Operačný program
Efektívna
verejná správa

Európska únia
Európsky sociálny fond

MINISTERSTVO
INVESTÍCIÍ, REGIONÁLNEHO ROZVOJA
A INFORMATIZÁCIE
SLOVENSKEJ REPUBLIKY

Výstup č. 1.2.2

Služby dátovej kvality

Zmluva o dielo č. 445/2022

Projekt:

Zlepšenie využívania údajov vo verejnej správe

ITMS kód projektu:

314011S979

KPMG

Tento projekt je podporený z Európskeho sociálneho fondu.

Kontrola a schválenie dokumentu

História verzií

Verzia	Autor	Dátum	Revízia

Revízia

	Recenzent	Dátum revízie
1		
2		
3		
4		
5		

Schválenia

	Schvaľovateľ	Podpis	Dátum schválenia
1			
2			
3			
4			
5			

Tento projekt je podporený z Európskeho sociálneho fondu.

Slovník

CMÚ	Centrálny model údajov
CDE	Critical Data Elements, kritické dátové prvky sú údaje, ktoré sú buď životne dôležité pre rozhodovanie v organizácii alebo sa považujú za vysoko citlivé.
Dataset	Množina údajov, množina dát, niekedy aj register
Dátový slovník	Zoznam dátových pravidiel pre všetky atribúty OE integrovaných v Talend v IS CSRÚ.
DataOps	Holistický prístup k správe údajov, ktorý presahuje technológiu a zameriava sa na kombináciu agilných metodológií, automatizácie a spolupráce medzi profesionálmi v oblasti údajov s cieľom zlepšiť kvalitu, rýchlosť a pridanú hodnotu činností súvisiacich s údajmi.
JSON-LD	Odľahčený formát Linked Data
Kampaň	Kampaň je hlavnou pracovnou jednotkou v prostredí Talend, v ktorej sa definujú všetky požadované konfiguračné prostriedky a realizujú sa činnosti na úpravu dát a zlepšovanie dátovej kvality.
MIRRI	Ministerstvo investícií, regionálneho rozvoja a informatizácie SR
OVM	Orgán verejnej moci
RA	Register adries
RDF	RDF (Resource Description Framework) - štandard výmeny dát na webe.
RPO	Register právnických osôb
SPARQL	Dopytovací jazyk (z angl. SPARQL Protocol and RDF Query Language)
Talend Data Stewardship konzola (DSC)	Talend Data Stewardship Console (DSC) je webové rozhranie pre správu dát, ktoré je poskytované v rámci Talend MDM Web User Interface ako doplnok alebo ponúkané ako samostatná aplikácia.
XML	Rozšíriteľný značkovací jazyk (z angl. eXtensible Markup Language), ktorý bol vyvinutý a štandardizovaný konzorciom W3C (World Wide Web Consortium) ako pokračovanie jazyka SGML a zovšeobecnenie jazyka

HTML. Umožňuje jednoduché vytváranie konkrétnych značkovacích jazykov na rôzne účely a široké spektrum rôznych typov údajov.

Tento projekt je podporený z Európskeho sociálneho fondu.

Obsah

1 Úvod a zhrnutie

Účelom tohto dokumentu je vytvoriť prehľad dostupných služieb dátovej kvality pripravovaných v rámci projektov Dátovej integrácie (DI) a Centrálna integračná platforma (CIP). Služby dátovej kvality sú popísané v tejto štruktúre:

- Charakteristika
- Aplikovanie
- Funkcie
- Použitie a
- Výstupy

Ďalšia časť dokumentu sa venuje odporúčania, ako by mali byť vytvorené používateľské príručky, čo je dôležité zohľadniť pri ich tvorbe a čo by mali obsahovať.

V nasledujúcej časti sa nachádza zhodnotenie prínosov pripravovaného sprístupňovania služieb dátovej kvality cez PaaS a výhod, ktoré prináša riešenie OpenMetadata pre budúce rozvíjanie tejto platformy. Zároveň identifikuje oblasti, na ktoré sa je možné zamerať pre rozvoj tejto platformy.

Posledná časť tohto dokumentu sa venuje automatizácii v oblasti dátovej kvality a riadenia údajov a za využitia strojového učenia a umelej inteligencie. Nachádzajú sa v nej odporúčania, ako k tejto problematike pristúpiť a správne ju rozvíjať, zároveň obsahuje aj inšpiratívny príklad z praxe.

Tento projekt je podporený z Európskeho sociálneho fondu.

2 Definícia služieb dátovej kvality

Služby dátovej kvality, rozvíjané v rámci projektov Dátovej integrácie (DI) a Centrálnej integračnej platformy (CIP) sú v dobe písania tohto dokumentu ešte rôznych štádiách vývoja. Niektoré služby dátovej kvality sú už dokončené a postupne ich jednotliví konzumenti, ktorí boli integrovaní, môžu začať využívať. V dôsledku rôznej realizačnej fázy jednotlivých služieb dátovej kvality existuje pri niektorých z nich obmedzené množstvo dokumentácie, z ktorej je možné pre nasledujúce kapitoly čerpať a teda popis niektorých služieb dátovej kvality nemusí v plnom rozsahu reflektovať ich konečný stav po ukončení projektov, ktorých výstupom sú tieto služby.

V súčasnosti sú dostupné alebo sa pripravujú tieto služby dátovej kvality, ktoré budú v podkapitolách nižšie bližšie opísané:

- DQ1 – Základná analýza dátovej kvality
- DQ2 – Validácia voči objektu evidencie (OE)
- DQ3 – Asistované čistenie
- DQ4 – Automatické stotožnenie
- DQ5 – Asistované stotožnenie
- DQ6 - 9 – Transformácia dátového súboru
- DQ10 – Vlastná služba dátovej kvality podľa špecifických potrieb konkrétneho OVM
- Nástroj na profilovanie dátových súborov OpenMetadata

2.1 Služby dátovej kvality projektu DI

V rámci projektu Dátovej integrácie (DI) bol rozvoj služieb dátovej kvality zameraný najmä na vytvorenie komplexného prostredia pre riadenie kvality dát obsahujúce nástroje pre:

- Tvorbu reportov dátovej kvality a stanovovanie biznis pravidiel na sledovanie kvality a čistenie dát
- Analýzu a čistenie dát ich štandardizáciou, elimináciou duplicít
- Monitoring a reportovanie stavu a vývoja zmien pri zlepšovaní kvality dát
- Sledovanie dátovej kvality cez preddefinované alebo vyvinuté reporty dátovej kvality cez riadené GUI prostredie

Na základe takto zadefinovaných funkcionalít vznikli služby dátovej kvality DQ1 až DQ5, ktoré sú bližšie opísané v kapitolách nižšie. Zároveň v rámci projektu vnikla aj služba dátovej kvality pre transformáciu dát, ktorá je rozčlenená do služieb dátovej kvality DQ6 až DQ9. Na rozdiel od ostatných služieb dátovej kvality, kde DQ1 až DQ5 je možné vykonať samostatne podľa potreby, pre úspešné vykonanie transformácie nie je možné vykonať DQ6 až DQ9 jednotlivo, ale musia byť vykonané ako celok v jasne stanovenom

Tento projekt je podporený z Európskeho sociálneho fondu.

poradí. Vysvetlenie nadväznosti služby dátovej kvality pre transformáciu dát je bližšie vysvetlená v nasledujúcich kapitolách venujúcich sa službám DQ6 až DQ9.

Využívanie služieb DQ1 až DQ5

Služby dátovej kvality vytvorené v rámci projektu Dátovej integrácie (DI) je možné sprístupniť cez používateľské rozhranie, ktoré sa pre jednotlivé služby dátovej kvality líši. Predtým ako dátový kurátor príslušného OVM môže začať využívať tieto služby je nutné, aby požiadal o pridelenie prístupu do používateľských rozhraní na základe toho, ktoré služby dátovej kvality bude využívať. Dátový kurátor môže žiadať o sprístupnenie týchto používateľských prostredí, v rámci ktorých môže využívať nasledujúce dátové služby:

Prístup do portálu IS CSRÚ:

- DQ1 – Základná analýza dátovej kvality
- DQ2 – Validácia voči objektu evidencie (OE)
- DQ4 – Automatické stotožnenie

Prístup do Talend dátovej stewardship konzoly (DSC):

- DQ3 – Asistované čistenie
- DQ5 – Asistované stotožnenie

Využívanie služieb DQ6 až DQ9

Služby dátovej kvality DQ6 až DQ9 spoločne tvoria transformačnú dátovú službu, pre ktorej úspešné vykonanie je nutné, aby prebehli všetky služby dátovej kvality DQ6 až DQ9 zreťazene, tak ako boli navrhnuté. V tomto zreťazenom procese sú služby DQ6 až DQ8 bežiace na pozadí a konečný užívateľ k nim nemá prístup, tieto služby slúžia na integrovanie a nastavenie potrebných vzťahov medzi dátovým súborom a ontológiami CMÚ. Konečný užívateľ má teda prístup len k dátovej službe DQ9, ktorá slúži na nahratie súboru, ktorý si konzument / OVM želá transformovať do formátu RDF. Po nahratí bude takmer v reálnom čase vrátený naspäť súbor vo formáte RDF, ktorý môže byť následne využitý na potrebný účel. Sprístupnenie služby DQ9 bude umožnené prostredníctvom užívateľského prostredia IS CSRÚ tak ako v prípade služieb dátovej kvality DQ1,DQ2 a DQ4.

Tento projekt je podporený z Európskeho sociálneho fondu.

2.1.1 DQ1 – Základná analýza dátovej kvality

Charakteristika

Procedúra DQ1 je zameraná na Data profiling – poskytnutie štatistík o stave zdrojových údajov a ich kvalite, možnosti využitia jednotlivých štatistík sú rozpísané nižšie. Procedúra DQ1 je určená pre získanie výpisu kontroly kvality objektov evidencie, ktoré sú uložené v IS CSRÚ. Avšak pre objekty evidencie, ktoré nie sú uložené v IS CSRÚ je možné využiť procedúru DQ1_custom, ktorá je určená pre kontrolu takýchto objektov evidencie, respektíve pre špecifické dátové súbory na základe potreby OVM. V prípade oboch služieb dátovej kvality DQ1 a DQ1_custom je výstupom report vo formáte PDF.

Aplikácia

Služba dátovej kvality DQ1 je využívaná v prípade, keď je potrebné zistiť, aké nedostatky v dátovej kvalite sa nachádzajú na OE. V rámci služby dátovej kvality profilovania údajov (DQ1) je možné aplikovať na všetky vybrané atribúty (stĺpce) rôzne identifikátory na odhalenia dátovej nekvality, zároveň OVM môže definovať vlastné identifikátory na základe ich biznis požiadaviek pre identifikáciu špecifických nekvalít. Na základe výstupu z tejto dátovej služby OVM získa prehľad o rozsahu dátovej nekvality na meranom OE.

Funkcie

Základná analýza dátovej kvality bude poskytovať dátovým kurátorom informácie o kvalite ich dát pomocou rôznych analýz:

1. **Tabuľkové analýzy:**
   - Analýza obsahu stĺpcov
   - Analýza unikátnych záznamov
2. **Analýza dát v stĺpcoch**
   - Analýza zameraná na vyhľadávanie opakujúcich sa vzorov v stĺpci
   - Analýza dát v stĺpcoch na základe zvolených identifikátorov

Tieto analýzy bude možné vykonať pomocou identifikátorov, ktoré sú buď štandardnými a teda sú zahrnuté v prostredí Talend (opísané nižšie) alebo špeciálnymi, ktoré bude nutné vytvoriť na základe špecifických potrieb dátových kurátorov alebo konzumentov dát. Špeciálne identifikátory by mali vychádzať z biznis pravidiel, z ktorých sa následne vytvoria identifikátory priamo v Talend alebo budú vytvorené a importované do prostredia Talend ako REGEX vo formáte csv. Návod ako vytvárať biznis pravidlá, a čo je potrebné zvážiť pri ich vytváraní, je možné nájsť vo výstupe „1.1.1 Štandardizácia dátovej kvality“.

Štandardné identifikátory v Talend:

3. **Základné štatistiky**
   - Počet riadkov
   - Počet nulových záznamov

Tento projekt je podporený z Európskeho sociálneho fondu.

- Celkový počet záznamov
- Počet unikátnych záznamov
- Počet duplicitných záznamov
- Počet prázdnych záznamov

4. **Textové štatistiky –** využívajú sa na identifikáciu podmienok, ktoré by mali spĺňať textové záznamy v danom stĺpci. Príkladom môže byť napríklad stĺpec, ktorý obsahuje dátum, pre dodržanie interoperability musí tento dátum dodržiavať určitý formát, ktorý je napríklad možné skontrolovať aj štandardnými textovými štatistikami v Talend. Konkrétne je možné vykonať kontrolu maximálnej a minimálnej dĺžky záznamov v stĺpci a teda odhaliť, či spĺňa jednu z mnohých podmienok na jeho formátovanie.
   - Minimálna dĺžka
   - Minimálna dĺžka s nulovým záznamom
   - Minimálna dĺžka s prázdnym záznamom
   - Minimálna dĺžka s nulovým a prázdnym záznamom
   - Maximálna dĺžka
   - Maximálna dĺžka s nulovým záznamom
   - Maximálna dĺžka s prázdnym záznamom
   - Maximálna dĺžka s nulovým a prázdnym záznamom
5. Štatistiky opakujúcich sa vzorov
   - Frekvencia opakujúcich sa vzorov

Po vykovaní procedúry jej výstupom je jeden alebo viacero reportov podľa počtu vybraných analýz, zároveň je možné vytvoriť aj evolučný report záznamu, pre zobrazenie vývoja dátovej kvality pre vybrané identifikátory.

Použitie

Pre použitie dátovej služby DQ1 je potrebné sa prihlásiť do prostredia IS CSRÚ. Do CSRÚ portálu sa prihlásite prostredníctvom URL liniek v prehliadači. Zadanie správnej linky je závislé aj od typu sieťového prepojenia medzi IS CSRÚ a prostrediami systémov konkrétneho OVM – a síce sa deje buď v rámci siete Govnet, alebo je riešené priamym prestupom z internetovej adresy, ktorú špecifikovalo OVM pre účel prepojenia na IS CSRÚ.

Tento projekt je podporený z Európskeho sociálneho fondu.

Obrázok 1 – Výber používateľského prostredia

Prihlasovanie do samotného portálu sa deje prostredníctvom poskytnutých prihlasovacích údajov, ktoré boli určeným osobám na stranách OVM poskytnuté v rámci dokumentov pre technické účty IS CSRÚ (LDAP, Používateľ pre CSRÚ portál).

Obrázok 2 – Prihlásenie do používateľského prostredia

Podrobný návod ako ďalej pokračovať pri použití služby dátovej kvality DQ1 a DQ1_custom je možné nájsť v dokumente „IS CSRÚ Portál – DQ Procedúry manuál“

Výstupy

Ako už bolo spomenuté vyššie, výstupom tejto dátovej služby je report vo formáte PDF (obrázok 4). V tomto reporte dátový kurátor nájde výsledky z merania dátovej kvality na všetkých dátových atribútoch, pre ktoré boli vybrané špeciálne alebo základné štatistické identifikátory v rámci tejto dátovej služby. Výstup z tejto dátovej služby je možné sprístupniť dvomi spôsobmi. Prvým z nich je zaslanie na e-mailovú adresu (adresy) priradenú k danej službe na portáli CSRU v súbore .zip, v ktorom sa nachádza report v PDF formáte. Druhým variantom je prihlásiť sa na sFTP pomocou FTP klienta s použitím správneho prístupového konta a následné stiahnutie súboru s reportom vo formáte PDF. Oba varianty sú z mapované na procesnom diagrame v kapitole 2.3.

Tento projekt je podporený z Európskeho sociálneho fondu.

Obrázok 3 – Príklad výstupu z reportu profilovania dát

Row Count	# COUNT
Row Count	3039

DQ RULE	% MATCH	% NOT MATCH	# MATCH	# NOT MATCH
BP_MVSR_RA_TO_LESS_NOW	3.36%	96.64%	102	2937

Tento projekt je podporený z Európskeho sociálneho fondu.

2.1.2 DQ2 – Validácia voči objektu evidencie (OE)

Charakteristika

Procedúra DQ2 je procedúra profilovania dát vykonávaná nad údajmi objektu evidencie, ktorý je uložený v IS CSRÚ.

Pri procedúre typu DQ2 ide o overenie špecifických biznis pravidiel vzťahujúcich sa k danému atribútu údajov, a to prostredníctvom validácie voči prvkom Dátového slovníka 1xAD vytvoreného v spolupráci s MIRRI pre oblasť verejnej správy, ktorý vychádza z ontológií zadefinovaných v Centrálnom modeli údajov. Pre DQ2 je dostupný výstup iba pre momentálne aktuálne údaje v CSRÚ. Avšak pre objekty evidencie, ktoré nie sú uložené v IS CSRÚ je možné využiť procedúru DQ2_Custom, ktorá sa môže vykonať nad objektmi evidencie, ktoré nie sú priamo uložené v IS CSRÚ, resp. sa použije na spracovanie špecificky poskytnutých datasetov od OVM. V danom prípade je od OVM potrebný vstupný .csv súbor podľa vopred dohodnutej štruktúry atribútov. V prípade oboch služieb dátovej kvality DQ2 a DQ2_custom je výstupom report vo formáte PDF.

