KPMG

Dátová kvalita vo verejnej správe

Koncept riadenia, merania a zlepšovania kvality dát

ITAPA 2019, Martin Florián

28.5.2019

Obsah

KPMG

1. Prečo Dátová kvalita?

Prečo riešime dátovú kvalitu?

Prečo riešime dátovú kvalitu?

Cieľom zlepšovania dátovej kvality nie je dátová kvalita sama o sebe, ale **vyššia kvalita života a lepšie kvalitnejšie služby verejnosti** - ktoré sú silno prepojené a závislé práve na primeranej dátovej kvalite. Dátová kvalita nie je abstraktný izolovaný cieľ, ale cesta a základňa na zvyšovanie kvality služieb, produktov a komunikácie s ňou spojených.

Aký je zmysel dátovej kvality?

Dáta sú realita a holé fakty. Čím kvalitnejšie dáta máme, tým **hodnovernejšie sú informácie a znalosti z nich odvodené**. Cieľom je nielen zvýšenie prvotných ukazovateľov dátovej kvality, ako sú presnosť, aktuálnosť a úplnosť dát, ale aj ich celková konzistentnosť, vierohodnosť, použiteľnosť, spoľahlivosť a integrita.

Prečo je potrebné meranie dátovej kvality?

Prečo meriame dátovú kvalitu?

To, čo chceme systematicky zlepšovať, musíme aj nejakým spôsobom **vyhodnocovať merať a kvantifikovať**. Aby sme mohli **objektivizovať** súčasný aj plánovaný stav a vedeli čo najobjektívnejšie sledovať trendy a dynamiku vývoja.

Prečo práve teraz?

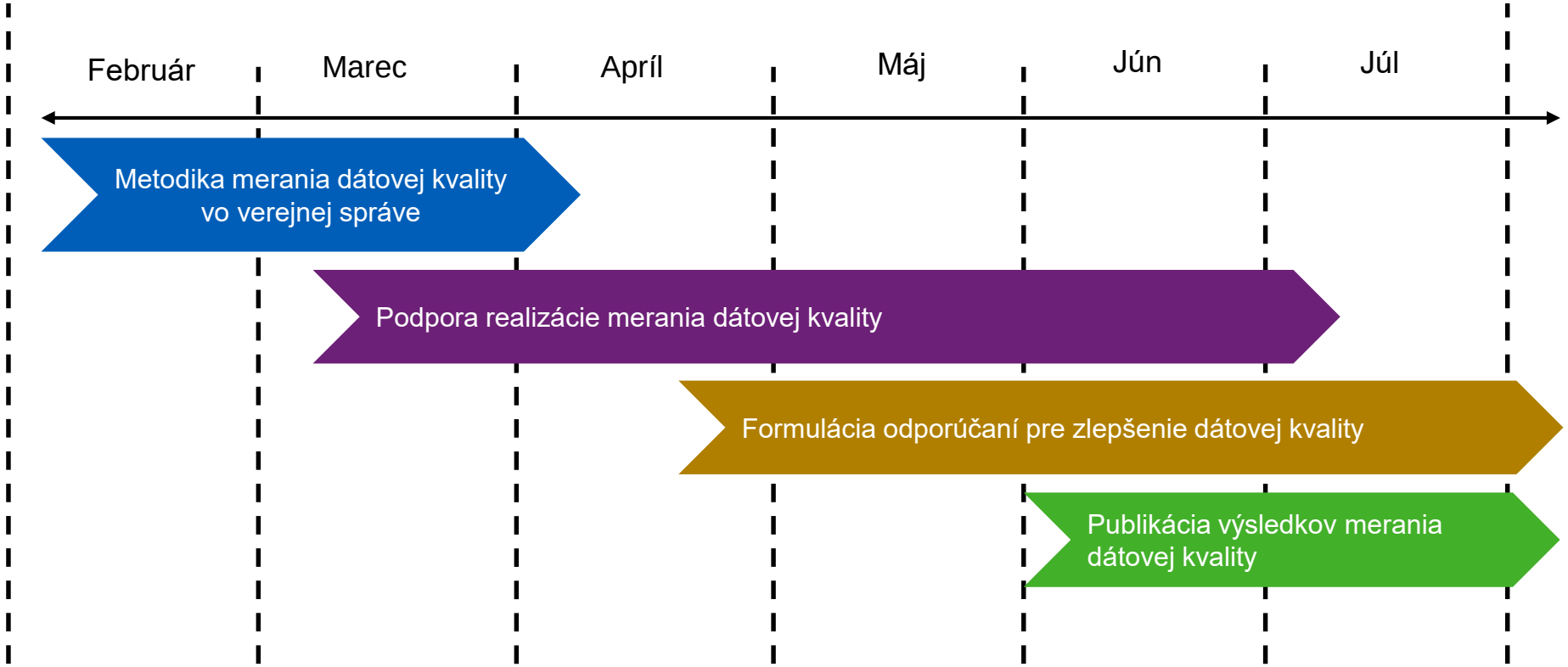
Dátová kvalita posledné roky získava stále väčšiu pozornosť a dôležitosť pre **exponenciálny nárast objemu štruktúrovaných aj neštruktúrovaných dát** a pre práve prebiehajúcu **digitálnu transformáciu**. V tomto prostredí už nestačí doteraz obvyklý intuitívny subjektívny ad hoc prístup ku kvalite dát, ale je nevyhnutné zaviesť **koncepčný prístup založený na overenej metodike**.

