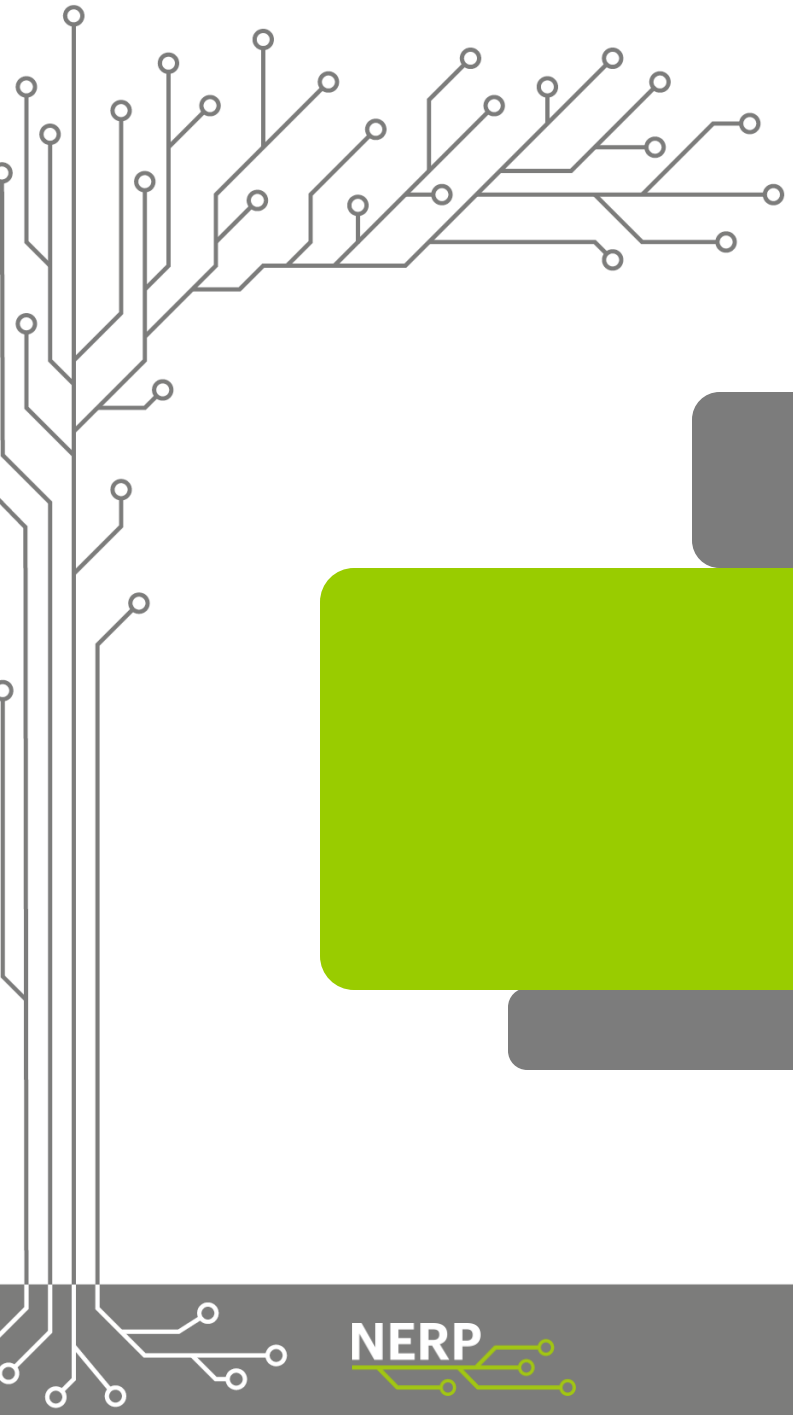


Evropská unie
Evropský sociální fond
Operační program Zaměstnanost

Miroslav Kukuc

Seminář NERP I.

20.10.2021, CENIA

Evropská unie
Evropský sociální fond
Operační program Zaměstnanost

NERP

NERP

Národní Enviromentální Reportingová Platforma

Optimalizace systému řízení příjmu, validace, zpracování a reportingu datových sad v rezortu životního prostředí

CZ.03.4.74/0.0/0.0/15_025/0016059

Evropská unie
Evropský sociální fond
Operační program Zaměstnanost

NERP

cenia

Program

Registrace	8:45	0:15
Začátek, úvod, info o konferenci	9:00	0:10
Výsledky ankety - Pavel Liška, Miro Kukuc	9:10	0:20
Formát a struktura DS pro strojové zpracování - Miro Kukuc	9:30	0:30
Otevřená data - Martin Černý	10:00	0:30
WS1 otevírame data, ale která?	10:30	0:20
Vyhodnocení WS1	10:50	0:10
Vizualizace - Kristýna Marvalová	11:00	0:30
Memorandum - Pavel Liška, Pavel Koukal	11:30	0:30
Zapojení poskytovatelů do envireportingu Jiří Přech	12:00	0:30
WS2 Co máme, co nevyužíváme, co nám chybí	12:30	0:10
Vyhodnocení WS2	12:40	0:10
Diskuze	12:50	0:10
Závěr	13:00	