Aplikácia

Služba dátovej kvality DQ2 je určená na kontrolu dátovej kvality voči zadefinovanému štandardu dátovej kvality špecifického OE pomocou ontológií v Centrálnom modeli údajov prenesených do prostredia Talend v podobe dátového slovníka. Táto služba teda slúži na overenie, či kontrolovaný dátový súbor spĺňa podmienky interoperability vo verejnej správe podľa zavedených štandardov.

Funkcie

Procedúra validácie údajov slúži na kontrolu záznamov voči vopred definovaným štandardom, ktoré sú určené pre jednotlivé stĺpce v rámci špecifického záznamu. Tieto štandardy sú zadefinované na základe biznis pravidiel, ktoré hovoria o tom, ako by mali vyzerať kvalitné dáta v špecifickom zázname. Definované biznis pravidlá sú následne vo forme REGEX ukladané v rámci Dátového slovníka. Pre zachovanie referenčnej integrity je v rámci dátového slovníka obsiahnutý aj Unikátny referenčný identifikátor (URI).

Pokiaľ sa atribúty objektu evidencie nachádzajú v dátovom slovníku, je možné využiť túto procedúru dátovej kvality na validáciu záznamu z daného objektu evidencie. Výstupom tejto procedúry je jeden alebo viacero reportov podľa rozsahu nastavenej validácie, zároveň je možné vytvoriť aj evolučný report záznamu, pre zobrazenie vývoja dátovej kvality voči nastaveným štandardom z dátového slovníku.

Použitie

Pre použitie dátovej služby DQ2 je potrebné sa prihlásiť do prostredia IS CSRÚ. Do CSRÚ portálu sa prihlásite prostredníctvom URL liniek v prehliadači. Postup prihlásenia je totožný s postupom pri dátovej službe DQ1. Podrobný manuál použitia dátovej služby DQ2 je možné nájsť v „IS CSRÚ Portál – DQ Procedúry manuál“.

Výstupy

Ako už bolo spomenuté vyššie výstupom tejto dátovej služby je report vo formáte PDF. V tomto reporte dátový kurátor získa prehľad o dodržaní štandardov v rámci jednotlivých záznamov objektu evidencie a teda informáciu o dátovej kvalite jednotlivých atribútov. Výstup z tejto dátovej služby je možné sprístupniť dvomi spôsobmi. Prvým z nich je

Tento projekt je podporený z Európskeho sociálneho fondu.

zaslanie na e-mailovú adresu (adresy) priradenú k danej službe na portáli CSRU v súbore .zip, v ktorom som nachádza report v PDF formáte. Druhým variantom je prihlásenie sa na sFTP pomocou FTP klienta s použitím správneho prístupového konta a následné stiahnutie súboru s reportom vo formáte PDF. Oba varianty sú z mapované na procesnom diagrame v kapitole 2.3.

Tento projekt je podporený z Európskeho sociálneho fondu.

2.1.3 DQ3 – Asistované čistenie

Charakteristika

Pri type procedúry DQ3 ide o sprístupnenie Talend Data Stewardship konzoly (DSC) dátovému kurátorovi špecifického OVM, v ktorej na základe zistenej nekvality môže vykonávať nápravné opatrenia s cieľom zvýšiť dátovú kvality datasetu.

Aplikácia

Procedúra Asistovaného čistenia kurátorom poskytuje kampaň, ktorú je možné využiť na riadenie dátovej kvality v prostredí Talend DSC – Kampaň na čistenie dát. V rámci tejto kampane je úlohou dátového kurátora rozhodnúť a správnosti resp. chybe v zázname a upraviť ho do požadovaného formátu pre špecifické pole.

Okrem tejto základnej kampane je možné vytvoriť aj ďalšie kampane, ktoré je možné zostrojiť podľa špecifického prípadu použitia. Avšak v súčasnosti sa neplánuje nasadenia inej kampane ako pre čistenie údajov.

Funkcie

- Analyzovanie obsahu jedného alebo viacerých stĺpcov databázy na základe štandardov z dátového slovníka a na základe výsledkov analýzy vytváranie úloh na kontrolu a úpravu záznamov v používateľskom prostredí konzoly, kde dátovému kurátor pomocou ponúknutých možností realizuje čistenie nekvalitných dátových záznamov a ich prípadnú úpravu.

- Analyzovanie obsahu databázy na základe biznis pravidiel, následne na základe výsledkov analýzy sa vygenerujú úlohy na kontrolu a úpravu záznamov v používateľskom prostredí konzoly, kde dátovému kurátor pomocou ponúknutých možností realizuje čistenie nekvalitných dátových záznamov a ich prípadnú úpravu.

- Využitie dátovej služby na generovanie reportov obsahujúcich záznamy, ktoré boli identifikované ako problémové v rámci analýz.

Použitie

Pre využitie služby dátovej kvality DQ3 je potrebné prihlásenie cez portál CSRÚ podobne ako v prípade DQ1 a DQ2, avšak v prípade DQ3 je nutné vybrať prihlásenie do „CSRÚ Data Stewardship konzoly“ (viď. obrázok xy). Pre sprístupnenie tejto služby je potrebné získať iný typ prístupu ako pri CSRÚ portáli, žiadosť o sprístupnenie tejto služby je potrebné komunikovať s Dátovou kanceláriou. Podrobný manuál použitia dátovej služby DQ3 je možné nájsť v „IS CSRÚ Portál – DQ Procedúry manuál“.

Tento projekt je podporený z Európskeho sociálneho fondu.

Obrázok 4 - Prihlásenie CSRÚ Data Stewardship konzola

Výstup

Výstupom z kampane je vždy dátový súbor vo formáte CSV, v ktorom sa nachádzajú nové dáta s vykonanými úpravami uskutočnenými v kampani asistovaného čistenia údajov.

Tento projekt je podporený z Európskeho sociálneho fondu.

2.1.4 DQ4 – Automatické stotožnenie

Charakteristika

Pri type DQ4 procedúry dochádza k stotožňovaniu vybraného objektu evidencie s aktuálne integrovanými referenčnými registrami v rámci IS CSRÚ, ktorými sú Register adries (RA), Register právnických osôb (RPO), Register pre evidenciu a monitorovanie pomoci (SEMP). Stotožnenie je možné sprístupniť pomocou používateľského rozhrania, v ktorom používateľ zvolí objekt evidencie (dátový súbor), ktorý bude stotožnený s vybraným referenčným registrom. Postup, ako realizovať proces zadania žiadosti a vykonanie dátovej služby DQ4 automatického stotožnenia sa nachádza v súbore „IS CSRÚ Portál – DQ Procedúry manuál“.

Aplikácia

Služba dátovej kvality DQ4 je využitá v prípade, keď dátový kurátor OVM potrebuje overiť správnosť údajov v OE voči referenčnému registru automatizovane a teda všetky problémové záznamy v OE sú automaticky nahradené záznamami z referenčného registra.

Funkcie

Úlohou procedúry automatického stotožnenia údajov je poskytnúť dátovému kurátorovi nástroj, pomocou ktorého si bude môcť overiť vecnú pravdivosť údaju. Toto stotožnenie bude realizované porovnaním záznamu so záznamom z referenčného registra. V rámci tejto procedúry dôjde k stotožneniu záznamov, kde v prípade nezhody záznamov bude v OE nahradená informácia údajom z referenčného registra. Stotožnenie je realizované automatizovane, dátový kurátor musí len vybrať objekt evidencie a referenčný register, s ktorým bude OE stotožnený.

Služba dátovej kvality DQ4 je využitá pri potrebe skontrolovať záznamy OE voči referenčnému registru a tak overiť ich pravosť a úplnosť. V prípade služby DQ4 ide o automatické stotožnenie, a teda dátový kurátor dostáva ako výstup z tejto dátovej procedúry súbor, v ktorom sú nesprávne údaje v OE nahradené správnymi údajmi z referenčného registra.

Použitie

Pre použitie dátovej služby DQ4 je potrebné prihlásenie sa do „Portálu CSRÚ“ rovnako, ako v prípade sprístupnenia dátovej služby DQ1 a DQ2. Podrobný manuál použitia dátovej služby DQ4 je možné nájsť v „IS CSRÚ Portál – DQ Procedúry manuál“.

Výstup

Výstupom z kampane je vždy dátový súbor vo formáte CSV, v ktorom sa nachádzajú nové dáta, ktoré prešli procesom nahradenia chybných dáta v objekte evidencie dátami z referenčného registra. V prípade stotožnenia s Registrom adries (RA) a Registrom právnických osôb (RPO) výstupom môže byť aj súbor vo formáte XLSX/XML/JSON.

Tento projekt je podporený z Európskeho sociálneho fondu.

2.1.5 DQ5 – Asistované stotožnenie

Charakteristika

Pri DQ5 procedúre dochádza k stotožňovaniu vybraného objektu evidencie s aktuálne integrovanými referenčnými registrami v rámci IS CSRÚ, ktorými sú Register adries (RA), Register právnických osôb (RPO). Stotožnenie je realizované v prostredí Talend Data Stewardship konzoly, v ktorej dátový kurátor vidí jednotlivé problémy odhalené pri stotožnení s referenčným registrom a sú mu ponúknuté možnosti ako riešiť tieto vzniknuté problémy v dátovej kvalite.

Aplikácia

Služba dátovej kvality DQ5 je využitá v prípade, keď dátový kurátor OVM potrebuje overiť správnosť údajov v OE voči referenčnému registru, avšak na rozdiel od DQ4 je potrebné, aby odhalené problémy riešil jednotlivo na základe vlastného úsudku.

Funkcie

Procedúra manuálneho stotožnenia údajov má rovnakú úlohu ako automatické stotožnenie údajov, a to overiť vecnú pravdivosť údaju na základe stotožnenie s objektom evidencie. Rozdielom je, že v tomto prípade dátový kurátor vykonáva deduplikáciu manuálne v rámci Talend data Stewardship konzoly.

Služba dátovej kvality DQ5 je využitá pri potrebe skontrolovať záznamy OE voči referenčnému registru, a tak overiť ich pravosť a úplnosť. V prípade služby DQ5 ide o asistované stotožnenie, a teda dátový kurátor má v prostredí Talend Data Stewardship konzoly sprístupnené problémy, odhalené pri stotožnení s referenčným registrom. Následne sú mu ponúknuté rôzne možnosti, ako ich riešiť.

Použitie

Pre použitie dátovej služby DQ5 je potrebné prihlásenie sa do „CSRÚ Data Stewardship konzoly“ rovnako, ako v prípade sprístupnenia dátovej služby DQ3. Podrobný manuál použitia dátovej služby DQ5 je možné nájsť v „IS CSRÚ Portál – DQ Procedúry manuál“.

Výstup

Výstupom z kampane je vždy dátový súbor vo formáte CSV,XLSX/XML/JSON, v ktorých sa nachádzajú nové dáta, ktoré prešli procesom nahradenia chybných dáta v objekte evidencie dátami z referenčného registra.

2.1.6 DQ6 – Vstupná analýza datasetu pre transformáciu

Charakteristika

Slúži na prípravu vstupného dátového súboru (XML/XSD) do formátu, ktorý bude možné ďalej spracovávať v transformačnom reťazci. Táto služba je „kreatívnou“ službou, a teda predstavuje krok, v rámci ktorého sú rozpoznávané jednotlivé atribúty datasetu, ich kvalita a vhodnosť na vykonanie transformácie dátového súboru. Táto služba nie je priamo sprístupnená dátovému kurátorovi OVM.

Aplikácia

Služba dátovej kvality DQ6 je využitá v prípade potreby transformovať dátový súbor do formátu RDF. Je prvou transformačnou službou dátovej kvality v reťazci a vždy musí byť využívaná spolu s ostatnými služba pre dosiahnutie úspešnej transformácie.

Funkcie

Predtým, ako vôbec táto dátová služba môže prebehnúť nad určitým dátovým súborom je nutné naplniť tieto podmienky:

- Dataset, ktorý je predmetom analýzy, musí byť k dispozícii vo forme aspoň jednej dátovej inštancie vo formáte XML
- Dataset, ktorý je predmetom analýzy, musí mať definovanú štruktúru vo formáte XSD schémy
- Dátový súbor s popisom metadát XML súboru, ktorý je predmetom analýzy, musí byť dostupný v popisnom formáte XLS (štruktúra datasetu a biznis popis položiek)

V tejto dátovej službe dochádza k analyzovaniu vstupného XML/XSD datasetu (Analýza biznis povahy datasetu) spoločne s analýzou ontológii súvisiacich s dátovým modelom transformovaného datasetu. Zároveň sa vykonáva aj identifikácia a interpretácia pôvodného formátu údajov, aby bolo zistené, čo je potrebné urobiť s údajmi na vykonanie transformácie do požadovaného stavu. Jednotlivé činnosti v tejto dátovej službe sú vykonané v tomto poradí:

1. Sprístupnenie popisov štruktúry a obsahu datasetu v XLS formáte
2. Analýza položiek/elementov datasetu
3. Analýza a pochopenie biznis povahy datasetu
4. Analýza schémy datasetu vo formáte XSD a validácia datasetu voči schéme
5. Analýza príkladu/príkladov konkrétnych inštancií datasetu vo formáte XML

Hlavným cieľom tejto fázy transformácie je zadefinovať, čo je nutné urobiť s údajmi, aby sa transformovali do požadovaného tvaru pre daný prípad použitia. Zvyčajne sa na to používa nástroj na profilovanie údajov. Profilovanie údajov sa vzťahuje na proces skúmania, analýzy, validácie a sumarizácie datasetov s cieľom získať prehľad o kvalite údajov. V rámci profilovania analytické algoritmy zisťujú charakteristiky datasetov, ako je priemer, minimum, maximum, percentil a frekvencia, aby bolo možné preskúmať údaje

do najmenších detailov. Následne algoritmy vykonávajú analýzy na odhalenie metadát vrátane rozdelenia frekvencie výskytu, kľúčových vzťahov, kandidátov na cudzie kľúče a funkčných závislostí. Nakoniec sa všetky tieto informácie použijú na odhalenie toho, ako sú tieto faktory v súlade so štandardmi a cieľmi dátového programu a dátovej kvality. Profilovanie údajov tak dokáže eliminovať nákladné chyby, ktoré sú bežné v databázach. Medzi tieto chyby patria nulové hodnoty (neznáme alebo chýbajúce hodnoty), hodnoty, ktoré by sa nemali ukladať, hodnoty s neobvykle vysokou alebo nízkou frekvenciou výskytu, hodnoty, ktoré sa neriadia očakávanými vzormi a hodnoty mimo bežného rozsahu. Takéto automatizované profilovanie poskytuje aj nasadený nástroj Talend a v čase písania tohto dokumentu sa implementuje v rámci projektu Dátovej integrácie – časť Dátová kvalita. V rámci tohto projektu sa robí aj ďalšia častá aktivita, ktorá je kľúčová pre profilovanie a dátovú kvalitu, a tou je definovanie takzvaných regulárnych výrazov („regular expressions (REGEX)“) pre vybrané dátové prvky - tieto informácie sú aj zaevidované priamo v dátovom slovníku nástroja Talend, ktorý sa využíva aj na spúšťanie kampaní pre zvyšovanie kvality údajov. Následne možno v spomínanom nástroji Talend získať profil datasetu, ktorý analyzuje percento údajov, ktoré spĺňajú daný vzor (napríklad REGEX pre e-mailovú adresu), ako aj ďalšie informácie (ako napríklad koľko položiek v datasete má nevyplnený údaj o e-mailovej adrese a koľko existuje duplicitných údajov).

Dátoví špecialisti používajú techniky profilovania údajov alebo skripty na zachytenie štruktúry a vlastností údajov a určenie, ako by sa mali upraviť. Teda, čo z nasledujúceho zoznamu je potrebné pre daný objekt evidencie alebo dataset spraviť:

- Zmeniť formát,
- Zmeniť štruktúru,
- Zmeniť dátový model,
- Aplikovať iný štandard,
- Zvýšiť kvalitu údajov bez zmeny obsahu údajov, napríklad odstránením chýbajúcich hodnôt, určením („casting“) a konverziou dátových typov pre kompatibilitu, úpravou dátumov a časov pomocou posunov a lokalizácie formátu, premenovaním schém, tabuliek a stĺpcov pre prehľadnosť, či vylepšením formátu danej hodnoty dátového prvku alebo obohatením údajov.

Použitie

V čase písanie tohto dokumentu nebol dostupný presný popis používania tejto služby. Popis ako je možné ju využiť bude dostupný po jeho sfinalizovaní v dokumente „IS CSRÚ Portál – DQ Procedúry manuál“.

Výstup

Výsledkom tejto dátovej služby je validácia správnosti vykonanej analýzy vstupného datasetu prostredníctvom doložených vstupných analytických podkladov a výstupného zoznamu identifikovaných ontológií. Keďže služba dátovej kvality DQ6 ako taká neprináša konečnú hodnotu, ale iba čiastkovú v rámci celého procesu transformácie pre úspešné vykonanie transformácie dátového súboru, získané výstupy z tejto dátovej služby slúžia ako vstupy pre vykonanie aktivít v rámci procedúry dátovej kvality DQ7, ktorá je ďalšou nadväzujúcou v procese transformácie.

Tento projekt je podporený z Európskeho sociálneho fondu.

2.1.7 DQ7 – Mapovanie na ontologické modely

Charakteristika

Mapovanie štruktúr XML/XSD datasetu na ontologické modely

Mapovanie údajov („data mapping“) zahŕňa extrakciu dátových prvkov z viacerých zdrojov a ich následné priradenie k cieľovým dátovým prvkom u konzumenta – v cieľovom informačnom systéme.