KPMG

2. Fázovanie projektu

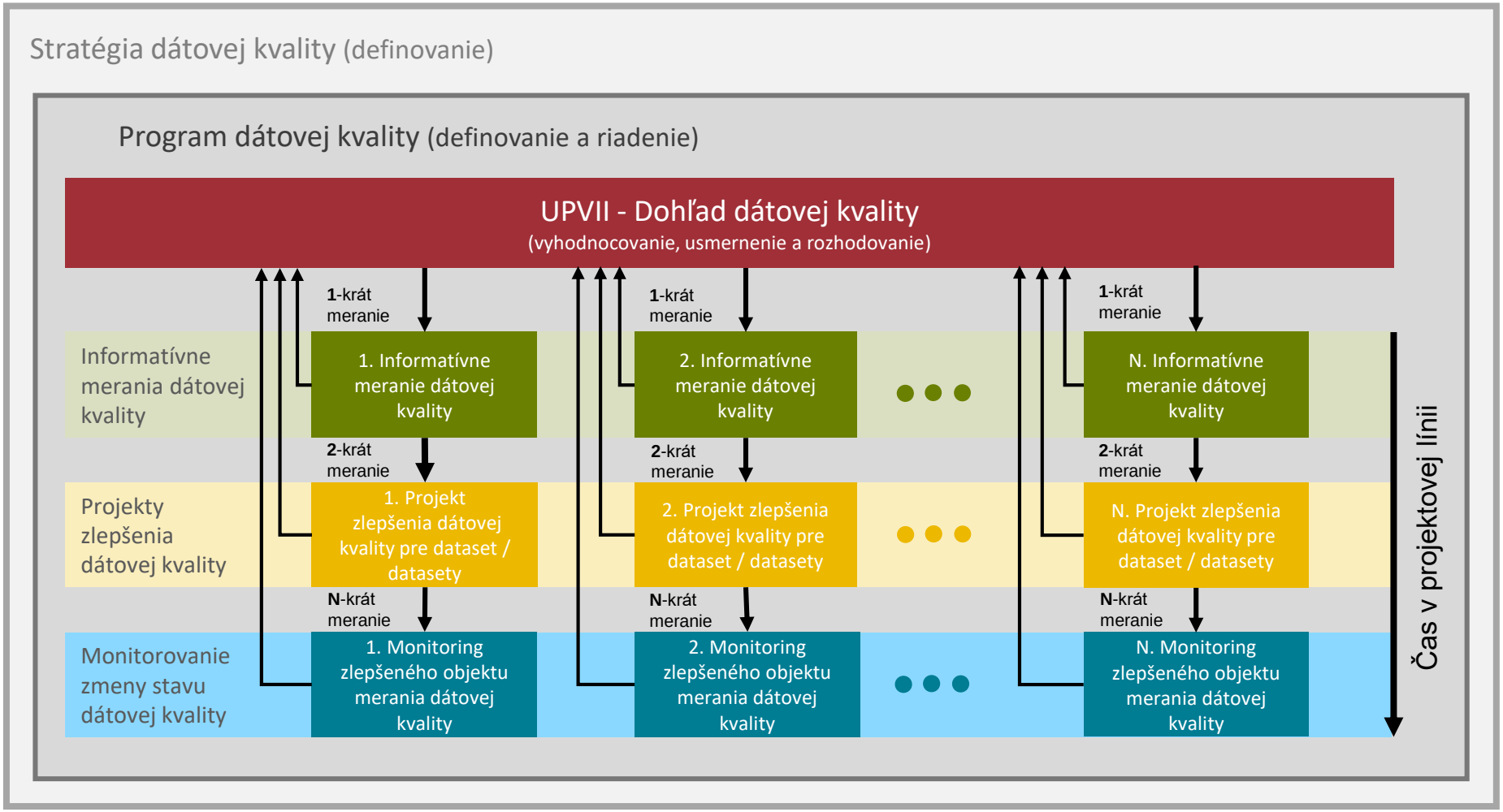
Fázovanie projektu merania dátovej kvality vo verejnej správe

Časové obdobie projektu: 8.2. - 7.8.2019



3. Náš prístup k riadeniu dátovej kvality

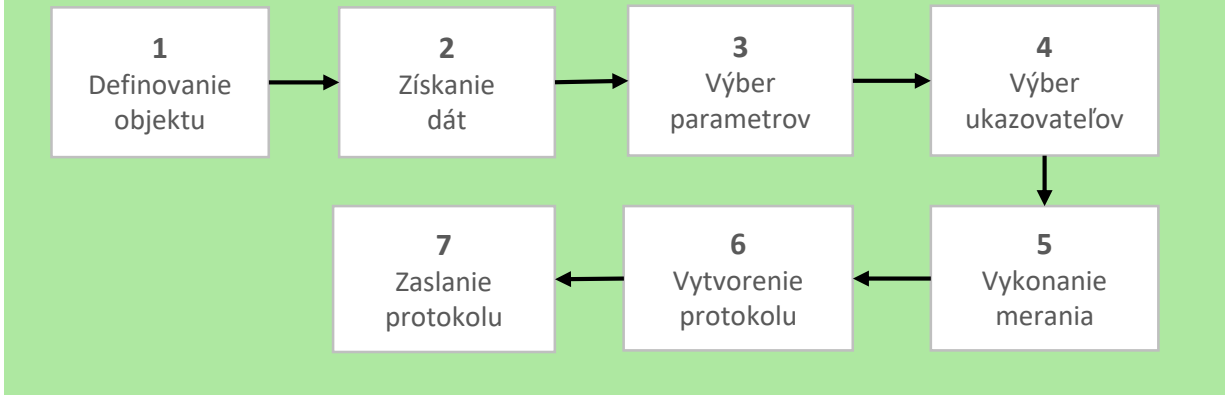
Zavádzanie dátovej kvality vo verejnej správe



