stratégie riadenia kvality údajov.	1	Licencie na riadenie kvality údajov k dispozícii	31.12.2019

Tabuľka 9: Návrh harmonogramu zavedenia opatrení a výstupov pre centrálnu úroveň

6.6 Návrh harmonogramu zavedenia opatrení a výstupov pre organizácie

Nasledujúca tabuľka popisuje odporúčania pre časový plán implementácie opatrení, vyjadrený v trvaní mesiacov od spustenia prác.

Oblasť	Opatrenie	Priorita	3 [mes.]	6 [mes.]	9 [mes.]	12 [mes.]	15 [mes.]	18 [mes.]	24 [mes.]	27 [mes.]	30 [mes.]
Dátové štandardy	Definovanie presného formátu pre každý zadávaný údajov, aby bola zabezpečená konzistencia všetkých zdrojov, ktoré sú v organizácií využívané.	1	☑								
Dátové štandardy	Vytvorenie lokálnych (rezortných dátových modelov) a zaviesť osobnú zodpovednosť pre túto oblasť centrálne aj na každom rezorte, plus jednotné pravidlá pre tvorbu a udržiavanie dátových modelov a dátovej architektúry.	1			☑						
Dátové štandardy	Zabezpečenie referencovania získavaných údajov organizáciou alebo údajov vedených v databázach organizácií na existujúce referenčné registre.	1				☑					

Oblasť	Opatrenie	Priorita	3 [mes.]	6 [mes.]	9 [mes.]	12 [mes.]	15 [mes.]	18 [mes.]	24 [mes.]	27 [mes.]	30 [mes.]
Dátové štandardy	Zadefinovanie jasných biznis pravidiel pre riadenie dátovej kvality vrátane väzby biznis pravidiel na legislatívu upravujúcu dotknuté údaje.	1		☑							
Dátové štandardy	Implementácia biznis pravidiel a pravidiel pre meranie dátovej kvality (vrátane nastavenia jednoznačných KPIs a ich prahových hodnôt ako aj stanovenie, pre ktoré atribúty / záznamy je potrebné merať dátovú kvalitu) v organizáciách a nastavenie ich riadenia.	1				☑					
Organizácia	Posilnenie kapacít pre organizačné zabezpečenie riadenia dátovej kvality vo forme dátových kurátorov.	1		☑							
Proces	Zabezpečenie merania dátovej kvality.	1					☑	☑	☑	☑	☑
Proces	Jasné definovanie povinných a nepovinných údajov pri jednotlivých záznamoch tak, aby bolo pri hodnotení údajov vidieť aká je štatistika pre povinné a nepovinné údaje.	1		☑							
Proces	Nastavenie jednoduchých procedúr pre transformovanie údajov do foriem využiteľných pre realizáciu procesov podporujúcich dátovú kvalitu alebo procesov tvorby reportov a analýz.	1			☑						
Proces	Zabezpečenie jednoduchšieho prístupu k datasetom potrebným pre meranie a riadenie dátovej kvality (napr. Prístupy do všetkých informačných systémom by mali byť na úrovni READ kedykoľvek. Export dát z databázy by mal byť pre zodpovedné osoby ľahšie možný a najlepšie vo formáte SQL).	1			☑						
Technológie	Implementácia a zabezpečenie nástroja na riadenie dátovej kvality (v súčasnosti existuje platforma Talend, ktorá je zabezpečená na úrovni PaaS služby v štátnom cloude).	1			☑						
Technológie	Nastavenie technologických pravidiel pre prácu s databázami (jasne definované pravidlá správy údajov, ktoré sa dajú podporiť existujúcimi technológiami).	1			☑						

Oblasť	Opatrenie	Priorita	3 [mes.]	6 [mes.]	9 [mes.]	12 [mes.]	15 [mes.]	18 [mes.]	24 [mes.]	27 [mes.]	30 [mes.]
Technológie	Nastavenie pravidiel, aby záznamy nemohli byť skompletované, ak neobsahujú všetky potrebné náležitosti (povinné polia) a ak nebudú dodržané definované pravidlá zápisu údajov do príslušných polí.	1			☑						
Dátové štandardy	Zabezpečenie súladu súčasného stavu s existujúcimi štandardami dátovej kvality.	2					☑				
Dátové štandardy	Vypracovanie a udržiavanie dátovej dokumentácie každého rezortu a zabezpečenie jej správy – je potrebné definovať a zaviesť jednotné dátové štandardy, jednotné formy dátových modelov, jednotné dokumentácie k metadátam, jednotné štruktúry a formy biznis pravidiel.	2						☑			
Organizácia	Striktné zadefinovanie a prípadne preformulovanie pravidiel zadávania vstupných údajov do rezortných systémov tak, aby bol v čo najväčšej možnej miere eliminovaný ľudský faktor. Znamená to technické a kapacitné posilnenie na úrovni vstupov.	2					☑				
Proces	Prehodnotenie a nastavenie pravidiel opráv chybných záznamov tak, aby bol proces čo najkratší s jasne nastavenými KPIs vedúcimi k motivácii na jeho výkon (napr. v prípade súdov - prideliť túto zodpovednosť na jeden konkrétny súd, ktorý bude riadne vyškolený aj motivovaný na opravu takýchto chybných alebo neúplných záznamov).	2					☑				
Proces	Zavedenie jasných pravidiel pre aktualizáciu záznamov a ich atribútov, ako aj implementácia procesov monitoringu a dodržiavania pravidiel.	2						☑			
Proces	Nastavenie procesov prípravy údajov na analytické spracovanie (forma, štruktúra potrebná pre analýzy systémoch ako napr. R, SAS, Python, ...).	2						☑			
Technológie	Nastavenie procesov profilovania a analyzovania údajov priamo v technologickom riešení.	2						☑			
Technológie	Nastavenie proaktívnych kontrol pre zdrojové systémy – kontrola bude prebiehať už pri vstupe údajov na základe definovaných technologických a procesných požiadaviek.	2						☑			
Technológie	Zabezpečenie technologickej podpory pre definované biznis pravidlá pri vstupe údajov do systému.	2						☑			