Pre úspešnú dátovú integráciu musia mať zdrojové a cieľové údaje podobné dátové modely. Je však zriedkavosťou, aby dve dátové úložiská mali rovnakú schému. Tu prichádza na rad mapovanie údajov, ktoré ukáže, kde sú nezrovnalosti medzi dvoma datasetmi.

Aplikácia

Služba dátovej kvality DQ7 je využitá v prípade potreby transformovať dátový súbor do formátu RDF. Je druhou transformačnou službou dátovej kvality v reťazci a vždy musí byť využívaná spolu s ostatnými službami pre dosiahnutie úspešnej transformácie.

Funkcie

Pri zmene dátového modelu ide spravidla o zosúladenie údajov zo zdroja s Centrálnym modelom údajov, ktorý je podrobne popísaný v dokumente 1.1.2 Štandardizácia pre modelovanie údajov. S tým súvisí aj aplikovanie iného štandardu, ak sú zdrojové údaje napríklad vo formáte XML alebo JSON, je potrebné ich previesť do štandardu RDF, obohatiť o ďalšie informácie z ontológií Centrálneho dátového modelu a následne serializovať napríklad do formátu JSON-LD.

Postup procesu vykonania DQ7 je nasledovný:

- Obohatenie vstupného XML datasetu s využitím metódy stotožnenia voči základným číselníkom
- Mapovanie štruktúr XML datasetu na identifikované ontologické modely z procedúry DQ6
- Vytvorenie predpisu pre požadovaný RDF formát (JSON-LD) výstupného datasetu

Použitie

V čase písanie tohto dokumentu nebol dostupný presný popis používania tejto služby. Popis bude dostupný po sfinalizovaní služby v dokumente „IS CSRÚ Portál – DQ Procedúry manuál“.

Výstup

Výstupom je predpis RDF formátu JSON-LD a obohatený XML súbor. Tak ako v prípade DQ6 aj DQ7 poskytuje iba čiastkovú hodnotu v rámci celého procesu transformácie, výstupy z tejto služby dátovej kvality slúžia ako vstup pre vykonanie aktivít v rámci procedúry dátovej kvality DQ8, ktorá je ďalšou nadväzujúcou v procese transformácie.

Tento projekt je podporený z Európskeho sociálneho fondu.

2.1.8 DQ8 – Transformácia datasetu do formátu RDF

Charakteristika

Mapovanie jednotlivých elementov, vrátane zápisu transformácie jazykom XSLT v rámci dátovej služby DQ8 prebehne v týchto krokoch:

- Navrhnutie XSLT transformácie, ktorá definuje transformáciu vstupného datasetu (v obohatenej/stotožnenej forme) do výstupnej podoby podľa definovaného JSON-LD predpisu.
- Overenie správnosti navrhnutej XSLT transformácie pomocou lokálnej (offline) verzie služby transformačného modulu.

Aplikácia

Služba dátovej kvality DQ8 je využitá v prípade potreby transformovať dátový súbor do formátu RDF. Je treťou transformačnou službou dátovej kvality v reťazci a vždy musí byť využívaná spolu s ostatnými služba pre dosiahnutie úspešnej transformácie.

Funkcie

Tento krok transformácie je úplne manuálny, poverená osoba podľa metodológie (bude dostupná až po skončení projektu DI) pripraví transformačný predpis XSLT. V rámci tejto služby dátovej kvality teda poverená osoba pripravuje XSLT transformačný predpis podľa definovaného JSON-LD predpisu cieľového formátu RDF tak, aby obohatený XML súbor mohol byť správne transformovaný. Poverená osoba pracuje vo vybranom integrovanom vývojovom prostredí, ktoré podporuje prehľadný vývoj softvéru, v tomto prostredí vytvára XSLT transformačný predpis, v rámci ktorého využíva vnorené časti kódu, ktoré sú bežne pre viaceré transformácie a teda je na tieto časti transformácie len odkazované.

Obrázok 5 – Príklad vnorených častí XSLT

```
<xsl:template match="Note">
    <skos:note rdf:datatype="http://www.w3.org/2001/XMLSchema#string">
        <xsl:value-of select="." />
    </skos:note>
</xsl:template>

<xsl:template match="Date">
    <dct:valid rdf:datatype="http://www.w3.org/2001/XMLSchema#date">
        <xsl:value-of select="."/>
    </dct:valid>
</xsl:template>
```

Použitie

V čase písanie tohto dokumentu nebol dostupný presný popis používania tejto služby. Popis ako je možné ju využiť bude dostupný po jeho sfinalizovaní v dokumente „IS CSRÚ Portál – DQ Procedúry manuál“.

Výstup

Tento projekt je podporený z Európskeho sociálneho fondu.

Výsledkom je validovaný XSLT predpis transformácie vstupného stotožneného a obohateného datasetu do výstupného formátu v JSON-LD. Tak ako v prípade DQ6 a DQ7 ani DQ8 neprináša konečnú hodnotu, vytvára a validuje XSLT transformáciu, ktorá bude použitá ako povinný vstup pre vykonanie procedúry DQ9, ktorá je poslednou nadväzujúcou dátovou službou, ktorá prináša konečnú hodnotu v procese transformácie, a teda poskytuje transformovaný súbor vo formáte RDF.

Tento projekt je podporený z Európskeho sociálneho fondu.

2.1.9 DQ9 – Nasadenie transformovaného datasetu vo formáte RDF

Charakteristika

Služba dátovej kvality DQ9 v rámci transformačného reťazca DQ6 až D9 je službou, s ktorou bude prichádzať do kontaktu koncový používateľ a teda dátový kurátor príslušného OVM. Pomocou tejto služby po integrovaní dátového súboru do transformačného modulu bude môcť dátový kurátor zaslať žiadosť o transformáciu XML súboru. Následne (takmer v reálnom čase) dostane odpoveď v podobe transformovaného súboru vo formáte RDF. Dátovú službu DQ9 je možné využiť na transformovanie len tých súborov, ktoré sú integrované v rámci transformačného modulu (DQ6 – DQ8). Podmienkou pre využitie dátovej služby DQ9 je aby:

- Dataset, ktorý je predmetom transformácie, v pôvodne poskytnutej zdrojovej verzii
- Transformačný predpis pre daný dataset bol uložený v úložisku transformácií vo forme XSLT súboru

Aplikácia

Služba dátovej kvality DQ9 je využitá v prípade potreby transformovať dátový súbor do formátu RDF. Táto služba môže byť využitá dátovým kurátorom OVM pomocou integrovania pripojenia tejto služby na IS OVM.

Funkcie

V poslednom kroku transformačného reťazca službou dátovej kvality DQ9 je vykonaná transformácia zdrojového XML súboru poskytnutého OVM. V rámci transformácie sa udejú procesy, ktoré boli nastavené v rámci služieb DQ6 až DQ8. Vstupom do transformácie je žiadosť informačného systému OVM (integrovaného na transformačný modul), ktorý si vyžiada transformovanie integrovaného XML dátového súboru v transformačnom module. Po vyžiadaní transformácie sa udejú nasledujúce kroky:

1. Je volaná služba obohatenia a stotožnenia, ktorú pokrývajú dátové služby DQ6 a DQ7. V rámci tohto kroku je zdrojový XML súbor obohatený, a teda sú doňho pridané chýbajúce číselníky a opravené chybné záznamy prípadne štruktúra XML súboru. Výstupom je nový obohatený XML dátový súbor.
2. Realizuje sa transformácie obohateného XML dátového súboru podľa vytvorenej XSLT transformácie v dátovej službe DQ8. Výstupom je súbor vo formáte JSON-LD.
3. Súbežne na pozadí sa vykoná kontrola transformovaného súboru, či spĺňa všetky požadované vlastnosti zadefinované v XSLT transformačnom predpise.
4. Pokiaľ pre transformovaný typ údajov je dostupná SHACL validácia je tiež vykonaná na pozadí.
5. Výstup je zaslaný systému, ktorý si vyžiadal transformáciu

Použitie

V čase písanie tohto dokumentu nebol dostupný presný popis používania tejto služby. Popis, ako je možné ju využiť bude dostupný po jeho sfinalizovaní v dokumente „IS CSRÚ Portál – DQ Procedúry manuál“.

Výstup

Výstupom dátovej služby DQ9 je dataset vo formáte JSON-LD, ktorý vznikol transformáciou vstupného datasetu.

Tento projekt je podporený z Európskeho sociálneho fondu.

2.1.10 DQ10 – Vlastná služba dátovej kvality podľa špecifických potrieb konkrétneho OVM

Ak existuje záujem zo strany OVM o vytvorenie špecifickej služby dátovej kvality inej, ako sú služby popísané vyššie, je potrebné, aby OVM kontaktoval Dátovú kanceláriu na Ministerstve investícií, regionálneho rozvoja a informatizácie a požiadal o vytvorenie takejto služby. OVM spolu s Dátovou kanceláriou spoločne vytvoria požiadavku na vytvorenie takejto dátovej služby. Nasleduje schvaľovací proces, na konci ktorého bude vytvorená výzva pre vytvorenie služby dátovej kvality. Príprava dátovej služby bude realizovaná dodávateľom projektu DI spolu s OVM, ktoré v počas prípravy dátovej služby bude bližšie špecifikovať svoje požiadavky.

V súčasnosti vzniká dátová služby DQ10, ktorá je pripravovaná pre Ministerstvo vnútra nad Registrom adries (RA), ktoré spravuje tento register. Táto služba bude zameraná na pravidelnú kontrolu záznamov v RA podľa definovaných biznis pravidiel, ktorej výstupom bude zoznam záznamov, ktoré nespĺňajú zadefinované biznis pravidlá. Táto služba dátovej kvality bude spúšťaná automatizovane a výstup z nej bude zasielaný na určené emailové adresy pracovníkom MV, ktorý budú zodpovední za nápravu zistených nesúladov v záznamoch. V čase písanie tohto výstupu implementácia tejto dátovej služby ešte nebola k dispozícii, a teda detaily jej fungovania a používania nie sú známe.

2.2 Služby dátovej kvality projektu CIP PaaS

Projekt CIP prináša jednu vlastnú službu dátovej kvality, a to nástroj na profilovanie dátových súborov viď. kapitola 2.2.2. Zároveň však sprístupňuje niektoré zo služieb dátovej kvality, ktoré boli vytvorené v rámci projektu Dátovej integrácie (DI).

2.2.1 Služby z projektu DI sprístupnené ako PaaS

Z projektu DI sú sprístupnené v CIP PaaS tieto služby dátovej kvality:

- DQ3 – Asistované čistenie (viď. kapitola 2.1.3)
- DQ4 – Automatické stotožnenie (viď. kapitola 2.1.4)
- DQ9 – Nasadenie transformovaného datasetu vo formáte RDF (viď. kapitola 2.1.9)

DQ9 – Nasadenie transformovaného datasetu vo formáte RDF

Pripravovaná služba DQ9 bude mať rozšírené možnosti sprístupnenia v rámci CIP PaaS. Klasická služba dátovej kvality DQ9 transformuje vstupný súbor do formátu JSON-LD. Upravená DQ9 bude umožňovať získať aj medzivýstupy tejto dátovej služby a zároveň transformáciu do formátu RDF. V upravenej dátovej službe DQ9 bude možné získať tieto výstupy samostatne (dátový súbor, na ktorý budú aplikované musí byť integrovaný do transformačného modulu):

- Využitie časti obohatenia dátového súboru
- Využitie časti stotožnenia dátového súboru
- Využitie na transformáciu do RDF / JSON-LD

V čase písania tohto dokumentu ešte nebolo dokončená finálna verzia sprístupnenia, a teda nie je možné presne špecifikovať spôsob, akým budú sprístupnené tieto služby cez webové používateľské rozhranie.

2.2.2 Nástroj na profilovanie dátových súborov OpenMetadata

Nástroj OpenMetadata[1] zabezpečí službu dátovej kvality, ktorá umožní dátovým kurátorom skúmať súbor údajov a pochopiť jeho hlavné charakteristiky. V rámci tejto dátovej služby bude poskytnutý balík dataprep.eda, ktorý zjednodušuje proces profilovania tým, že umožňuje používateľovi preskúmať dôležité vlastnosti pomocou rozhraní API. Každé API následne umožní používateľovi analyzovať súbor od vysokej úrovne až po nízku úroveň a z rôznych perspektív.

V rámci služby dátovej kvality budú poskytnuté nasledujúce funkcie:

- **Analýza distribúcií –** Analyzuje rozdelenie stĺpcov, skúma distribúciu stĺpcov a štatistiky množín údajov. Zistí typ a potom vytvorí rôzne grafy a štatistiky, ktoré sú vhodné pre príslušný typ. Používateľ môže voliteľne označiť jeden alebo dva

[1] OpenMetadata Documentation: Get Help Instantly (open-metadata.org)

Tento projekt je podporený z Európskeho sociálneho fondu.

stĺpce záujmu ako parameter. V prípade že je označený jeden stĺpec, jeho rozdelenie sa vykreslí rôznymi spôsobmi a vypočíta sa štatistika stĺpcov. Ak sú značené dva stĺpce vygenerujú sa grafy zobrazujúce vzťahy medzi týmito dvoma stĺpcami.

- **Analýza korelácií –** Analyzuje korelácie medzi stĺpcami rôznymi spôsobmi pomocou viacerých korelačných metrík. Štandardne zobrazuje korelačné matice s rôznymi metrikami. Používateľ môže voliteľne odovzdať jeden alebo dva stĺpce záujmu ako parametre. V prípade, že odošle jeden stĺpec, vypočíta sa a zoradí sa korelácie medzi týmto stĺpcom a všetkými ostatnými stĺpcami. Ak sú zadané dva stĺpce vykreslí sa bodový graf a regresná čiara.

- **Analýza chýbajúcich hodnôt –** Analyzuje chýbajúce hodnoty a ich vplyv na dátový súbor. V predvolenom nastavení vygeneruje rôzne grafy, ktoré zobrazujú množstvo chýbajúcich hodnôt pre každý stĺpec a akékoľvek základné vzory chýbajúcich hodnôt v množine údajov. Na pochopenie vplyvu chýbajúcich hodnôt v jednom stĺpci na ostatné stĺpce môže používateľ zadať názov stĺpca ako parameter, na základe čoho sa vygeneruje distribúcia každého stĺpca s chýbajúcimi hodnotami daného stĺpca a bez nich, čo umožní dôkladne pochopiť ich vplyv

- **Vytvorenie reportu profilovania –** Generuje komplexnú správu profilu súboru údajov v týchto kategóriách:

    - **Prehľad –** Zisťovanie typov stĺpcov v dátovom rámci
    - **Premenné –** Typ premennej, jedinečné hodnoty, rozdielny počet, chýbajúce hodnoty
    - **Kvartilové štatistiky –** Minimálna hodnota, prvý kvartil, medián, tretí kvartil, maximum, rozsah, medzikvartilový rozsah
    - **Deskriptívne štatistiky –** Priemer, režim, štandardná odchýlka, súčet, absolútna odchýlka, variačný koeficient, špicatosť, šikmosť
    - **Analýza textu –** Dĺžka textu, vzorka a písmeno
    - **Korelácie –** Zvýraznenie vysoko korelovaných premenných, Spearmanova, Pearsonova a Kendallova matica
    - **Chýbajúce hodnoty –** Stĺpcový graf, tepelná mapa a spektrum chýbajúcich hodnôt

- **Analýza časových radov údajov**

Rozdiel medzi DQ1 a OpenMetadata

Nástroj OpenMetadata je službou dátovej kvality určenou na profilovanie údajov podobne ako služba dátovej kvality DQ1 – Základná analýza dátovej kvality. Hlavným rozdielom je, že pri profilovaní cez OpenMetadata dátový kurátor získava nástroj na profilovanie, v ktorom si musí nastaviť všetky merania na jednotlivé atribúty v dátovom súbore, ktoré má záujem skontrolovať a musí nastaviť všetky súvisiace konfigurácie. Naopak, pri službe dátovej kvality DQ1 – Základná analýza dátovej kvality ide o asistované profilovanie údajov, a teda všetky nastavenia sú vykonané za dátového

kurátora podľa jeho potrieb a pre získanie reportu stačí zaslať na Portáli CSRÚ len žiadosť o vygenerovanie reportu, ktorý mu bude zaslaný. Avšak, keďže ide o asistované profilovanie údajov je nutné, aby bola zadaná výzva na vytvorenie profilovania pre požadovaný objekt evidencie predtým, ako je možné cez portál zaslať žiadosť o vygenerovanie reportu. Výhodou DQ1 sú teda minimálne nároky na znalosti dátové kurátora v oblasti profilovania dát. Pri OpenMetadata sa u dátového kurátora predpokladajú značné znalosti s meraním dátovej kvality a profilovaním údajov, aby bol schopný vytvoriť report. V tomto prípade je samozrejme výhodou plná flexibilita riadenia nástroja a plné využitie jeho možností na pokročilú dátovú analýzu.

Príkladom uvedených rozdielov je aj report, ako výstup z týchto služieb. V prípade DQ1 – Základná analýza dátovej kvality je dostupný len statický report v PDF. OpenMetadata na rozdiel od DQ1 poskytuje interaktívny report, ktorý je vygenerovaný vo webovom rozhraní a užívateľovi umožňuje vytváranie rôznych pohľadov na analýzu dátovej kvality a dynamické prispôsobovanie výstupu postupnému získavaniu vhľadu do charakteristík skúmaného súboru.