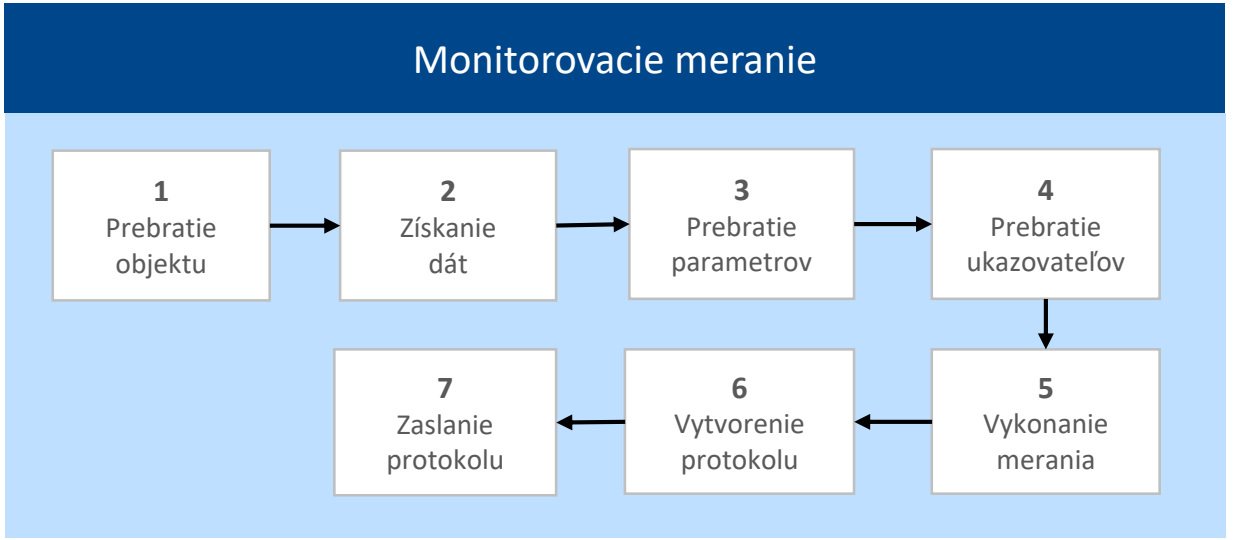
4. Náš prístup k meraniu a monitorovaniu dátovej kvality

7-krokový postup merania a monitorovania DK

Informatívne meranie



Monitorovacie meranie



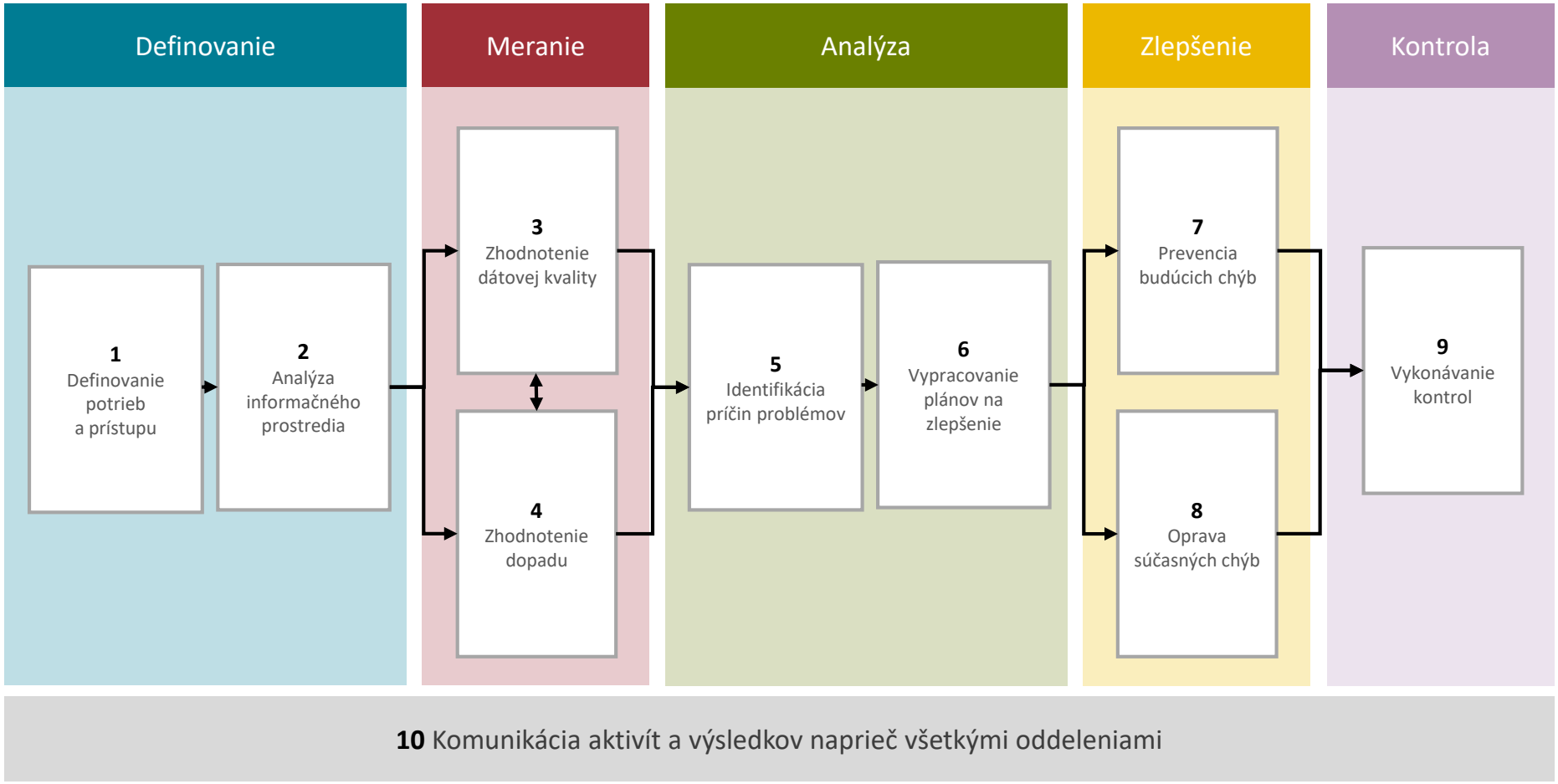
7-krokový postup merania a monitorovania DK