Explainer

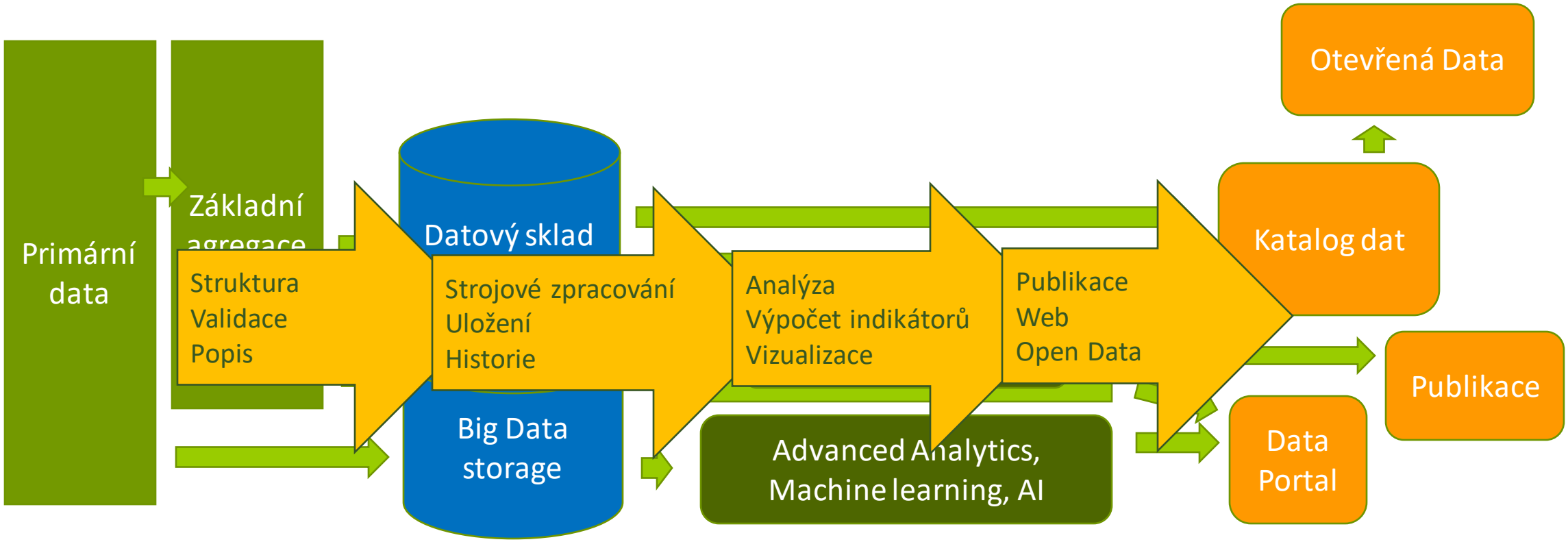
Účel a obsah vzdělávaní

Účelem školení je vzdělávání organizací státní správy k vyšší digitalizaci, zpracování, využívání a zpřístupnění dat o životním prostředí

- Data
- Legislativa
- Otevřená data
- Procesy
- Životní prostředí

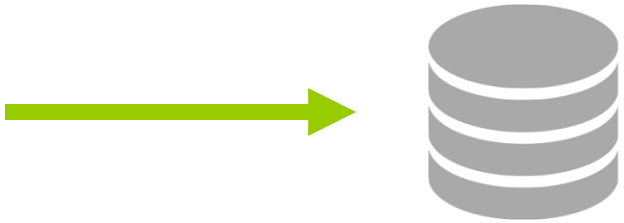
Procesy zpracování dat

Data workflow

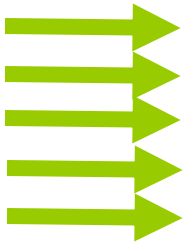


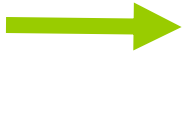
Náročnost zpracování

1.

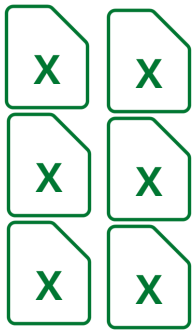


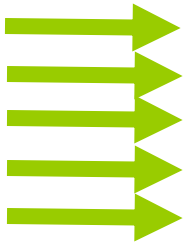
2.





3.