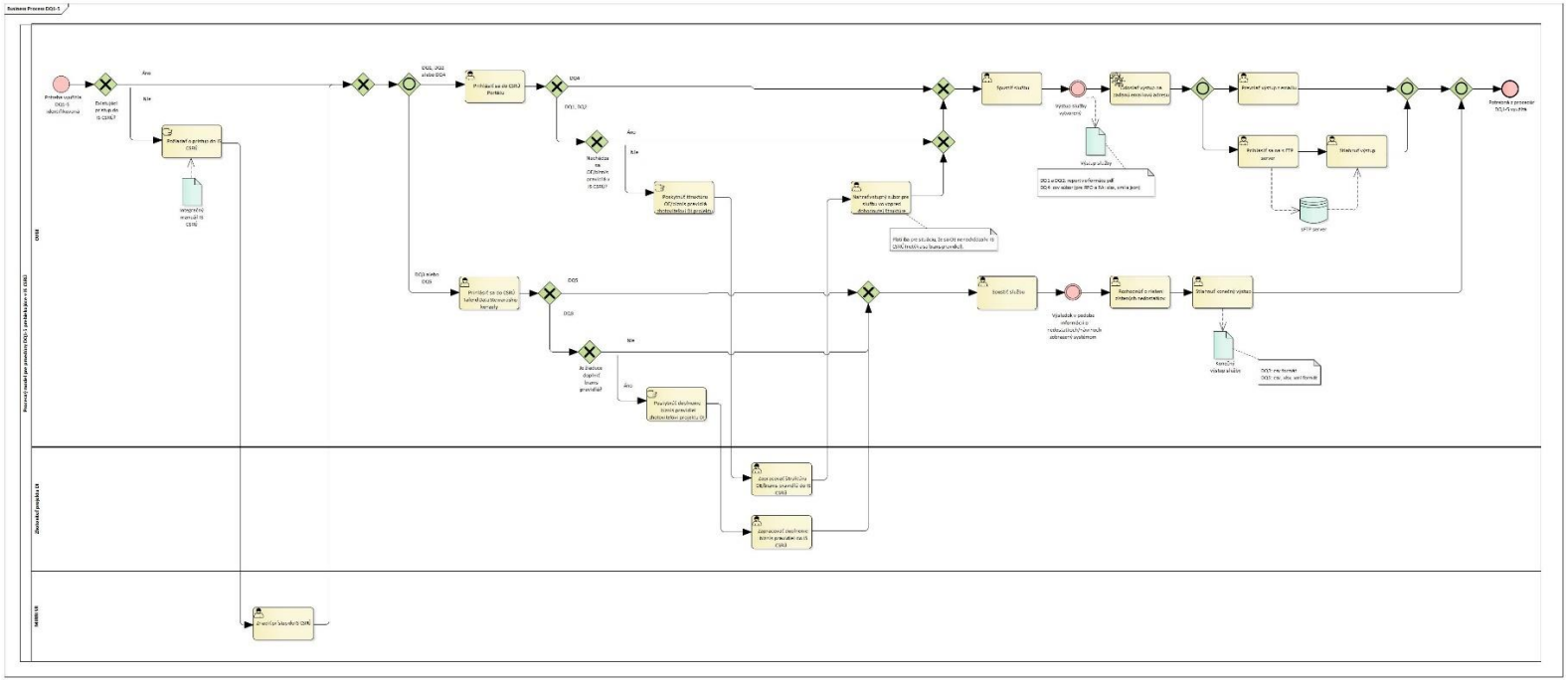
Tento projekt je podporený z Európskeho sociálneho fondu.

2.3 Zhrnutie postupu využívania služieb dátovej kvality

Služby dátovej kvality DQ1 až DQ5

Nižšie na obrázku je zmapovaný procesný diagram služieb dátovej kvality DQ1 až DQ5. V procesnom diagrame je zmapovaný proces od zistenia potreby získania prístupu k určitej dátovej službe až po úspešné spustenie služby a následné získanie požadovaného výstupu. Pre lepšiu čitateľnosť procesného diagramu je ho možné nájsť v prílohe 1: *Služby dátovej kvality DQ1 až DQ5*

Obrázok 6 - Procesný diagram služieb dátovej kvality DQ1 - DQ5

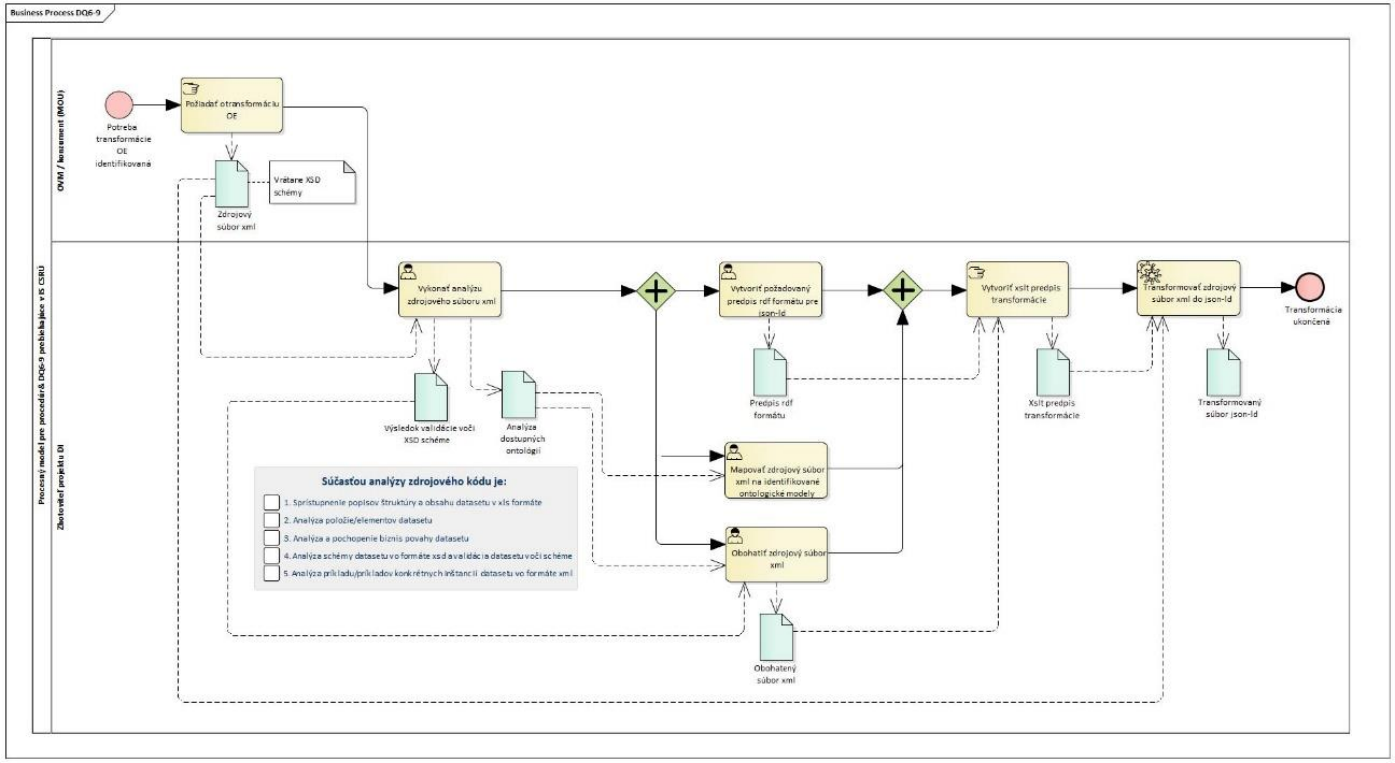


Tento projekt je podporený z Európskeho sociálneho fondu.

Služby dátovej kvality DQ6 až DQ9

Služba dátovej kvality zameraná na transformáciu dátových súborov, ktorá je zložená zo služieb dátovej kvality DQ6 až DQ9 je zmapovaná procesným diagramom na obrázku 9. V procesnom diagrame je zobrazený proces transformácie od vytvorenia požiadavky na transformovanie špecifického dátového súboru, až po ukončenie procesu transformácie, kedy je možné transformovať daný súbor opakovane. Pre lepšiu čitateľnosť procesného diagramu je ho možné nájsť v prílohe 2: *Služby dátovej kvality DQ6 až DQ9*

Obrázok 7 - Procesný diagram služieb dátovej kvality DQ6 - DQ9



Tento projekt je podporený z Európskeho sociálneho fondu.

3 Podpora využívania nástrojov dátovej kvality v praxi

Táto kapitola sa venuje navrhnutiu šablóny pre zmapovanie služieb dátovej kvality z hľadiska používateľského využitia, pričom bude zahŕňať niektoré technické aspekty dátových služieb, ktoré je potrebné, aby koncoví používatelia - dátoví kurátori jednotlivých OVM poznali. Taktiež sa v tejto kapitole nachádza návrh, ako riadiť podporu využívania nástrojov dátovej kvality, vrátane spôsobu sprístupňovania príručiek, návodov a iných zdrojov pre dátových kurátorov.

3.1 Riadenie podpory využívania nástrojov dátovej kvality

Je dôležité navrhnúť jednotný prístup ku komunikácii a koordinácii dátových kurátorov jednotlivých OVM pre efektívne využívanie služieb dátovej kvality. Vytvorenie efektívneho modelu riadenia podpory využívania služieb dátovej kvality musí naplniť niekoľko predpokladov, ktoré zabezpečia, že dátoví kurátori budú mať k dispozícii efektívnu podporu. Je ich možné rozdeliť do týchto kategórií:

- Pravidelná komunikácia s dátovými kurátormi
- Poskytnutie používateľských príručiek
- Poskytnutie konzultácii a technickej podpory
- Školenie pre dátových kurátorov OVM

Predpokladáme, že vhodným riešením na zabezpečenie koordinovanej a štruktúrovanej podpory je vytvorenie webového prostredia, v ktorom budú pre dátových kurátorov dostupné potrebné informácie, čím sa dosiahne efektívne využívanie služieb dátovej kvality. V súčasnosti sa síce plánuje, že v rámci používateľského rozhrania IS CSRÚ, v ktorom budú sprístupnené služby dátovej kvality budú zahrnuté rôzne návody na to, ako tieto služby využívať, avšak tieto návody neposkytnú dátovým kurátorom komplexný pohľad, keďže množstvo informácií v tomto rozhraní nebude dostupných. Zároveň mnohí dátoví kurátori nebudú mať prístup do tohto prostredia predtým, ako získajú špecifický prístup, a teda nebudú môcť zhodnotiť, do akej miery ich potreby budú naplnené získaním prístupu k službám dátovej kvality. Preto je dôležité vytvorenie dedikovanej webovej stránky zameranej na komplexnú podporu a informovanie dátových kurátorov v oblasti využívania služieb dátovej kvality. Za vhodné riešenie je možné považovať napríklad vytvorenie webového prostredia v rámci existujúceho portálu datalab.digital. Na tomto portáli by bol vytvorený priestor, zameraný na navigovanie dátových kuriérov v oblasti dátových služieb minimálne v rozsahu, ktorý je opísaný nižšie.

Pravidelná komunikácia s dátovými kurátormi

Hlavný predpokladom pre zvyšovanie dátovej kvality vo verejnej správe je vytvorenie a udržiavanie komunikačných kanálov so všetkými dátovými kurátormi OVM. V rámci toho je potrebné mapovať a udržiavať aktuálny zoznam informácií o dátových kurátoroch, ktorý by mal obsahovať údaje o ich zručnostiach, skúsenostiach, realizovaných v oblasti dátovej kvality a iných údajoch, ktoré je dobré poznať pre efektívne rozvíjanie spolupráce v oblasti dátovej kvality. Na zabezpečenie komunikačných kanálov s dátovými kurátormi môžu byť využité tieto spôsoby:

Tento projekt je podporený z Európskeho sociálneho fondu.

- Push/Pull komunikácia (mailing list, subscription, ...)
- Pravidelné online stretnutia dátových kurátorov v primeranej frekvencii (prípadne iná forma pravidelnej komunikácie, vrátane osobných stretnutí)
- **Webové rozhranie –** webové rozhranie by malo poskytnúť dátovým kurátorom dostatočné množstvo informácií o službách dátovej kvality vrátane informácií, ktoré sú kľúčové pre ich sprístupnenie, využívanie a riešenie žiadostí a rôznych vzniknutých technických alebo iných problémov. Špecificky by mali byť dátovým kurátorom sprístupnené tieto informácie a funkcionality:
  - Odpovede na často kladené otázky (FAQ)
  - Zverejnenie relevantných typov príručiek (z odseku nižšie)
  - Poskytnutie diskusných fór pre dátových kurátorov
  - Newsletter s novinkami (pripravované zmeny, pripravované výzvy, realizovanie školení, možnosti realizovania integrácií, nové trendy a iné)
  - Sekcia pre informovanie o bežiacich projektoch na jednotlivých OVM v oblasti dátovej kvality s priestorom vytvoreným na spoluprácu a zdieľanie znalostí a skúseností dátových kurátorov

Vytvorenie takéhoto zázemia a prostriedkov pre komunikáciu a spoluprácu zainteresovaných strán (MIRRI, OVM, odborníci z praxe a iní) je kľúčové pre rozvoj v oblasti dátovej kvality, využívanie a rozširovanie existujúcich služieb dátovej kvality. Udržiavanie pravidelnej komunikácie a usmerňovanie dátových kurátorov by mohlo priniesť väčší záujem o efektívnejšie využívanie služieb dátovej kvality.

Poskytnutie používateľských príručiek

Je dôležité dátovým kurátorom poskytnúť zrozumiteľné návody v dostatočnom detaile pre všetky dostupné služby a činnosti s nimi súvisiace. Príprave návodov služieb dátovej kvality sa venuje kapitola 3.2, ktorá navrhuje šablónu pre príručky dátových služieb tak, aby poskytovala dostatočné množstvo informácii v zrozumiteľnej forme. Keďže existujú rôzne druhy konzumentov príručiek je potrebné rozlišovať niekoľko typov príručiek a podľa toho aj prispôsobovať ich obsah.

Typy príručiek, ktorých vytvorenie je dobré zohľadniť:

- **Návod na použitie alebo Používateľská príručka:** Návod na použitie (Používateľská príručka) je typ príručky, ktorý poskytuje základné pokyny na používanie služby určeným spôsobom, prípadne metodické usmernenia.
- **Servisná príručka:** príručka, ktorá popisuje, ako sa starať o softvér v rôznych fázach životného cyklu.
- **Technický popis / príručka:** technické dokumenty, ktoré informujú o správnom používaní alebo prevádzke v značnej miere detailu.
- **Školiaca príručka:** Príručka sprevádzajúca školenie, zachytávajúca obsah školenia ako podporný materiál.

Tento projekt je podporený z Európskeho sociálneho fondu.

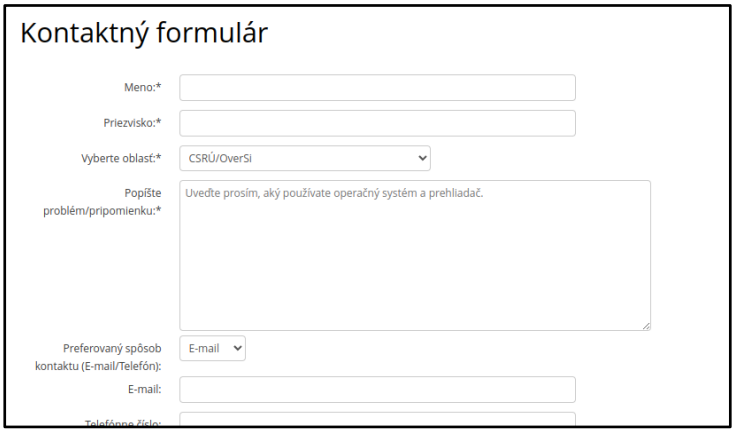
Poskytnutie konzultácii a technickej podpory

Dôležitým aspektom riadenia podpory využívania nástrojov je aj poskytnutie dostačujúcej technickej podpory a poradenstva v prípade akýchkoľvek vzniknutých problémov. V súčasnosti je poskytovanie konzultácii rozdelené do viacerých kontaktných bodov podľa typu požiadavky, ktorú dátový kurátor OVM potrebuje riešiť.

Existujúce komunikačné kanály podpory:

- Za podporu prevádzky CSRÚ nesie zodpovednosť NASES, logovanie požiadaviek bežnej prevádzkovej podpory je realizované prostredníctvom stránky Servicedesk[2]. V kontaktom formulári na stránke je potrebné zvoliť v kategórii "Vyberte oblasť" možnosť "CSRÚ/OverSi" a následne popísať vzniknutý problém.

Obrázok 8 – Kontaktný formulár NASES



- Za podporu v oblasti realizovania zmien, úpravu integračných štruktúr a podobne je zodpovedná MIRRI *(v súčasnosti konkrétne pán Michal Pažítka - michal.pazitka@mirri.gov.sk)*.

Kontaktné osoby v rámci komunikačných kanálov by mali byť zodpovedné za poskytovanie informácii a navigovanie dátových kurátorov a následné eskalácie technických problémov v prípade ich vzniku, na čo by mali byť patrične vyškolené.