Komplexné meranie súčasného stavu

Kontrolne meranie po implementácii zlepšení

5. Náš prístup k zlepšeniu dátovej kvality

10-krokový postup zlepšenia dátovej kvality



KPMG

6. Parametre dátovej kvality

Parametre dátovej kvality

Parameter	Popis
1. Presnosť	Ako dobre dáta zodpovedajú skutočnostiam v reálnom svete
2. Správnosť	Hovorí, či formát údajov zodpovedá definovaným pravidlám
3. Kompletnosť	Hovorí, či dáta obsahujú dostatok informácií
4. Unikátnosť	Hovorí o duplicitách v záznamoch
5. Aktuálnosť	Hovorí, či sú dáta aktuálne alebo staršieho dátumu
6.Strojová spracovateľnosť	Hovorí, ako jednoducho je možné údaje automatizovane spracovať
7. Referenčná integrita	Zameriava sa na použitie dát naprieč verejnou správou a previazanie údajov v rámci aj mimo nej pre ďalších konzumentov údajov

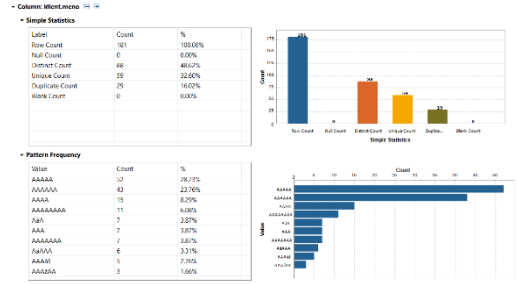
7. Príklad merania

Unikátnosť - unikátnosť atribútu

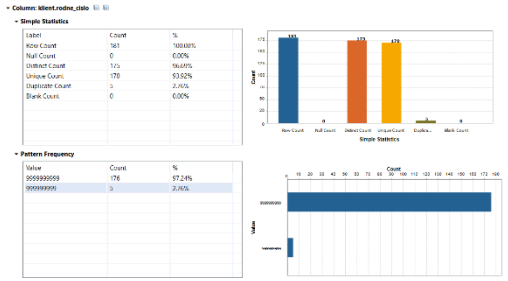
Názov ukazovateľa	Unikátnosť atribútu
Definícia	Nič sa nezaznamená viac ako 1 krát na základe toho, ako sa vec identifikuje
Metrika	Analýza počtu vecí hodnotených v reálnom svete v porovnávaní s počtom záznamov v datasete
Rozsah	Meraný voči všetkým záznamom v rámci jedného datasetu
Jednotka metriky	Percento
Príbuzné parametre	Konzistencia
Voliteľnosť	Záleží na okolnostiach
Príklad	Škola ma 120 súčasných študentov a 380 bývalých študentov (t.j. celkovo 500). Avšak databáza študenta ukazuje 520 odlišných študentských záznamov. Môže obsahovať meno Jožo Mrkvička a Jožko Mrkvička ako samostatný záznam aj napriek skutočnosti, že v škole existuje len 1 študent menom Jožko Mrkvička. To znamená, že unikátnosť je 500/520 = 96,2%.
Pseudokód (výpočtový vzorec)	(Number of things in real world) / (Number of records describing different things)