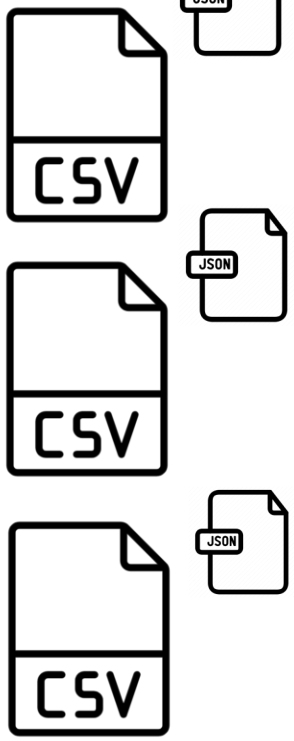
cenia

METADATA

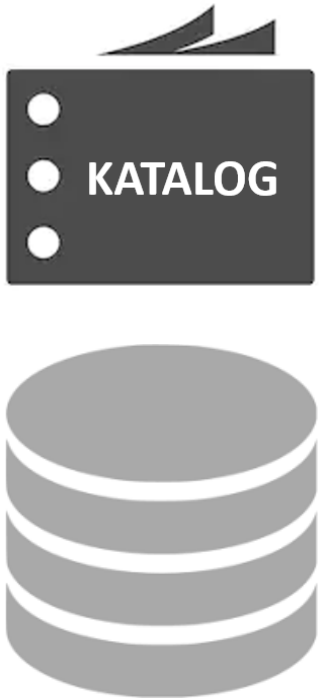
DATA

Vstupní soubor
(Zdrojový excel od poskytovatele)

Listy=
Atomizované (homogenní)
datové sady



Strojově zpracovatelné
datové sady



Důležité aspekty zpracování dat

- Otevřená rozhraní
- Číselníky a jednoznačnost dimenzí
  - (datum – vzniku?, aktualizace?, platnosti?, loadu?)
  - Referenční číselníky
  - Vazby
- Standardizace formátu na strojově zpracovatelné
- Čistota dat – standardizace, čištění, validace
- Zdroj pravdy

Metadata

Metadata – data o datech

- **Popisují**, vysvětlují a lokalizují **data** a jejich zdroj
- **Zjednodušují** získávání, používání a **správu dat**

❑**Descriptive metadata**

Identifikace dat a důvod zpracování (**časové a geografické pokrytí**)

❑**Structural metadata**

Popis struktur a datových modelů (**schéma datové sady**)

❑**Administrative metadata**

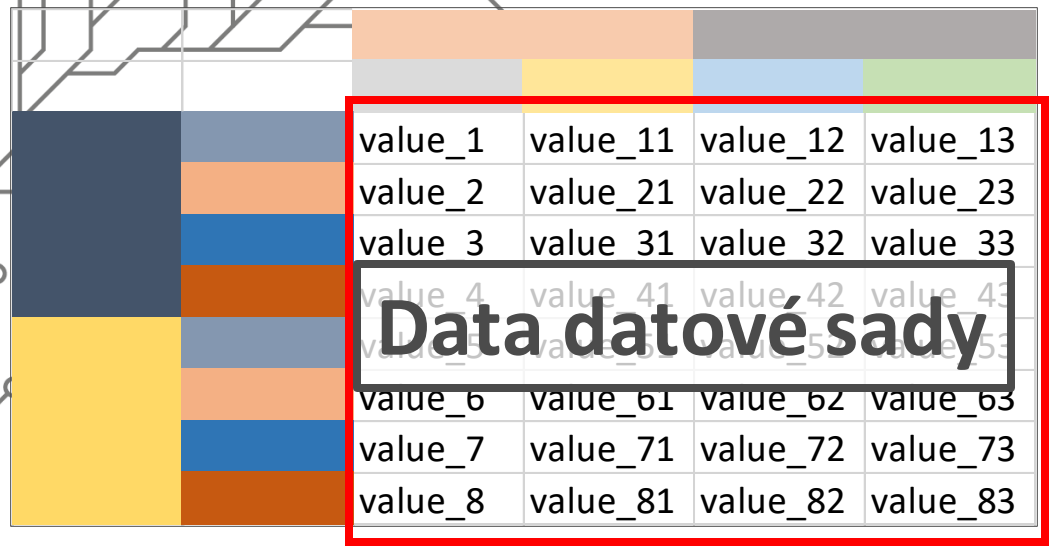
Umožňují spravovat data (**podmínky užití dat,** …)

Klíčový význam u dat z nesourodých zdrojů

*Pro zabezpečení **interoperability mezi systémy** a propojování dat, by metadata měla být mapována na **standardní model** (např. DCAT, Dublin Core).*