Školenie pre dátových kurátorov OVM

V rámci projektu dátovej integrácie pri vzniku služieb dátovej kvality boli realizované školenia na používanie týchto služieb. Tieto školenia boli realizované MIRRI (Dátovou kanceláriou) za spolupráci dodávateľov projektu DI pre dátových kurátorov OVM, ktorí prejavili záujem. Novým dátovým kurátorom alebo kurátorom OVM, ktorí sa iniciálnych školení nezúčastnili je potrebné sprostredkovať možnosť získať takéto školenia v určitej frekvencii pokiaľ takýto záujem z ich strany bude existovať. Alternatívou je i zverejnenie video záznamov z týchto školení a ich následná pravidelná aktualizácia.

[2] https://helpdesk.slovensko.sk/new-incident/

Tento projekt je podporený z Európskeho sociálneho fondu.

Dátovým kurátorom by mali byť ponúknuté rôzne druhy školení podľa ich aktuálnych znalostí a potrieb vyplývajúcich z využívania služieb dátovej kvality. Po zohľadnení tejto skutočnosti je možné rozdeliť školenia do týchto kategórií:

- Školenia pre dátových kurátorov, ktorí už skoleniami v minulosti prešli. Školenia pre túto cieľovú skupinu by mali byť zamerané na:
  - Rozdielové školenia - zamerané na služby dátovej kvality pri realizovaní zmien,
  - Školenia zamerané na nové služby dátovej kvality zavádzané v budúcnosti (napr. OpenMetadata)
  - Všeobecné školenia zamerané na zvyšovanie dátovej kvality vo verejnej správe
- Súbor školení pre nových dátových kurátorov by mal obsahovať tieto školenia:
  - Školenia zamerané na existujúce služby dátovej kvality
  - Metodické školenia (riadenie dát, meranie kvality, zvyšovanie kvality dát a pod.)

3.1.1 Riadenie sprístupňovania príručiek, návodov a iných zdrojov

Ako bolo v kapitole vyššie spomenuté v odseku *„Poskytnutie používateľských príručiek“,* je potrebné sprístupňovať rôzne druhy príručiek. Aby príručky boli využívané efektívne, je dobré mať stanovenú vhodnú stratégiu na ich sprístupňovanie a správu.

Systém sprístupňovania používateľských príručiek, návodov a iných zdrojov informácií je efektívne vykonávaný pokiaľ zohľadňuje a spĺňa tieto body:

- Príručky a návody a iné zdroje informácií sú sprístupnené z jedného miesta a na toto umiestnenie je odkazované zo všetkých relevantných lokácií.
- Miesto, kde sú používateľské príručky umiestnené by malo byť prehľadné a všetky príručky, návody a iné znalostné zdroje by mali byť jasne pomenované, aby referovali na to, čo sa v nich nachádza
- Existuje poverená osoba, ktorá má za úlohu, kontrolovať, aktualizovať a sprístupňovať používateľské príručky konečným používateľom.
- Príručky, návody a ostatné zdroje sú pravidelne kontrolované a aktualizované pri zmenách alebo zistených nezrovnalostiach
- Existuje systém zberu spätnej väzby a poverená osoba spätnú väzbu pravidelne vyhodnocuje a v prípade potreby vykonáva potrebné úpravy
- Pri zmenách v používateľských príručkách je potrebné odlišovať jednotlivé verzie príručiek a ponechať dostupné aj ich staršie verzie po dostatočne dlhú dobu.

- Je dôležité pre používateľov uviesť jedno kontaktné miesto, ktoré pri vzniku akýchkoľvek problémov pri sprístupňovaní príručiek, s obsahom príručiek alebo akýmkoľvek iným pridruženým problémom budú môcť kontaktovať a získať pomoc pri riešení.

3.2 Odporúčania a požiadavky na príručky pre služby dátovej kvality

Táto kapitola definuje informácie, ktorých obsiahnutie v používateľských príručkách služieb dátovej kvality je považované za dôležité. Avšak pri písaní príručiek a obsiahnutí odporúčaných informácii je dôležité dodržiavať aj súbor všeobecných základných odporúčaní vzťahujúcich sa na písanie používateľských príručiek.

- **Zrozumiteľné formulácie –** Nepoužívať v používateľských príručkách žargón alebo skratky, ktoré nie sú vysvetlené. Tieto jazykové voľby sťažujú používanie používateľských príručiek a zabraňujú ich efektívnemu používaniu. Je dôležité sa zamerať na jednoduché formulácie viet, ktoré sú zrozumiteľné a skratky alebo nejasné výrazy musia byť dostatočne vysvetlené.

- **Vizuálne príklady –** Je dôležité zahrnúť vizuálne prvky v používateľských príručkách, keďže častokrát je jednoduchšie pochopiť, ako niečo funguje z vysvetlenia na vizuálnom prvku, ako z textu samotného. Zároveň slúžia na vizuálne delenie textu, cez ktorý sa potom používateľovi ľahšie naviguje.

- **Logická nadväznosť –** Využitie jasnej hierarchickej štruktúry nadpisov a podnadpisov, aby používateľ pochopil, čo sa nachádza v každej časti používateľskej príručky. Je dôležité začínať všeobecnými základmi a až následne prejsť do vysvetlenia viac špecifických a komplikovanejších častí.

- **Prepojenie na všetky ostatné relevantné dokumenty a užívateľské príručky** V častiach používateľskej dokumentácie, kde dochádza k prelínaniu sa s inou dokumentáciu alebo iná dokumentácia poskytuje detailnejšie informácie, ktoré nie sú predmetom používateľskej príručky, je dôležité sa vhodným spôsobom odkazovať na túto dokumentáciu.

Špecifikácia dizajnu:

- Prehľad: Táto časť poskytuje prehľad o službách dátovej kvality a ich hlavných cieľoch v rámci udržiavania presnosti, spoľahlivosti a konzistentnosti údajov v informačnom ekosystéme organizácie.

- Účel: je potrebné uviesť účel služby dátovej kvality a zdôrazniť jej úlohu pri zvyšovaní integrity údajov, uľahčovaní rozhodovania a zlepšovaní celkovej výkonnosti OVM.

- Zdroj vstupných údajov: uvedú sa typy a formáty zdrojov vstupných údajov, ktoré dokáže spracovať dokumentovaná služba (XML, XSD, CSV alebo iné relevantné formáty údajov, metaúdajov a pod.).

- Profilovanie údajov: opíšu sa štatistické analýzy a techniky profilovania používané na posúdenie aspektov kvality údajov, ako je úplnosť údajov, overenie

typu údajov, analýza vzorov a zisťovanie duplicitných hodnôt. Uvedie sa či ide o jednorazovú analýzu alebo opakujúci sa proces na neustále zlepšovanie.

- Overovanie biznis pravidiel: Popíše sa proces overovania správnosti špecifických biznis pravidiel týkajúcich sa atribútov údajov na základe CMÚ a RDF tvarov alebo iných relevantných zdrojov, vrátane prístupu k identifikácii biznis pravidiel pre prvky OE a prístupu k spracovaniu neprepojených atribútov.

- Čistenie údajov: Vysvetlenie, ako funguje proces čistenia údajov, a podrobný popis funkcie a vlastností konzoly GUI stevarda, ktorá umožňuje používateľom využívať službu a vykonávať čistenie údajov na základe vopred definovaných pravidiel alebo pravidiel priebežne definovaných používateľom.

- Stotožnenie údajov: Objasnenie metód používaných na automatické stotožnenie údajov na základe záznamov referenčných registrov, ako je napríklad porovnávanie totožnosti alebo podobnosti. Vysvetlí sa proces deduplikácie v rámci súborov údajov a podporované výstupné formáty (napr. CSV, XML, JSON, XLSX).

- Asistované stotožnenie: opis kritérií a techník na vyhodnotenie údajov, ktoré nebolo možné automaticky stotožniť, so zameraním najmä na "previazané" údaje a opisom dodatočných pravidiel validácie, ktoré možno v prípade potreby uplatniť prostredníctvom konzoly GUI stewarda.

- Analýza vstupného súboru údajov pre transformáciu: podrobné vysvetlenie procesu analýzy vstupného súboru údajov XML/XSD s ohľadom na jeho biznis povahu a súvisiace ontológie s cieľom zosúladiť ho s dátovým modelom transformovaného súboru údajov.

- Mapovanie ontologických modelov: Popis techník a metodík používaných na mapovanie štruktúr dátových súborov XML/XSD na ontologické modely, čím sa zabezpečuje bezproblémová integrácia údajov a lepšie pochopenie vzťahov medzi údajmi.

- Transformácia súboru údajov do formátu RDF: kroky spojené s transformáciou súboru údajov do formátu RDF, špecifikuje sa mapovanie jednotlivých prvkov pomocou jazyka XSLT a výhody použitia RDF na ich vyjadrenie.

- Nasadenie transformovaného súboru údajov vo formáte RDF:

    - Proces nasadenia: Vysvetliť postupy a požiadavky na nasadenie upraveného súboru údajov do testovacieho prostredia.

    - Testovanie a ladenie: podrobne opísať kroky testovania a ladenia procesu transformácie s cieľom zabezpečiť presnosť a spoľahlivosť údajov.

    - Overovanie a schvaľovanie: opísať proces validácie vrátane možného použitia SHACL, ak je k dispozícii a získania súhlasu na produkčné nasadenie.

    - Produkčné nasadenie: opísať záverečné kroky nasadenia transformovaného súboru údajov vo formáte RDF, čím sa zabezpečí, že bude dostupný na produkčné použitie.

- Vlastná služba DQ: Poskytnúť komplexný postup, čo je potrebné spraviť pre získanie takejto služby dátovej kvality, ktorá bude prispôsobená špecifickým potrebám a požiadavkám rôznych OVM.

Funkčná špecifikácia:

- Vstupné požiadavky: Špecifikujte špecifické požiadavky na vstupné údaje vrátane formátov súborov, štruktúr údajov a akýchkoľvek obmedzení pre bezproblémové vykonanie každej služby dátovej kvality.

- Výstupný formát: Podrobne opíšte rôzne výstupné formáty generované jednotlivými službami dátovej kvality, vrátane PDF reportov, súborov CSV, dokumentov XML, súborov JSON alebo tabuliek XLSX spolu s opisom obsahu a účelu každého výstupu.

- Konzola GUI steward: Vypracujte návod o funkciách a vlastnostiach konzoly GUI steward so zameraním na jej jednoduché používanie, aplikáciu pravidiel, možnosti prispôsobenia a na to, ako umožňuje koncovým používateľom interakciu s procesmi čistenia a stotožnenia údajov.

- Ontologické modely: Poskytnite podrobný opis ontologických modelov používaných na účely mapovania a integrácie s dôrazom na ich úlohu pri zlepšovaní kvality údajov a porozumení sémantických vzťahov.

- Validácia SHACL (ak sa uplatňuje): Ak sa používa validácia SHACL, opíšte pravidlá a proces validácie a uveďte jej význam pre zabezpečenie kvality údajov a súladu s predpismi vo fáze nasadenia.

- Proces nasadenia: Uveďte technické požiadavky a kroky potrebné na nasadenie transformovaného súboru údajov vo formáte RDF vrátane kontroly verzií, zálohovania údajov a akýchkoľvek špecifických aspektov pre bezproblémové nasadenie.

Špecifikácia prístupu a používateľského rozhrania:

- Používateľský prístup: Definujte prístupové práva a roly pre vývojárov dátových služieb a zabezpečte, aby mali potrebné oprávnenia na efektívne využívanie služieb dátovej kvality.

  - Proces sprístupnenia: Popíšte kroky žiadosti o sprístupnenie služby dátovej kvality. Popis by mal zahŕňať potrebné informácie a dokumentáciu potrebnú na vytvorenie žiadosti.

  - Proces schvaľovania: Vysvetlite, ako žiadosti o sprístupnenie kontrolujú a schvaľujú správcovia. Uveďte kritériá schvaľovania a očakávaný časový rámec spracovania žiadostí.

  - Podmienky: Uveďte všetky podmienky alebo kvalifikácie, ktoré musia používatelia splniť pred získaním prístupu k službám dátovej kvality. Môžu zahŕňať školenia, certifikáty alebo iné potrebné kvalifikácie.

  - Kontrola prístupu: Uveďte zavedené mechanizmy kontroly prístupu používateľov k citlivým údajom a funkciám. Opíšte, ako sa prostredníctvom kontroly prístupu zachováva súkromie a bezpečnosť údajov.

Tento projekt je podporený z Európskeho sociálneho fondu.

- Zásady používania hesiel: Uveďte požiadavky na heslá a zásady, ktoré musia používatelia dodržiavať pri vytváraní svojich účtov. Informácie by mali zahŕňať aspoň pravidlá týkajúce sa zložitosti hesla, vypršania platnosti a obnovenia hesla.

- Integrácia API (ak sa uplatňuje): Ak sú k dispozícii rozhrania API na integráciu služieb dátovej kvality do existujúcich systémov, uveďte koncové body, metódy overovania a formáty údajov podporovaných jednotlivými rozhraniami API.

- Systémové požiadavky: Uveďte softvérové, hardvérové a infraštruktúrne požiadavky na efektívne využívanie služieb dátovej kvality a zabezpečte, aby boli vývojári dobre vybavení potrebnými zdrojmi.

- Spracovanie údajov: Zdôraznite dôležitosť bezpečnosti údajov, ochrany súkromia a poskytnite usmernenia o tom, ako sa bude s citlivými údajmi zaobchádzať počas využívania služieb dátovej kvality a ako sa bude zachovávať dôvernosť.

Tento projekt je podporený z Európskeho sociálneho fondu.

4 Návrh pilotného zavedenia služieb dátovej kvality vo vládnom cloude

Zavedenie služieb dátovej kvality centrálne do IS CSRÚ je realizované prostredníctvom projektu CIP, ktorý rozvíja zavádzania služieb cez Platformu ako službu (Platform-as-a-Service, PaaS). V rámci tejto platformy bude sprístupnená časť služieb dátovej kvality z projektu DI a nová služba profilovania údajov tak, ako sa uvádza v kapitole 2.2. Jednou z dôležitých výhod je však prinesenie dátového katalógu, na ktorom bude do budúcna možné postaviť nové služby dátovej kvality, a tak rozvíjať a zlepšovať dátovú kvality vo verejnej správe.

4.1 OpenMetadata na CIP

OpenMetadata v projekte CIP prináša okrem nástroja na profilovania údajov aj dátový katalóg, ktorý je zásadný pre ďalší rozvoj dátovej kvality vo verejnej správe. Dátový katalóg prinesie možnosti, ako riadiť dátovú kvalitu a prinášať aj nové služby na jej zabezpečenie a zvyšovanie. Dátový katalóg OpenMetadata prinesie nasledujúce funkcie a výhody:

Zjednotené metadátové modely – Rozličné zdroje údajov sa líšia, a teda spôsob ukladania ich metadát sa taktiež líši. Pre zabezpečenie efektívnej práce s údajmi musí existovať jeden model metadát, ktorý bude využiteľný naprieč celou verejnou správou a nielen v jednom špecifickom riešení. Získaním jednotného modelu metadát prostredníctvom nasadenia OpenMetadata bude možné realizovať centralizáciu, konfiguráciu a udržiavanie rôznych konfigurácií. Sprístupnenie metadátových modelov pre tvorbu nových služieb dátovej kvality a sprístupnenie pre systémy jednotlivých OVM bude oveľa dostupnejšie a efektívnejšie.

Otvorené a štandardizované rozhrania API pre integrácie – OpenMetadata sprístupňuje pomocou API jednotný dátový model pre rôzne nadväzujúce systémy, aplikácie a služby dátovej kvality. Otvorené API, ktoré je možné využiť sa opierajú o dobre štruktúrovanú opísanú schému podľa špecifikácii JSON schém.

Rozšíriteľný model metadát – Je možné očakávať, že v budúcnosti sa z rôznych biznis dôvodov a nových zákonných povinností zmení štruktúra dát, ktorú bude potrebné uchovávať, čo následne je potrebné reflektovať v upravenej štruktúre modelu metadát. OpenMetadata tu prináša riešenie, v ktorom nie je nutné upravovať pôvodný model metadát, ale prostredníctvom funkcie rozšíreného modelu metadát umožňuje vykonávať zmeny, ktoré je potrebné zaviesť do špecifických modelov.

Prijímanie metadát založené na princípe „pull“ – Prijímanie metadát je založená na princípe pull, čo znamená, že za extrakciu metadát zodpovedá metadátový nástroj (engine) a nie zdroj údajov. Úloha extrakcie a transformácie metadát do jednotného metadátového modelu pripadá teda na nástroj na katalogizáciu údajov, podobne ako to robí nástroj ETL pri vytváraní dátových jazier a dátových skladov (data lake, data warehouse).

Ukladanie metaúdajov do grafov – Sprístupňuje centralizovaného ukladania metadát, ktoré sú "aktívne organizované ako graf spájajúci údaje" so všetkými tímami, nástrojmi a procesmi. Grafy metadát môžu využívať nadväzujúce aplikácie a umožniť tak sprístupnenie funkcií ako: katalogizácia údajov, správa údajov, lineárnosť údajov,

Tento projekt je podporený z Európskeho sociálneho fondu.

automatizované udržovanie kvality údajov a testovanie, profilovanie údajov, sledovanie údajov a mnohé iné.

Medzi ďalšie výhody patria možnosti integrácie rôznych riešení, ktoré podporujú:

- Strojové učenie
- Služby reťazeného spracovania
- Reporty a informačné panely
- Služby na prácu s metadátami

4.2 Rozvoj služieb dátovej kvality v PaaS

Rozvoj služieb v PaaS by mal byť v súčasnosti zameraný na prípravu budúceho rozšírenia, a to zlepšením funkcionalít a zjednodušením prípravy metadát a dát, ktoré je potrebné realizovať pred činnosťami zameranými na zlepšovanie dátovej kvality. Pridávanie ďalších služieb nad rámec tých, ktoré už v súčasnosti vznikli by malo byť realizované na základe biznis požiadaviek konzumentov alebo vlastníkov dát, tak ako v prípade pripravovanej služby DQ10 v rámci projektu DI (viď. kapitola 2.1.10).