Register Úpadcov - Klient

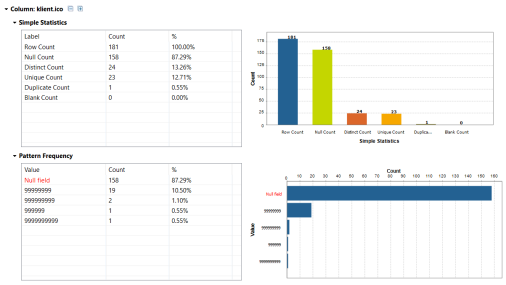
1. Meno



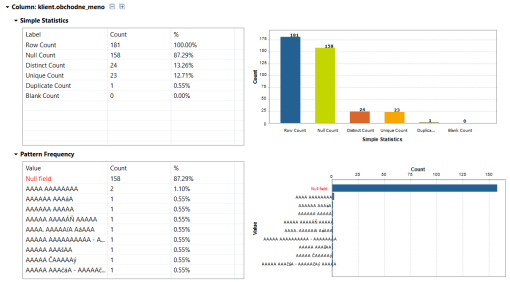
2. Rodné číslo



3. IČO

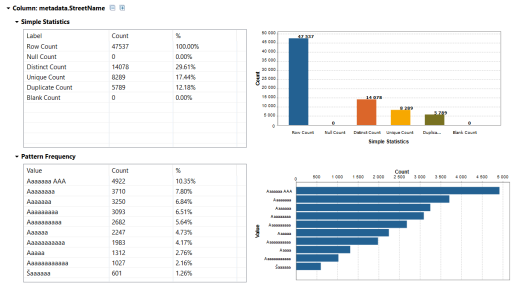


4. Obchodné meno

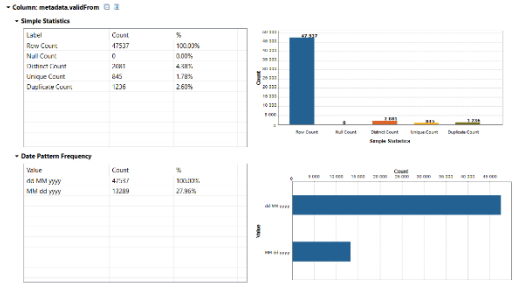


Register Adries - Street Name

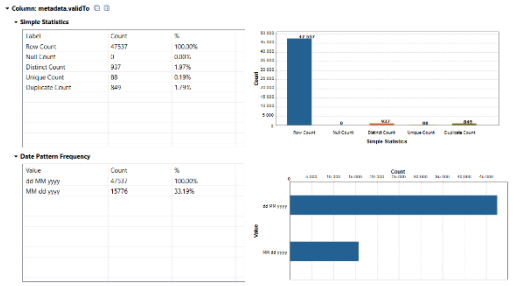
1. Street Name



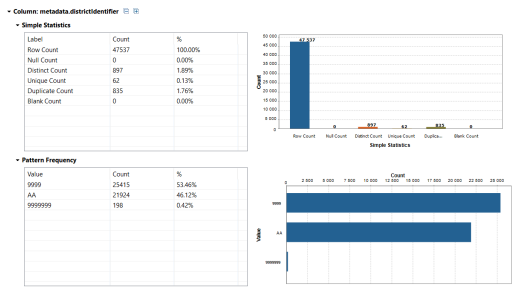
2. Valid From



3. Valid To



4. District Identifier



KPMG

Ďakujeme za pozornosť

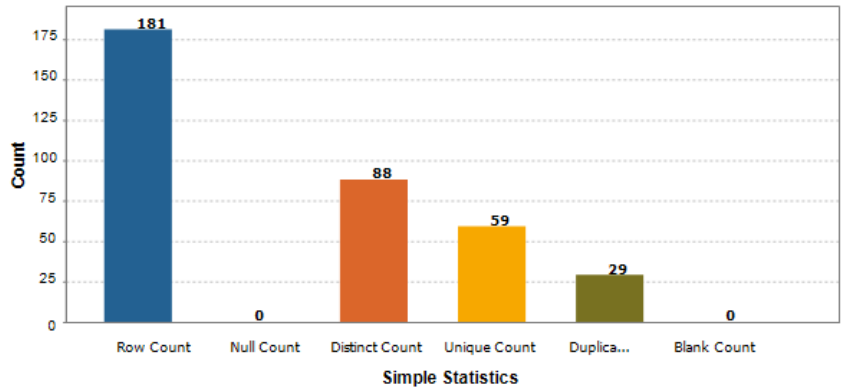
Klient - Meno

Späť

Column: klient.meno

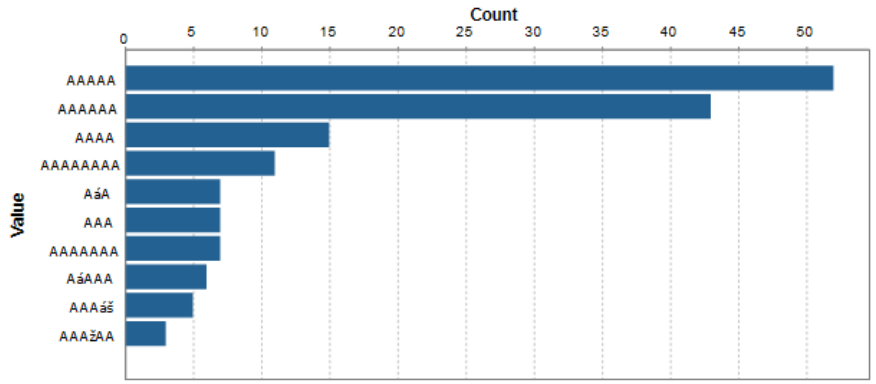
Simple Statistics

Label	Count	%
Row Count	181	100.00%
Null Count	0	0.00%
Distinct Count	88	48.62%
Unique Count	59	32.60%
Duplicate Count	29	16.02%
Blank Count	0	0.00%



Pattern Frequency

Value	Count	%
AAAAA	52	28.73%
AAAAAA	43	23.76%
AAAA	15	8.29%
AAAAAAAA	11	6.08%
AáA	7	3.87%
AAA	7	3.87%
AAAAAAA	7	3.87%
AáAAA	6	3.31%
AAAáš	5	2.76%
AAAŽAA	3	1.66%