Oblasť	Opatrenie	Priorita	3 [mes.]	6 [mes.]	9 [mes.]	12 [mes.]	15 [mes.]	18 [mes.]	24 [mes.]	27 [mes.]	30 [mes.]
Technológie	Implementácia nástroja / algoritmov na definovanie duplicity záznamov pri vstupe údajov do systému alebo na podporu procesu deduplikácie záznamov.	2							☑		
Organizácia	Implementovanie zmeny v oblasti dátovej kultúry a interpretácie údajov, ktoré má organizácia vo „vlastníctve“ vrátane zavedenia biznis metadát.	3								☑	
Proces	Zabezpečenie lepšej osvety pre zadávanie nepovinných údajov a vysvetlenie pridanej hodnoty zadania nepovinného údaja do systému.	3								☑	
Technológie	Aplikovanie systémov AI pre technologické riadenie kvality údajov.	3								☑	
Technológie	Identifikovanie možností a následného zabezpečenia technologickej podpory pre elimináciu faktora „človek“ vo všetkých procesoch a fázach práce s údajmi.	3								☑	

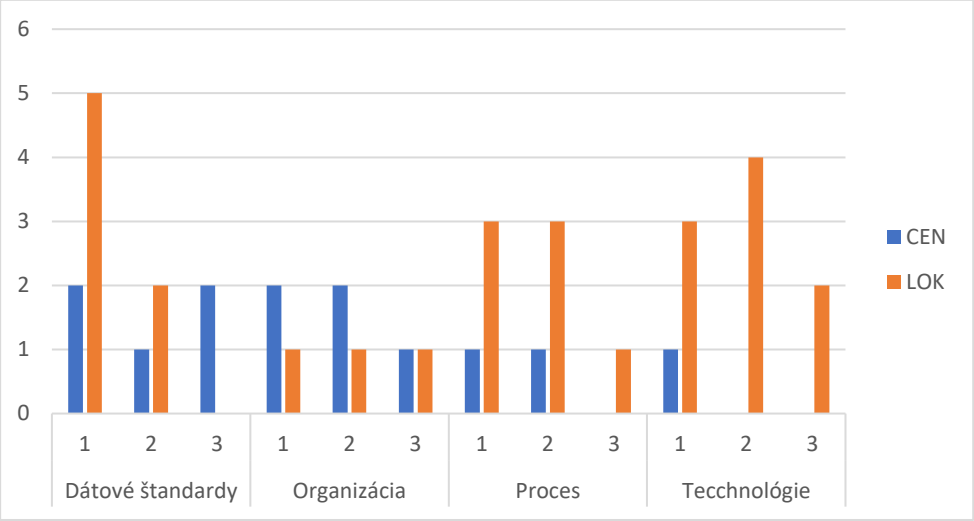
Tabuľka 10: Návrh harmonogramu zavedenia opatrení a výstupov pre organizácie

6.7 Prehľad a štatistika opatrení

V nasledujúcej tabuľke a grafe je uvedený prehľad opatrení (súhrn počtov) podľa priorít, oblastí a zodpovednosti.

Priorita / Oblasť	Centrálne	Lokálne	Spolu
Priorita 1	6	12	18
Dátové štandardy	2	5	7
Organizácia	2	1	3
Proces	1	3	4
Technológie	1	3	4
Priorita 2	4	10	14
Dátové štandardy	1	2	3
Organizácia	2	1	3
Proces	1	3	4
Technológie	NA	4	4
Priorita 3	3	4	7
Dátové štandardy	2		2
Organizácia	1	1	2
Proces	NA	1	1
Technológie	NA	2	2
Spolu	**13**	**25**	**39**

Tabuľka 11: Prehľad štatistiky navrhnutých opatrení



Obrázok 4: Grafické zobrazenie štatistiky prehľadu navrhnutých opatrení

7 Zoznamy

7.1 Zoznam tabuliek

7.2 Zoznam obrázkov