Rozvoj služieb PaaS by mal byť realizovaný v týchto oblastiach, ktorých vývoj bude dostatočne podporený zavedením OpenMetadata:

- Analyzovanie metadát na základe opakujúcich sa vzorov v dátovom súbore, pre uľahčenie vykonávania činností na zlepšovanie dátovej kvality
- Umožnenie ukladania biznis pravidiel v PaaS, možnosť ich exportovania / importovania a využitia v iných systémoch
- Nástroj na prípravu biznis pravidiel v GUI
- Automatická integrácie existujúcich biznis pravidiel v dohodnutej štruktúre do OpenMetadata

Ďalšie inšpirácie ohľadom rozvoja PaaS je možné čerpať z prípadu použitia z nasledujúcej kapitoly.

Tento projekt je podporený z Európskeho sociálneho fondu.

5 Automatizácia procesov dátovej kvality

Kvalita údajov je kľúčovým aspektom digitálnej transformácie akejkoľvek vlády, pretože umožňuje lepšie rozhodovanie, poskytovanie služieb a spoluprácu. Dosiahnutie a udržanie vysokej kvality údajov však môže byť náročné, najmä v zložitých a fragmentovaných dátových prostrediach. Automatizácia môže byť preto výkonným nástrojom na zefektívnenie a zlepšenie procesov kvality údajov, ako je zber údajov, overovanie, integrácia, analýza a podávanie správ.

Automatizácia kvality údajov je jedným z kľúčových trendov a osvedčených postupov v moderných stratégiách správy údajov. Automatizácia procesov kvality údajov môže pomôcť znížiť chyby, ušetriť čas, náklady a zvýšiť hodnotu údajov pre verejné služby a tvorbu politík.

Vo vládnom segmente na celom svete existuje množstvo digitálnych iniciatív a prípadov použitia, ktoré sú úspešné najmä vďaka automatizácii procesov dátovej kvality. Ako **príklady** uvádzame:

Výber daní: Automatizácia overovania a validácie daňových priznaní, odhaľovanie chýb a podvodov a zníženie manuálneho spracovania a tým aj nákladov. Automatizované nástroje môžu pomôcť zjednodušiť proces podávania daňových priznaní pre daňovníkov a znížiť chyby alebo nezrovnalosti v daňových priznaniach alebo platbách. Napríklad v Estónsku daňový úrad používa online systém, ktorý predvypĺňa daňové priznania informáciami z rôznych zdrojov.

Zdravotná starostlivosť a sociálne služby: Automatizácia integrácie a analýzy zdravotných údajov z rôznych zdrojov, zabezpečenie presnosti a konzistentnosti údajov môže viesť k zlepšeniu výsledkov starostlivosti o pacientov a poskytovania sociálnych služieb. Automatizované nástroje môžu zlepšiť výsledky zdravotnej starostlivosti tým, že umožnia rýchlejšiu diagnostiku, lepšie odporúčania na liečbu, presnejšie účtovanie alebo efektívnejšie prideľovanie zdrojov. Napríklad v Singapure používa ministerstvo zdravotníctva umelú inteligenciu na analýzu lekárskych záznamov a nárokov na poistné plnenia s cieľom identifikovať potenciálne chyby prípadne i podvody.

Automatizácia dátovej kvality umožňuje odhaľovanie podvodov alebo anomálií vo veľkých objemoch transakcií alebo záznamov, ktoré by inak ľudskí inšpektori mohli nechať bez povšimnutia. Napríklad v Austrálii ministerstvo sociálnych vecí používa pokročilé analýzy na identifikáciu potenciálnych prípadov podvodov pri platbách sociálnych dávok.

5.1 Návrh postupov automatizácie procesov dátovej kvality

Existujú určité všeobecné kritériá na posúdenie pripravenosti organizácií na automatizáciu procesov dátovej kvality:

- Organizácia má jasnú víziu a stratégiu využívania údajov na dosiahnutie svojich cieľov.
- Je vyčlenený dostatočný rozpočet a zdroje na investovanie do potrebných nástrojov a technológií na automatizáciu procesov kvality dát.
- Je vyhradený tím alebo oddelenie zodpovedné za dohľad nad procesmi kvality údajov a ich vykonávanie.

Tento projekt je podporený z Európskeho sociálneho fondu.

- Organizácia stanovila jasné úlohy a zodpovednosti za riadenie kvality údajov naprieč rôznymi funkciami a úrovňami riadenia.
- Existuje silná kultúru dátovej gramotnosti a povedomia medzi svojimi zamestnancami a ďalšími zainteresovanými stranami.
- Sú definované merateľné ciele a ukazovatele na hodnotenie kvality svojich údajov.
- Organizácia zaviedla efektívne komunikačné kanály a mechanizmy spätnej väzby na podávanie správ a riešenie problémov s kvalitou údajov.
- Sú prijaté štandardizované formáty, protokoly a konvencie na vytváranie, ukladanie, prenos a prístup k svojim údajom.
- Je zabezpečený súlad s príslušnými zákonmi, nariadeniami, politikami a etickými normami týkajúcimi sa jej údajov.
- Organizácia podporuje mentálne nastavenie k neustálemu zlepšovaniu procesov kvality údajov.

Vyššie uvedené je možné dosiahnuť použitím nasledujúcich postupov:

- Zistite, ako zlepšená kvalita údajov ovplyvňuje rozhodnutia a fungovanie organizácie.
- Vytvorte zoznam existujúcich problémov s kvalitou údajov, ktorým organizácia čelí, a zmapujte, ako ovplyvňujú KPI.
- Definujte ukazovatele kvality údajov a ich prahové hodnoty na posúdenie a zlepšenie kvality dát v kľúčových oblastiach vrátane presnosti, relevantnosti, úplnosti, včasnosti a konzistentnosti.
- Zaveďte prísne kontroly prichádzajúcich a spracovávaných údajov.
- Zaveďte zásady a kontroly pre zdroje údajov tretích strán.
- Opravte zlyhania kvality údajov.
- Posilnite svoje tímy zavádzaním kultúry DataOps.

Pri zvažovaní automatizácie procesov by organizácie mali vykonať určité prípravné kroky a vytvoriť komplexné kontrolné / riadiace zoznamy (check lists), aby sa zabezpečil správny výber techniky a zabezpečila sa kvalita výsledku automatizácie. Zatiaľ čo tradičný prístup k dátovej kvalite obsahuje veľa pomalých, manuálnych procesov, pokročilé riešenia umožňujú posunúť efektivitu a konzistenciu na novú úroveň, ak sú správne navrhnuté a vyladené. Niektoré organizácie verejného sektora stále buď využívajú manuálne procesy alebo používajú jednoduchú automatizáciu, pokiaľ ide o kvalitu údajov. Vo väčšine prípadov v Európe sa však už testujú alebo aj prakticky využívajú výhody nových technológií a riešení založených na strojovom učení či všeobecne umelej inteligencii (ML/AI).

Nižšie uvedená tabuľka ukazuje, ako sa niektoré dôležité parametre časom menia v závislosti od prístupu ku kvalite údajov. Je dobré si všimnúť, že bez ohľadu na stupeň pokročilosti automatizácie, trend jednoznačne smeruje k preferencii „prevencie pred liečbou“. Inými slovami, organizácie sa snažia zabezpečiť, aby vznikali len kvalitné dáta a redukovala (ak nie priam eliminovala) sa potreba dodatočného čistenia dát po ich vzniku či integrácii. Je však zrejmé, že aj v tomto prípade treba rozlišovať typy spracovávaných dát. Vyššie uvedené poznámky sú (mali by byť) oveľa viac aplikovateľné a aplikované na referenčné údaje (referenčné registre, objekty evidencie) než je tomu v prípade „transakčných“ dát (takými sú napr. vo vyššie uvedených prípadoch lekárske záznamy či nároky na poistné plnenia).

Úroveň automatizácie	Manuálne postupy	Ľahká automatizácia	Pokročilá automatizácia
Popis	Manuálny, reaktívny proces monitorovania Problémy nie sú identifikované včas; často sú identifikované mesiace (roky) po ich vzniku Neformálne riadenie riešenia problémov	Automatizované, proaktívne monitorovanie Problémy sa identifikujú včas, často do 24 hodín Schopnosť predchádzať problémom vďaka lepšiemu monitorovaniu Formálny proces riadenia riešenia problémov	Používa sa pokročilý proces monitorovania Komplexné indikátory stavu a dashboardy Priebežné monitorovanie a upozornenia pre správcov údajov Schopnosť predchádzať problémom vďaka lepšiemu monitorovaniu Formálny proces riadenia riešenia problémov
Náklady	Rastú	Bez zmeny	Klesajú
Zapojenie expertov	Rastie	Rastie	Určitý pokles
Čas dodania	Rastie	Bez zmeny	Určitý pokles
Udržateľnosť	Klesá	Bez zmeny	Určitý rast

Pri nastavovaní očakávaní ohľadom možného nasadenia automatizácie procesov zlepšovania kvality dát v prostredí slovenského verejného sektora je dôležité vziať do úvahy pripravenosť jednotlivých organizácií na takúto zmenu. Bez potreby rozsiahleho dokazovania je možné skonštatovať, že mnoho organizácií verejnej správy na takýto krok zatiaľ nie je pripravených. Preto nižšie uvádzame návrh na automatizáciu tak tradičným spôsobom, ako i automatizáciu založenú na strojovom učení (ML).

5.1.1 Automatizácia tradičným prístupom

Tradičný prístup ku kvalite údajov sa opiera o hypotézy na definovanie a testovanie pravidiel na základe známych parametrov kvality údajov:

- Sekvenčný prístup ku „kritickým dátovým prvkom“ (CDE, critical data elements) na tvorbu biznis definícií a pravidiel dátovej kvality
- Zoznamy pravidiel dátovej kvality sú odovzdané IT na implementáciu
- Implementácia nových pravidiel sa riadi prehusteným kalendárom vývoja IT, ktorý často vedie k manuálnemu, nekontrolovanému, individualizovanému profilovaniu dát a reportovaniu o výsledkoch
- Toto síce vyhovuje pre rozsiahle IT a transformačné projekty súvisiace s údajmi, ale nedáva tímom v prvej línii alebo prevádzkovým tímom kontrolu nad ich údajmi a tým, čo definuje „dobrú kvalitu“

Organizácie čelia rôznym výzvam v rámci tradičného prístupu, keď skúšajú automatizovať procesy dátovej kvality. Medzi najdôležitejšie problémy patria:

- Nedostatok jasných štandardov a metrík kvality údajov.
- Chýbajúce štruktúrované postupy/procesy správy údajov.
- Nízka dátová gramotnosť a informovanosť.

Tento projekt je podporený z Európskeho sociálneho fondu.

- Nedostatok integrácie a synchronizácie údajov.
- Medzery v bezpečnosti údajov a ochrane súkromia.
- Podcenené zálohovanie a obnova dát.
- Nepostačujúca orchestrácia a koordinácia v roztrieštených iniciatívach, zdrojoch a službách dátovej kvality.
- Málo zažitá kultúra neustáleho zlepšovania.
- Rezistentnosť voči zmenám

S cieľom riešiť vyššie uvedené výzvy by iniciatíva na automatizáciu procesu kvality údajov mala zahŕňať nasledujúce kroky:

Stanoviť štandardy kvality údajov. Definujte kritériá a metriky, ktoré merajú kvalitu údajov, ako je presnosť, úplnosť, konzistentnosť, aktuálnosť a relevantnosť. Tieto štandardy by mali byť v súlade s cieľmi a požiadavkami organizácie.

Implementovať prísne kontroly prichádzajúcich údajov. Overte si kvalitu údajov z externých alebo interných zdrojov pred ich prijatím do dátových kanálov alebo systémov. Pomocou pravidiel a filtrov overenia údajov skontrolujte chyby, anomálie alebo nezrovnalosti v prichádzajúcich údajoch.

Vytvoriť rámec správy údajov. Priraďte úlohy a zodpovednosti za riadenie kvality údajov v celej organizácii. Definujte zásady, postupy a osvedčené prístupy na vytváranie, ukladanie, prístup a používanie údajov. Vytvorte program správy údajov, aby ste zaistili zodpovednosť a vlastníctvo kvality údajov.

Vytvoriť proces auditu údajov. Pravidelne monitorujte a merajte kvalitu údajov pomocou automatizovaných nástrojov a správ. Identifikujte a zdokumentujte hlavné príčiny problémov s kvalitou údajov a ich vplyv na výsledky a fungovanie organizácie.

Implementovať procesy čistenia dát. Používajte automatizované nástroje a techniky na opravu, štandardizáciu, obohatenie a deduplikáciu údajov. Aplikujte pravidlá na čistenie údajov a transformácie na zlepšenie presnosti, konzistencie a úplnosti údajov.

Stanoviť jediný zdroj pravdy. Konsolidujte a integrujte údaje z rôznych zdrojov a systémov do centralizovaného úložiska alebo platformy, ktorá poskytuje konzistentný a spoľahlivý pohľad na údaje. Používajte automatizované nástroje a techniky na zabezpečenie integrácie, synchronizácie a kvality údajov v rôznych systémoch.

Implementovať proces zálohovania a obnovy údajov. Chráňte integritu a dostupnosť údajov vytváraním pravidelných záloh údajov a ich ukladaním na bezpečné miesto. Použite automatizované nástroje a techniky na obnovenie údajov v prípade straty, poškodenia alebo katastrofy.

Aktualizovať a udržiavať štandardy bezpečnosti údajov. Zabezpečte dôvernosť a súkromie údajov implementáciou vhodných bezpečnostných opatrení, ako je šifrovanie, autentifikácia, autorizácia a kontrola prístupu. Používajte automatizované nástroje a techniky na zistenie a zabránenie neoprávnenému alebo škodlivému prístupu alebo úprave údajov.

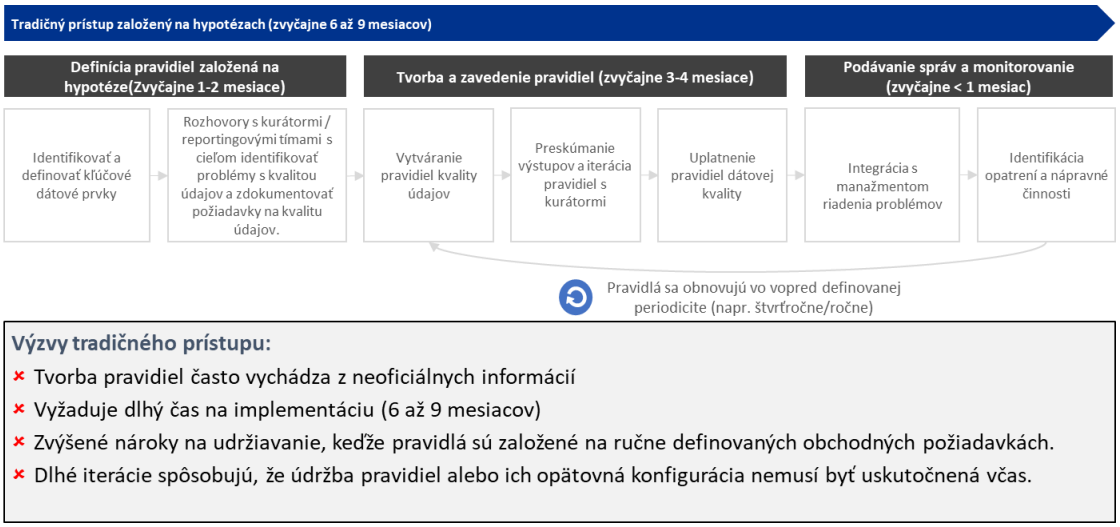
Zvážiť cloudový model pre poskytovanie služieb dátovej kvality. Používajte cloudové služby a platformy (pokiaľ to je možné z hľadiska bezpečnostných a pod. obmedzení), ktoré poskytujú škálovateľné, spoľahlivé a nákladovo efektívne riešenia na správu kvality údajov.

Tento projekt je podporený z Európskeho sociálneho fondu.

Neustále kontrolovať a zlepšovať procesy kvality údajov. Vyhodnoťte efektívnosť a efektivitu automatizovaných procesov kvality údajov pomocou spätnej väzby, metrík a benchmarkov. Identifikujte medzery, výzvy a príležitosti na zlepšenie v procesoch kvality údajov. Implementujte zmeny a vylepšenia na optimalizáciu výkonu a hodnoty údajov.