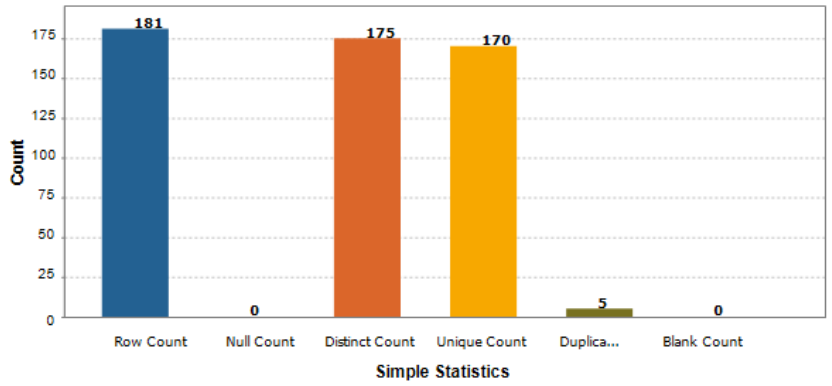
Klient - Rodné číslo

Späť

Column: klient.rodne_cislo

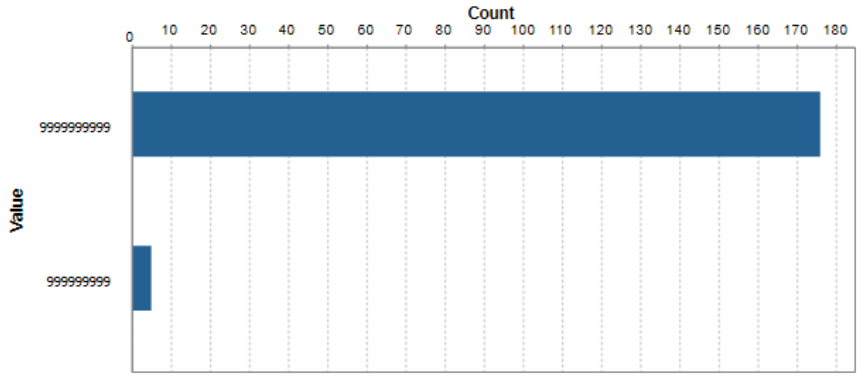
Simple Statistics

Label	Count	%
Row Count	181	100.00%
Null Count	0	0.00%
Distinct Count	175	96.69%
Unique Count	170	93.92%
Duplicate Count	5	2.76%
Blank Count	0	0.00%



Pattern Frequency

Value	Count	%
9999999999	176	97.24%
999999999	5	2.76%



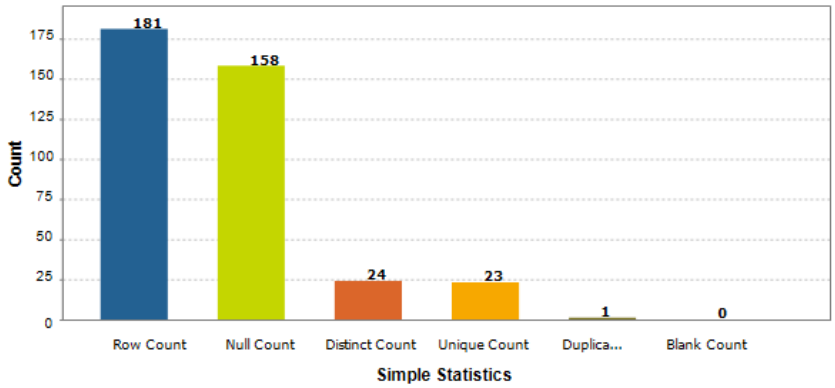
Klient - IČO

Späť

Column: klient.ico

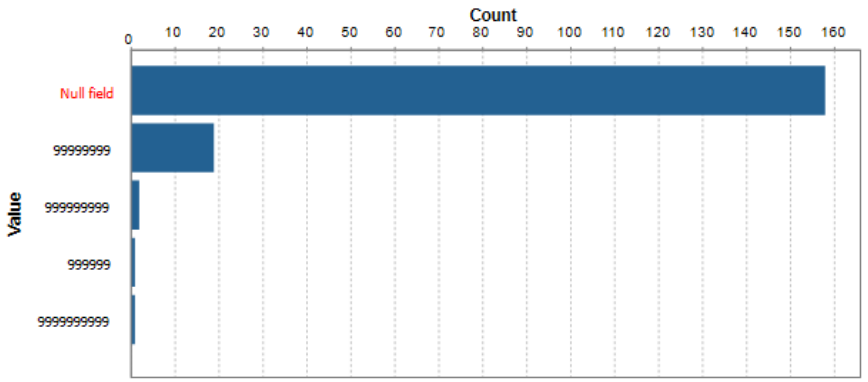
Simple Statistics

Label	Count	%
Row Count	181	100.00%
Null Count	158	87.29%
Distinct Count	24	13.26%
Unique Count	23	12.71%
Duplicate Count	1	0.55%
Blank Count	0	0.00%



Pattern Frequency

Value	Count	%
Null field	158	87.29%
99999999	19	10.50%
999999999	2	1.10%
999999	1	0.55%
9999999999	1	0.55%



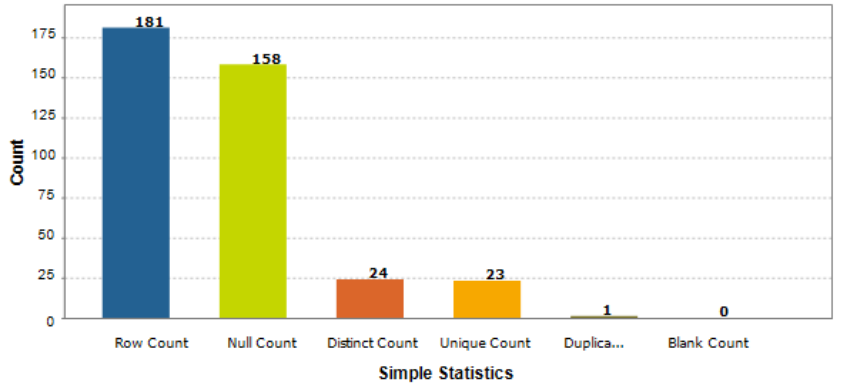
Klient - Obchodné meno

Späť

Column: klient.obchodne_meno

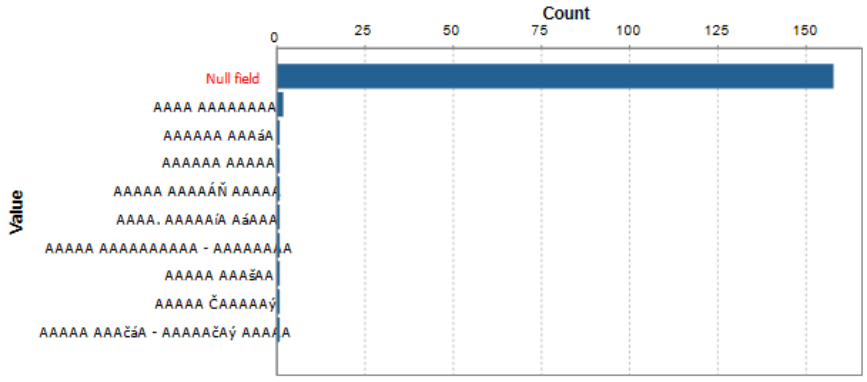
Simple Statistics

Label	Count	%
Row Count	181	100.00%
Null Count	158	87.29%
Distinct Count	24	13.26%
Unique Count	23	12.71%
Duplicate Count	1	0.55%
Blank Count	0	0.00%



Pattern Frequency

Value	Count	%
Null field	158	87.29%
AAAA AAAAAAAA	2	1.10%
AAAAAA AAAáA	1	0.55%
AAAAAA AAAAA	1	0.55%
AAAAA AAAAÁŇ AAAAA	1	0.55%
AAAA. AAAAAíA AáAAA	1	0.55%
AAAAA AAAAAAAAAA - A...	1	0.55%
AAAAA AAAšAA	1	0.55%
AAAAA ČAAAAAý	1	0.55%
AAAAA AAAčáA - AAAAAč...	1	0.55%



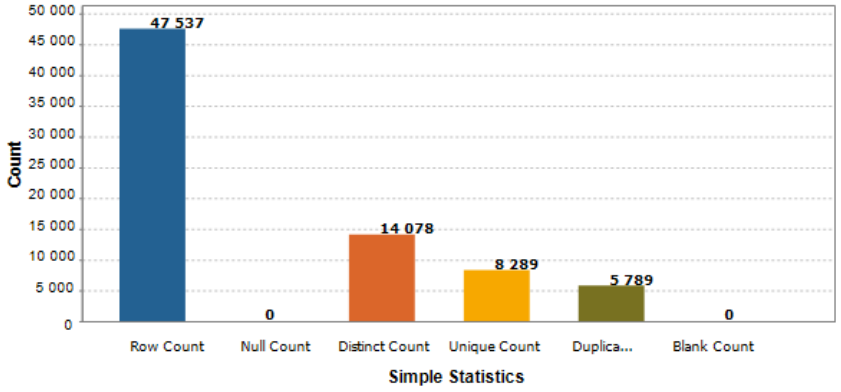
Street Name

Späť

Column: metadata.StreetName

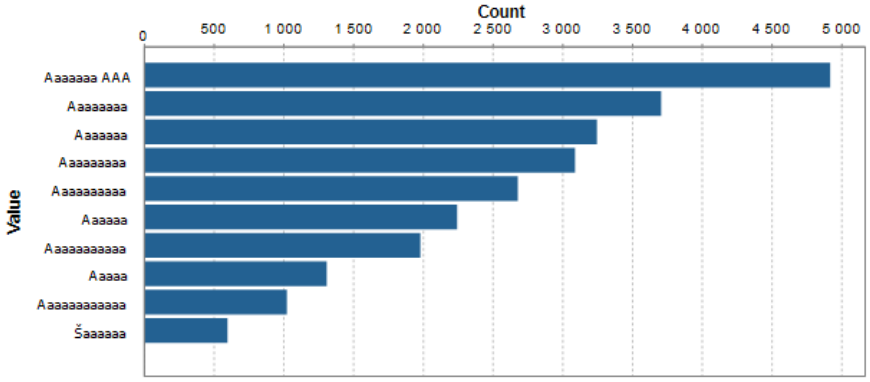
Simple Statistics

Label	Count	%
Row Count	47537	100.00%
Null Count	0	0.00%
Distinct Count	14078	29.61%
Unique Count	8289	17.44%
Duplicate Count	5789	12.18%
Blank Count	0	0.00%



Pattern Frequency

Value	Count	%
Aaaaaaa AAA	4922	10.35%
Aaaaaaaa	3710	7.80%
Aaaaaaa	3250	6.84%
Aaaaaaaaa	3093	6.51%
Aaaaaaaaaa	2682	5.64%
Aaaaaa	2247	4.73%
Aaaaaaaaaaa	1983	4.17%
Aaaaa	1312	2.76%
Aaaaaaaaaaaa	1027	2.16%
Šaaaaaa	601	1.26%