Nasledovná schéma zachytáva postup tradičného spôsobu automatizácie:

Obrázok 9 Schéma postupu tradičného spôsobu automatizácie



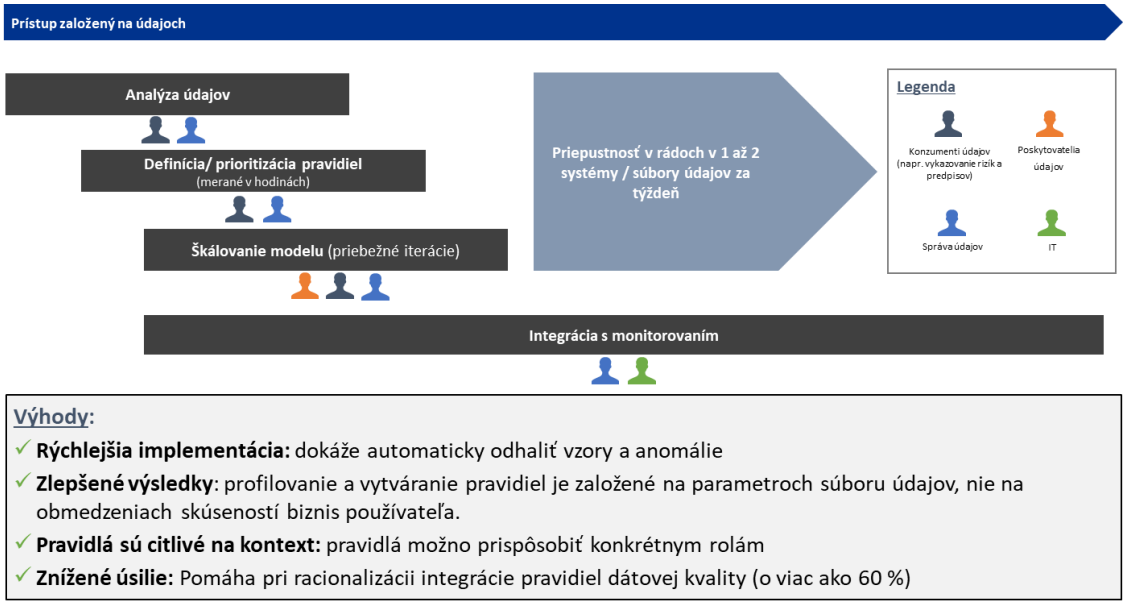
5.1.2 Automatizácia založená na strojovom učení (prístup založený na údajoch)

Jedná sa o zvyšovanie kvality údajov založené na skúmaní vlastností samotných spracovávaných údajov pri dynamickej interakcii a rozhodovaní o pravidlách, čo dramaticky skracuje čas cyklu a zároveň zlepšuje výsledky.

Prístup ku kvalite údajov založený na údajoch (data driven approach) a vychádzajúci z potrieb organizácie je charakterizovaný takto:

- Spotrebitelia (konzumenti) a tvorcovia údajov definujú a implementujú profilovanie, táto úloha teda nie je presunutá na IT oddelenia
- Používajú sa intuitívne nástroje na iteratívny spôsob definovania pracovných tokov (workflows) kvality údajov
- Pracovné postupy sa vytvárajú na základe všeobecne overených hypotéz vlastných danému odvetviu a sú založené na charakteristikách spracovávaných mnnožín údajov
- Správcovia údajov sú poučení a oprávnení vytvárať pravidlá, ktoré sa premietnu automaticky do interaktívnych reportingových dashboardov
- Využíva sa pokročilá analýza a monitorovanie kvality dát prostredníctvom automatizácie a strojového učenia

Obrázok 10 Schéma prístupu založeného na údajoch



Úspešné iniciatívy v rámci prístupu k automatizácii dátovej kvality založenom na údajoch zahŕňajú tieto kroky:

1. **Posúdenie súčasného stavu kvality údajov** v rôznych zdrojoch údajov, systémoch a oblastiach.

   Identifikujte hlavné problémy kvality údajov, ako sú chýbajúce hodnoty, duplicity, nekonzistentnosti, chyby alebo neaktuálne informácie. Používajte metriky a ukazovatele na meranie vplyvu problémov s kvalitou údajov na výkonnosť a ciele organizácie.

2. **Definícia požadovaného stavu kvality údajov** na základe vízie, poslania a cieľov organizácie.

   Stanoviť jasné a merateľné normy a kritériá kvality údajov pre každý dátový prvok, proces a výstup. Zosúladiť očakávania kvality údajov s potrebami a požiadavkami používateľov údajov a zainteresovaných strán.

3. **Stanovenie priorít iniciatív na zlepšenie kvality údajov** na základe hodnoty a uskutočniteľnosti každého projektu.

   Zvážte naliehavosť, dôležitosť, zložitosť a náklady na riešenie každého problému kvality údajov. Na porovnanie a výber najsľubnejších príležitostí na automatizáciu použite bodovací alebo klasifikačný systém.

4. **Návrh riešenia automatizácie kvality údajov** pomocou vhodných nástrojov a technológií, ako sú robotická automatizácia procesov (RPA), strojové učenie (ML), spracovanie prirodzeného jazyka (NLP) alebo umelá inteligencia (AI).

   Špecifikujte vstupy, výstupy, pracovné postupy, pravidlá a výnimky pre každú automatizovanú úlohu alebo proces.

Tento projekt je podporený z Európskeho sociálneho fondu.

5. **Vývoj a testovanie riešení automatizácie kvality údajov** v kontrolovanom prostredí pred ich nasadením do produkcie.

   Zabezpečte, aby riešenia boli spoľahlivé, presné, efektívne, bezpečné a v súlade s príslušnými normami a predpismi. Vykonajte užívateľské akceptačné testovanie (UAT) s cieľom overiť, či riešenia spĺňajú očakávania a potreby používateľov údajov a zainteresovaných strán.

6. **Nasadenie a monitorovanie riešení automatizácie kvality údajov** v prevádzkovom prostredí.

   Sledujte a merajte výkonnosť a výsledky riešení pomocou kľúčových ukazovateľov výkonnosti (KPI) a informačných panelov (dashboard). Identifikujte a promptne riešte všetky problémy alebo chyby, ktoré môžu vzniknúť počas vykonávania automatizovaných úloh alebo procesov.

7. **Pravidelné vyhodnocovanie a optimalizácia riešenia automatizácie kvality údajov,** aby sa zabezpečilo, že prinášajú očakávanú hodnotu a výhody.

   Zbierajte spätnú väzbu od používateľov údajov a zainteresovaných strán s cieľom posúdiť ich spokojnosť a návrhy na zlepšenie. Realizujte cykly neustáleho zlepšovania s cieľom aktualizovať a zdokonaľovať riešenia na základe meniacich sa potrieb a požiadaviek organizácie.

8. **Dokumentácia a komunikácia riešenia automatizácie kvality údajov** s cieľom zvýšiť ich viditeľnosť a prijatie v rámci organizácie.

   Vytvorte jasnú a komplexnú dokumentáciu, ktorá opisuje účel, rozsah, funkcie a výsledky každého riešenia. Komunikujte o hodnote a osvedčených postupoch používania riešení s príslušným publikom (stakeholders) prostredníctvom rôznych kanálov, ako sú informačné bulletiny, webové semináre, workshopy alebo prípadové štúdie.

9. **Riadenie a spravovanie riešenia automatizácie kvality údajov** s cieľom zabezpečiť ich udržateľnosť a škálovateľnosť v čase.

   Stanovte úlohy a zodpovednosti za dohľad nad riešeniami a ich údržbu v rámci organizácie. Definujte zásady a postupy na riadenie zmien, incidentov, rizík alebo problémov súvisiacich s riešeniami.

10. **Rozšírenie a integrácia riešení automatizácie kvality údajov** s cieľom využiť ich plný potenciál v rôznych oblastiach a funkciách organizácie.

    Pravidelne zbierajte vyhodnocujte podnety na zlepšenie a požiadavky na ďalšie rozšírenie automatizácie, porovnajte ich s pripravenosťou organizácie / organizačných zložiek a disponibilnými zdrojmi a na základe prioritizácie ich postupne implementujte po zvládnuteľných častiach.

Bežné chyby pri pokročilých iniciatívach v oblasti automatizácie

Existuje niekoľko bežných chýb, ktoré môžu ohroziť úspech projektov automatizácie kvality údajov aj v relatívne vyspelých prostrediach:

Organizácie môžu **zanedbať vplyv automatizácie na ľudí**, ktorí údaje používajú alebo spravujú. Môžu napríklad neposkytovať primerané školenia, komunikáciu alebo mechanizmy spätnej väzby pre používateľov údajov a zainteresované strany. To môže mať za následok nízku úroveň prijatia, odpor alebo nespokojnosť s automatizačnými riešeniami

Organizácie sa môžu **príliš upnúť na konkrétny automatizačný nástroj,** ako je napr. robotická automatizácia procesov (RPA) a snažiť sa ho použiť na každý problém s kvalitou údajov. Rôzne problémy s kvalitou údajov si však môžu vyžadovať rôzne riešenia, napríklad strojové učenie, spracovanie prirodzeného jazyka alebo umelú inteligenciu.

Organizácie sa môžu pokúsiť automatizovať **príliš veľa procesov kvality údajov naraz** bez toho, aby zvážili hodnotu a uskutočniteľnosť každého projektu. To môže mať za následok plytvanie zdrojmi, nedodržanie termínov alebo zlé výsledky.

Používatelia si môžu myslieť, že na automatizáciu procesov kvality údajov môžu použiť nízkokódové alebo bezkódové aplikácie **bez zapojenia IT**. To však môže viesť k **problémom so zabezpečením údajov, správou a integráciou**, ako aj k chybám a zlyhaniam počas aktualizácií systému.

Organizácie môžu **uvažovať o automatizácii, aby zakryli medzery alebo nedostatky** vo svojich existujúcich dátových procesoch namiesto ich prepracovania alebo nahradenia. To však môže len predĺžiť životnosť neoptimálnych starších systémov a zakryť základné neefektívnosti

Organizácie **nemusia mať jasnú a konzistentnú definíciu toho, čo sa považuje za vysokokvalitné údaje** na ich účely. To môže sťažovať meranie a zlepšovanie kvality údajov, ako aj zosúladenie automatizačných riešení s obchodnými výsledkami.

Organizácie sa môžu ponáhľať so zavádzaním svojich automatizačných riešení vrátane riešení tretích **strán bez toho, aby ich riadne otestovali a vyhodnotili.** To môže viesť k chybám, zlyhaniam alebo neočakávaným výsledkom, ktoré môžu ohroziť kvalitu a integritu údajov

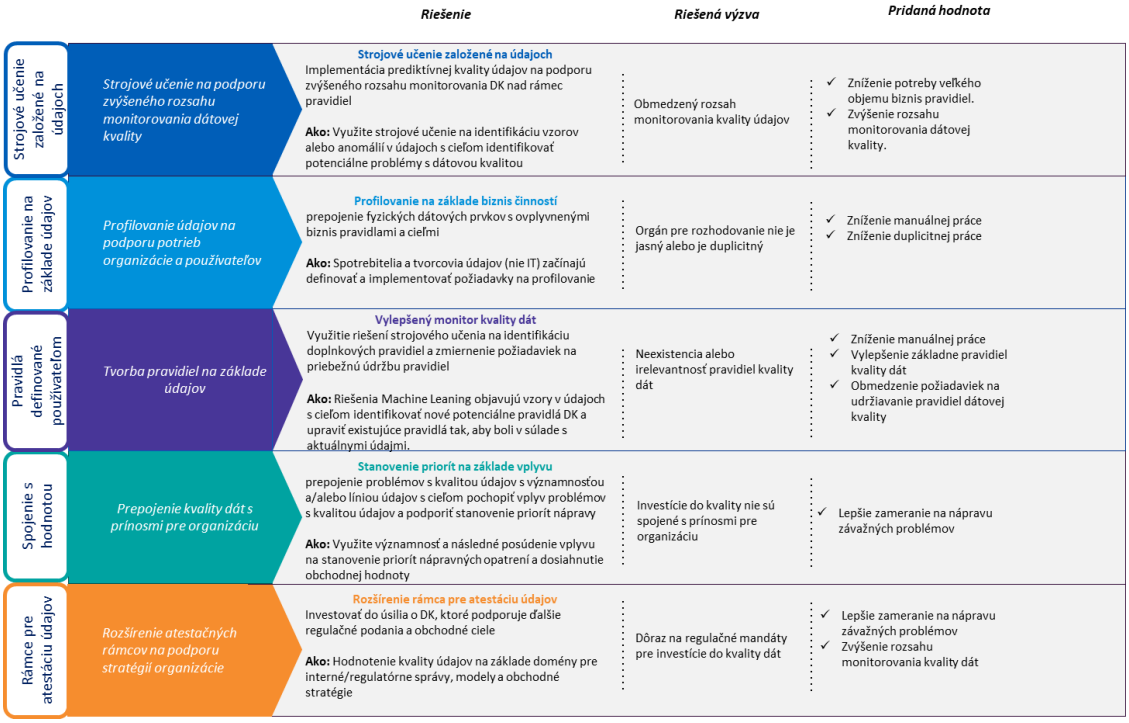
Organizácie **nemusia naplno využiť potenciál** svojich automatizačných riešení ich rozšírením alebo integráciou s inými systémami alebo doménami. To môže viesť k nevyužitiu príležitostí na zdieľanie údajov a spoluprácu.

Organizácie nemusia d**ostatočne dobre zdokumentovať alebo komunikovať svoje automatizačné riešenia**, aby sa zvýšila ich viditeľnosť a prijatie v rámci organizácie. To môže mať za následok zmätok, nedorozumenie alebo zdvojenie úsilia.

Nové riešenia/prístupy a ich prínosy pre organizácie

Nižšie sú uvedené vybrané úvahy o dôležitých dôsledkoch nových techník pre automatizáciu kvality údajov.

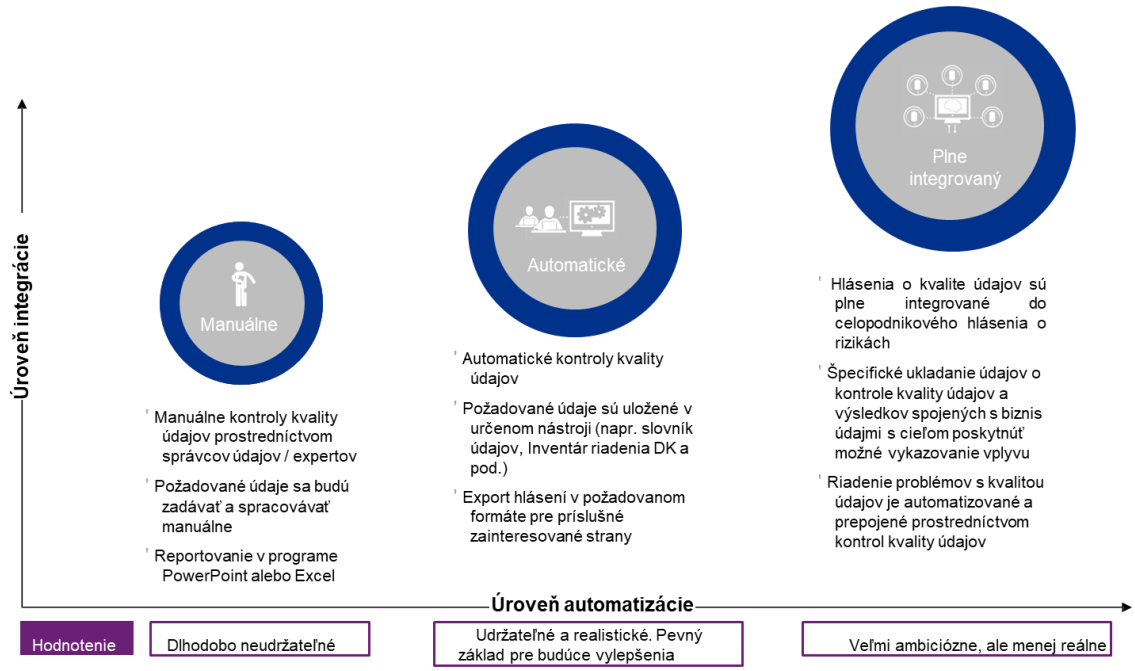
Obrázok 11 Nové techniky pre automatizáciu kvality údajov.



Rozsah automatizácie procesov dátovej kvality.

Nasledujúca schéma porovnáva jednotlivé prístupy v dimenziách 1. úroveň automatizácie a 2. úroveň integrácie. Vyplýva z neho, že vyvážený prístup podporuje individuálnu úroveň integrácie a automatizácie kvality údajov podľa špecifických potrieb organizácie.

Obrázok 12 Prístupy k automatizácii v dimenziách



Tento projekt je podporený z Európskeho sociálneho fondu.

V rámci existujúceho dátového prostredia v organizácii/OVM je dôležitosť metadát a kvality kmeňových údajov veľmi významná. Referenčná integrita je kľúčovým predpokladom automatizácie kvality údajov. Pri niektorých registroch, ktoré sa skladajú/pochádzajú z viac ako 100 zdrojov, by sa úvahy o kvalite údajov mali začať od zlepšenia/overenia pôvodných súborov údajov/DB. To je predpokladom úspešnej automatizácie procesu kvality údajov, ako aj identifikáciu najvhodnejších oblastí pre implementáciu automatizácie. Vo väčšine prípadov je veľmi ťažké riadiť/zlepšovať kvalitu údajov v cieľovom registri / cieľovej databáze, ak kvalita zdrojových údajov nie je overená v zdrojovej databáze a neexistujú formálne požiadavky, politiky a kontroly nad prichádzajúcimi údajmi, najmä nad kmeňovými údajmi.

Dôrazne sa odporúča pokračovať v úsilí o vytvorenie jednotného zdroja pravdy (referenčné údaje / registre), predovšetkým pre jednotnosť a konzistentnosť všetkých používaných kmeňových údajov. Zachytenie správnych údajov priamo pri zdroji je oveľa efektívnejšie ako čistenie/oprava cieľového registra.

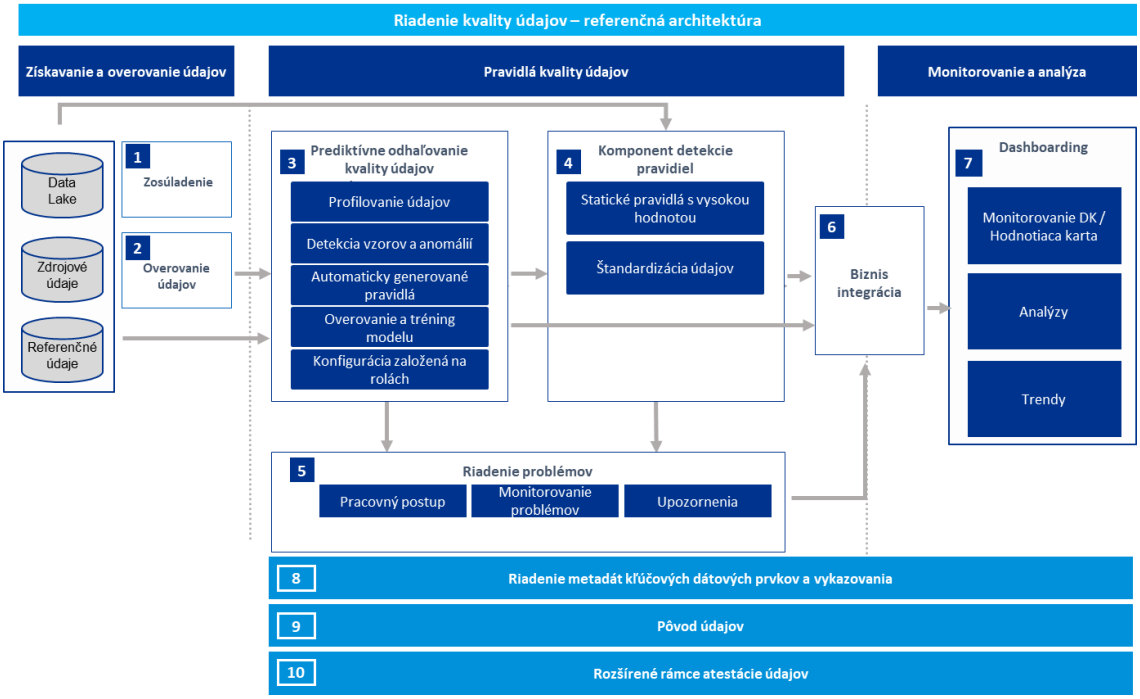
Vzhľadom na súčasný stav spracovania údajov v zapojených subjektoch/OVM a prebiehajúce projekty týkajúce sa údajov navrhujeme pokračovať v automatizácii procesov dátovej kvality v modeli PaaS/SaaS s využitím možností prebiehajúcich projektov.

Iniciatíva Dátová kvalita ako služba by mala byť doplnená / podporené kapacitami Centra Excelentnosti (CoE). Takéto kapacity sú potrebné na pilotovanie automatizačných riešení a vývoj pilotných projektov (Proof of Concept, PoC, napr. pokročilé techniky automatizácie kvality dát, ktoré môžu zahŕňať funkcie generatívnej AI/LLM na monitorovanie odľahlých hodnôt a anomálií v súboroch údajov/registroch), rozširovanie a šírenie úspešných nástrojov a osvedčených postupov. Tieto iniciatívy by malo viesť a dohliadať na ne MIRRI. Pre viac detailov odkazujeme v tejto súvislosti na výstup 1.2.1.

Prebiehajúce projekty umožnia automatizáciu vybraných úloh/procesov, najmä profilovanie údajov, kontroly konzistencie, monitorovanie dostupnosti údajov tradičným spôsobom.

Pokiaľ ide o zavádzanie nových technológií (ako napríklad AI/ML) v produktívnom prostredí, najmä v štátnych organizáciách treba brať do úvahy aj nadchádzajúce nariadenia EÚ o AI, ktoré budú schválené v najbližších mesiacoch. Nové pravidlá stanovia obmedzenia pre umelú inteligenciu v závislosti od úrovne rizika spojeného s rôznymi technológiami. Niektoré riešenia budú zakázané alebo ich využívanie výrazne obmedzené (pozri Akt EÚ o umelej inteligencii: prvé nariadenie o umelej inteligencii).

Obrázok 13 Zjednodušená referenčná architektúra riadenia kvality dát.



Na obrázku vyššie sú znázornené možné oblasti automatizácie s výhľadom do roku 2028. V iných častiach výstupov sme poskytli podrobné úvahy o súčasnom a navrhovanom stave kvality dát, prístupe k výberu automatizačných nástrojov, a kľúčových parametroch a indikátoroch kvality. Odporúčame vykonávať pravidelnú analýzu prebiehajúcich projektov automatizácie a ich výsledkov, blokátorov a výziev. Je potrebné spomenúť, že úplná automatizácia všetkých procesov súvisiacich s kvalitou dát je nereálna a je veľmi dôležité monitorovať výkonnosť nasadených riešení na základe schválených kritérií/KPI, najmä pokiaľ ide o nové technológie.

Tento projekt je podporený z Európskeho sociálneho fondu.

5.2 Vzorový prípad automatického zvyšovania kvality dát

5.2.1 Automatizácia overovania kvality dát pomocou Great Expectations v Urban Institute

Cieľom Centra pre politiku financovania bývania Urban Institute[3] je poskytovať prísne a spoľahlivé politické analýzy s využitím viacerých zdrojov údajov z evidencie nehnuteľností. Na podporu tímu pre politiku financovania bývania a ďalších výskumníkov zaviedol tím dátovej vedy v Urban Institute robustný systém zabezpečenia kvality údajov, ktorý automatizuje, iteruje a urýchľuje kontroly kvality[4].

Výzvy:

- Manuálne čistenie údajov je časovo náročné a náchylné na chyby, čo vedie k neefektívnosti projektových cyklov.
- Nedostatok automatizovaných kontrol kvality údajov môže viesť k nekonzistentnej a nespoľahlivej analýze údajov.
- Dátoví vedci musia stráviť veľa času čistením údajov a ručným kódovaním jednorazových kontrol kvality.

Riešenie:

Mestský inštitút vyvinul komplexný systém kvality údajov, ktorý pozostáva zo štyroch hlavných zložiek:

1. Infraštruktúra dátových potrubí: Výskumníci napíšu kód na generovanie dátových súborov, pošlú ho na GitHub a po schválení sa spustí na miestnom serveri. Dátové súbory sa vygenerujú a odošlú do vedra S3, čím sa spustí automatizovaná cloudová pipeline na kontrolu kvality.
2. Systém metaúdajov: Systém metadát pomáha definovať premenné, hodnoty premenných a typy údajov, stanovuje pravidlá pre vzhľad údajov a uľahčuje identifikáciu chýb a spoluprácu výskumníkov.
3. Great Expectations: Great Expectations je open-source nástroj na kontrolu kvality údajov, ponúka automatizované jednotkové testy (pipeline testing) na kontrolu súborov údajov podľa definovaných očakávaní kvality údajov. Výskumníci môžu konfigurovať rôzne očakávania, ako sú názvy stĺpcov, chýbajúce bunky, typy údajov, očakávané hodnoty a štatistické charakteristiky. Nástroj poskytuje viac ako 50 zabudovaných očakávaní a umožňuje používať vlastné funkcie očakávaní.
4. Front-End webová stránka: Na webovej stránke systému sú uvedené všetky dostupné súbory údajov spolu s históriou ich kontroly kvality a správami. Výskumníci si môžu jednoducho prezerať a zdieľať správy.

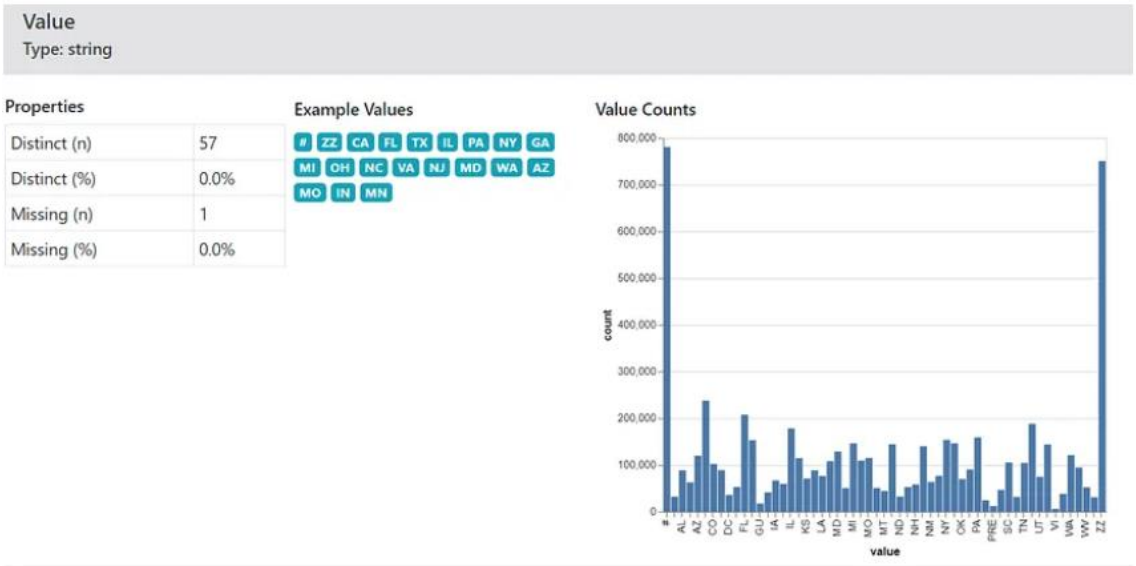
[3] O Urban Institute | Urban Institute
[4] Automatizácia kontrol kvality údajov s GreatExpectations | podľa Data@Urban | Médium

Tento projekt je podporený z Európskeho sociálneho fondu.

Výhody Great Expectations:

- Kompatibilita: Great Expectations sú použiteľné pre rôzne formy údajov a môžu sa používať s nástrojmi a systémami, ako sú Pandas, Spark data frames, S3, MySQL a PostgreSQL.
- Komplexné jednotkové testy: Výskumníci môžu konfigurovať rôzne očakávania na overenie vlastností údajov a zabezpečiť tak ich kvalitu.
- Jednoduché používanie: Po nastavení sa kontroly vykonávajú automaticky vždy, keď sú k dispozícii nové údaje.
- Dodanie výsledkov: Great Expectations generuje správy HTML s podrobnými informáciami, čo zjednodušuje ladenie. Môže tiež zverejňovať súhrnné výsledky kontroly na kanáloch Slack.
- Funkcia profilovania: Great Expectations dokáže automaticky generovať súhrnné štatistiky pre rýchly prehľad údajov. Nižšie je znázornený príklad správy o profilovaní Great Expectations.

Obrázok 14 Príklad správy z profilovaní v Great Expectations



Implementácia:

Systém kontroly kvality údajov v Urban Institute vyžaduje tri hlavné prvky: súbor očakávaní (vo formáte JSON), údaje uložené v S3 a kontext údajov nakonfigurovaný prostredníctvom súboru great_expectations.yml.

Tento projekt je podporený z Európskeho sociálneho fondu.

Výsledky:

Automatizovaný systém kontroly kvality údajov zabezpečuje integritu údajov, zefektívňuje následnú analýzu politík a znižuje počet manuálnych zásahov. Urban Institute pracuje na tom, aby bol systém použiteľný aj pre iné súbory údajov, čo by prinieslo prospech viacerým používateľom údajov a politickej práci v rámci organizácie.

Záverom možno konštatovať, že zavedením programu Great Expectations a automatizáciou kontrol kvality údajov Urban Institute zlepšil správu údajov, skrátil čas potrebný na ich čistenie a zvýšil integritu údajov na účely analýzy politík a výskumných činností.

5.2.2 Aplikácia na situáciu MIRRI

Aj keď sa príklad Urban Institute na zlepšenie kvality údajov prostredníctvom automatizovaných procesov môže líšiť od Centrálnej integračnej platformy (CIP) v MIRRI na Slovensku (rozsah, zložitosť, správa údajov), stále existuje niekoľko spôsobov, ako možno zásady a techniky nimi uplatňované aplikovať na CIP:

1. Automatizované kontroly kvality údajov: Cenným príkladom pre CIP môže byť využívanie systému Great Expectations na automatizované kontroly kvality údajov. Great Expectations je výkonný open-source nástroj na kvalitu údajov založený na jazyku Python, ktorý možno integrovať do prostredia CIP "Platforma ako služba" (PaaS). Vytvorením grafického rozhrania na profilovanie údajov a generovanie správ o kvalite údajov môže CIP umožniť správcom údajov a výskumným pracovníkom jednoducho posúdiť kvalitu svojich údajov. Táto automatizácia zníži manuálne úsilie a zvýši efektívnosť procesov zabezpečenia kvality údajov v rámci platformy CIP. Podobne ako Urban Institute využíva Great Expectations na automatizované jednotkové testy a profilovanie, CIP môže využívať pokročilé techniky profilovania na identifikáciu potenciálnych anomálií údajov, chýbajúcich hodnôt a štatistických charakteristík, čo umožní efektívne čistenie a validáciu údajov.

2. Efektivita a úspora času: Automatizované kontroly kvality údajov, ako ukázal Urban Institute, môžu viesť k výraznému zvýšeniu efektívnosti a úspore času projektu CIP. Automatizácia procesov kvality údajov umožňuje overovať prichádzajúce údaje v reálnom čase, čím sa zabezpečí, že sa budú ukladať a používať na analýzu len kvalitné údaje. To zefektívňuje proces validácie údajov, čím sa znižuje čas strávený manuálnym čistením a validáciou údajov. V dôsledku toho sa môžu správcovia údajov a výskumní pracovníci viac sústrediť na analýzu údajov a rozvoj politiky.

3. Správa a štandardizácia údajov: Systém metadát Urban Institute zavádza jasné riadenie údajov definovaním premenných, typov údajov a očakávaní kvality. Podobne môže projekt CIP vytvoriť a presadzovať štandardy správy údajov v rôznych zdrojoch údajov z rôznych OVM. To zahŕňa definovanie štandardizovaných formátov údajov, konvencií pomenovania a pravidiel kvality údajov. Zavedenie a najmä dôsledné presadzovanie štandardizácie údajov v rámci platformy CIP (aj s väzbou na Centrálny model údajov, CMÚ) zabezpečí, že všetky údaje budú dodržiavať konzistentné normy kvality, čím sa zvýši integrita údajov a zníži riziko chýb.

Tento projekt je podporený z Európskeho sociálneho fondu.

4. Monitorovanie a podávanie správ: V prípade Urban Institute sa uprednostňuje monitorovanie a vykazovanie zlepšenia kvality údajov. Služba kvality údajov CIP s grafickým rozhraním pre správy o kvalite údajov môže vytvárať pravidelné správy o stave a zmenách kvality údajov. Tieto správy poskytnú správcom údajov a výskumníkom cenné informácie o stave ich údajov. Projekt CIP môže zaviesť automatizované mechanizmy podávania správ, ktoré budú zainteresované strany informovať o akýchkoľvek problémoch alebo anomáliách v kvalite údajov, čo uľahčí rýchlu nápravu.

5. Prispôsobenie a integrácia: Prispôsobenie očakávaní Urban Institute poukazuje na dôležitosť prispôsobenia kontrol kvality údajov konkrétnym požiadavkám. Podobne aj projekt CIP môže prispôsobiť kontroly kvality údajov na základe jedinečných potrieb rôznych OVM, konzumentov i zdrojov údajov. Umožnením vlastných funkcií a pravidiel pre kontroly kvality údajov môže CIP vyhovieť rôznym formátom a štruktúram údajov z rôznych zdrojov. Integrácia riadenia kvality údajov do existujúcej platformy na integráciu údajov zabezpečuje bezproblémovú validáciu údajov ako súčasť procesu toku údajov.

6. Centralizácia a spolupráca: Centrálna integračná platforma projektu CIP konsoliduje údaje z viacerých zdrojov, čím podporuje spoluprácu a zdieľanie údajov. Tento centralizovaný prístup je v súlade so systémom metadát Urban Institute, ktorý podporuje spoluprácu medzi výskumníkmi definovaním štandardov a očakávaní v oblasti údajov. Centralizáciou riadenia kvality údajov v rámci platformy CIP majú všetky zainteresované strany prístup k štandardizovanému a jednotnému prostrediu kvality údajov. Spolupráca medzi OVM sa zefektívňuje, čo vedie k zlepšeniu konzistentnosti a kvality údajov.

Na záver možno konštatovať, že prípad využitia Urban Institute na zlepšenie kvality údajov prostredníctvom automatizovaných procesov poskytuje cenné poznatky a techniky uplatniteľné na projekt CIP v MIRRI. Prijatím automatizovaných kontrol kvality údajov, dôrazom na profilovanie a analýzu údajov a podporou správy a štandardizácie údajov môže CIP dosiahnuť zlepšenie kvality údajov, efektívnosti a spolupráce medzi vládnymi agentúrami. Zavedenie takýchto postupov v oblasti kvality údajov prispeje k presnejším a spoľahlivejším údajom na účely analýzy politík a rozhodovania v podmienkach slovenskej verejnej správy.

Tento projekt je podporený z Európskeho sociálneho fondu.

Error! Reference

Kontaktujte nás

Rudolf Sedmina
partner

E rsedmina@kpmg.sk

Some or all of the services described herein may not be permissible for KPMG audit clients and their affiliates or related entities.

www.kpmg.com

Tento projekt je podporený z Európskeho sociálneho fondu.