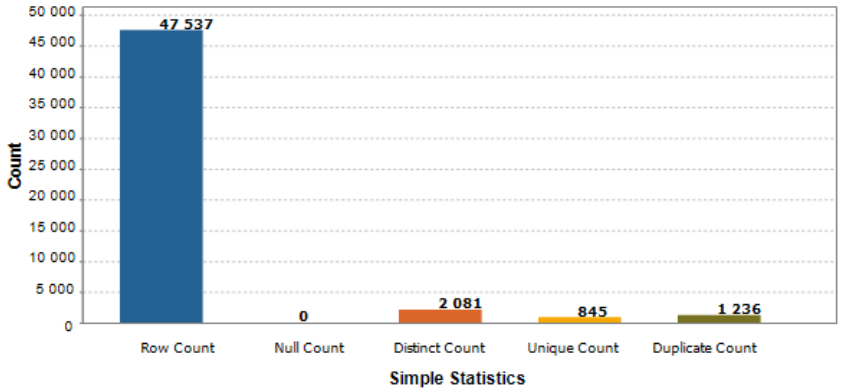
Valid From

Späť

Column: metadata.validFrom

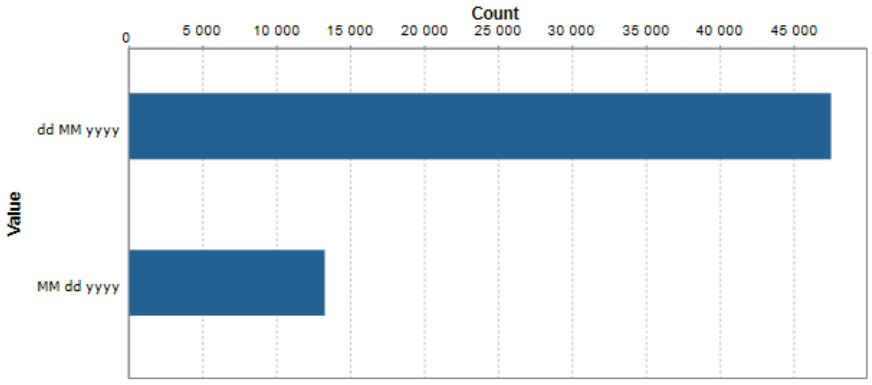
Simple Statistics

Label	Count	%
Row Count	47537	100.00%
Null Count	0	0.00%
Distinct Count	2081	4.38%
Unique Count	845	1.78%
Duplicate Count	1236	2.60%



Date Pattern Frequency

Value	Count	%
dd MM yyyy	47537	100.00%
MM dd yyyy	13289	27.96%



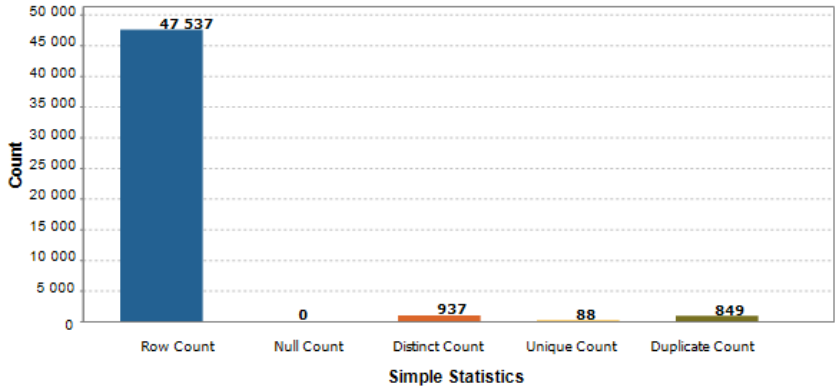
Valid To

Späť

Column: metadata.validTo

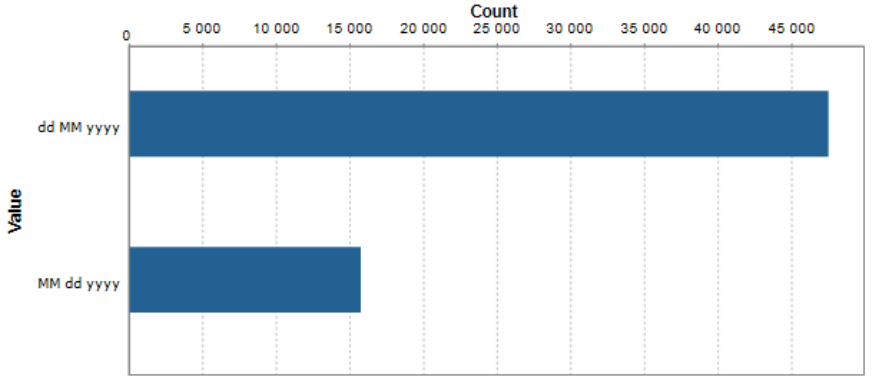
Simple Statistics

Label	Count	%
Row Count	47537	100.00%
Null Count	0	0.00%
Distinct Count	937	1.97%
Unique Count	88	0.19%
Duplicate Count	849	1.79%



Date Pattern Frequency

Value	Count	%
dd MM yyyy	47537	100.00%
MM dd yyyy	15776	33.19%



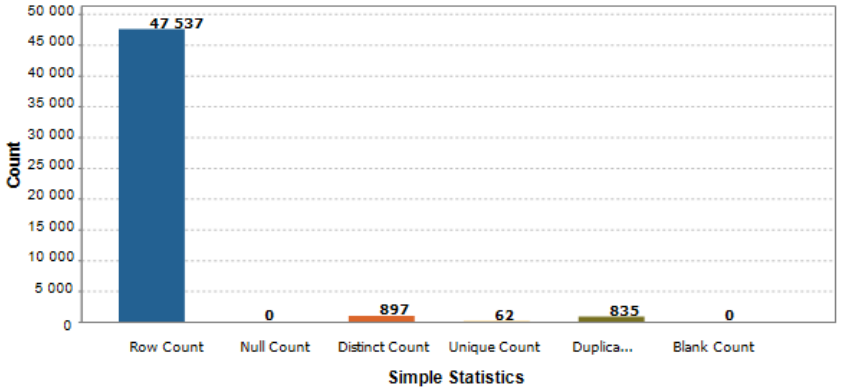
District Identifier

Späť

Column: metadata.districtIdentifier

Simple Statistics

Label	Count	%
Row Count	47537	100.00%
Null Count	0	0.00%
Distinct Count	897	1.89%
Unique Count	62	0.13%
Duplicate Count	835	1.76%
Blank Count	0	0.00%



Pattern Frequency

Value	Count	%
9999	25415	53.46%
AA	21924	46.12%
9999999	198	0.42%