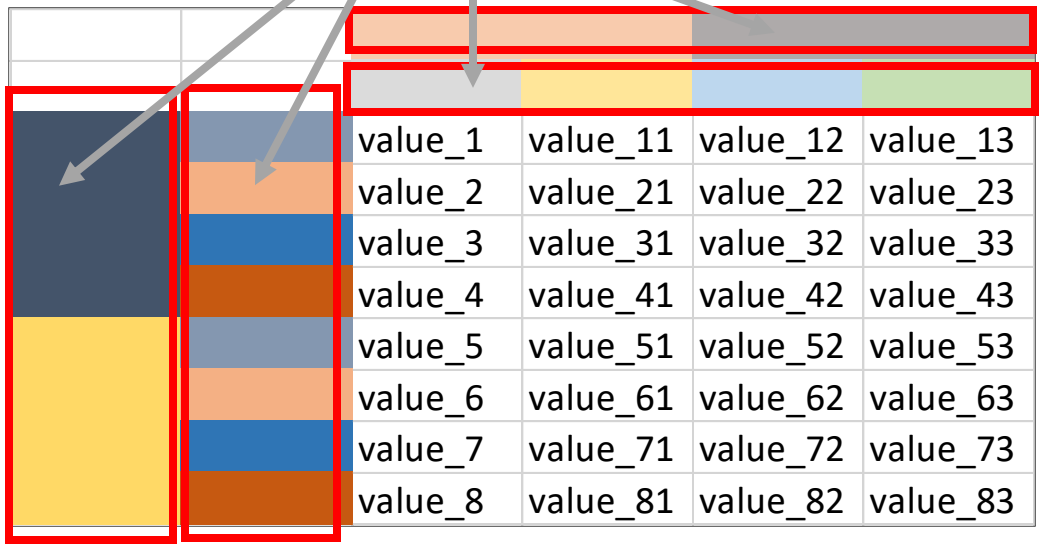
Metodika, formáty dat

Dimenze datové sady



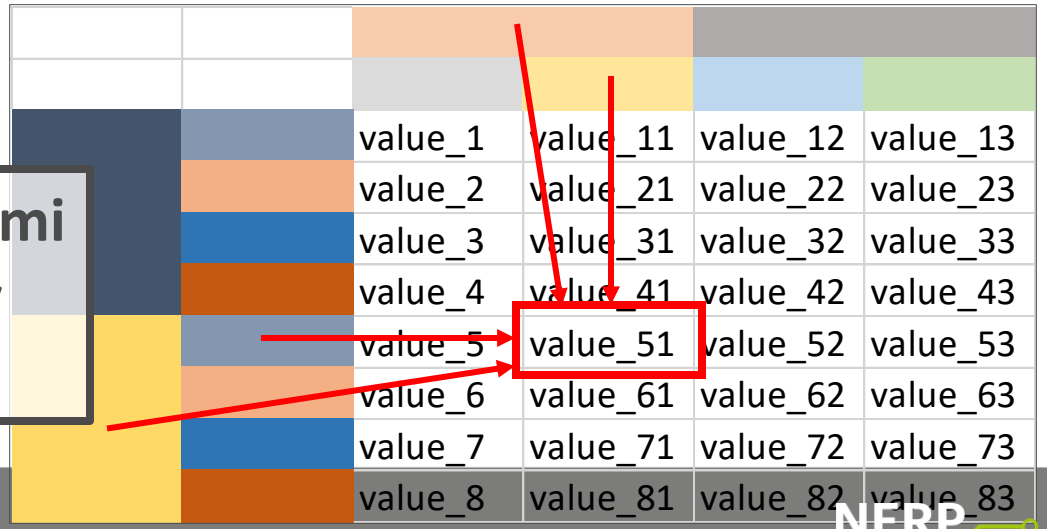
value_1	value_11	value_12	value_13
value_2	value_21	value_22	value_23
value_3	value_31	value_32	value_33
value_4	value_41	value_42	value_43
value_5	value_51	value_52	value_53
value_6	value_61	value_62	value_63
value_7	value_71	value_72	value_73
value_8	value_81	value_82	value_83

Data datové sady



value_1	value_11	value_12	value_13
value_2	value_21	value_22	value_23
value_3	value_31	value_32	value_33
value_4	value_41	value_42	value_43
value_5	value_51	value_52	value_53
value_6	value_61	value_62	value_63
value_7	value_71	value_72	value_73
value_8	value_81	value_82	value_83

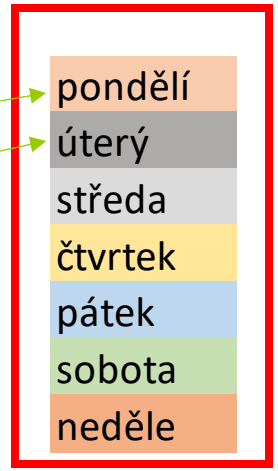
Číslo je identifikováno hodnotami na dimenzích a dalšími atributy (např. jednotka)



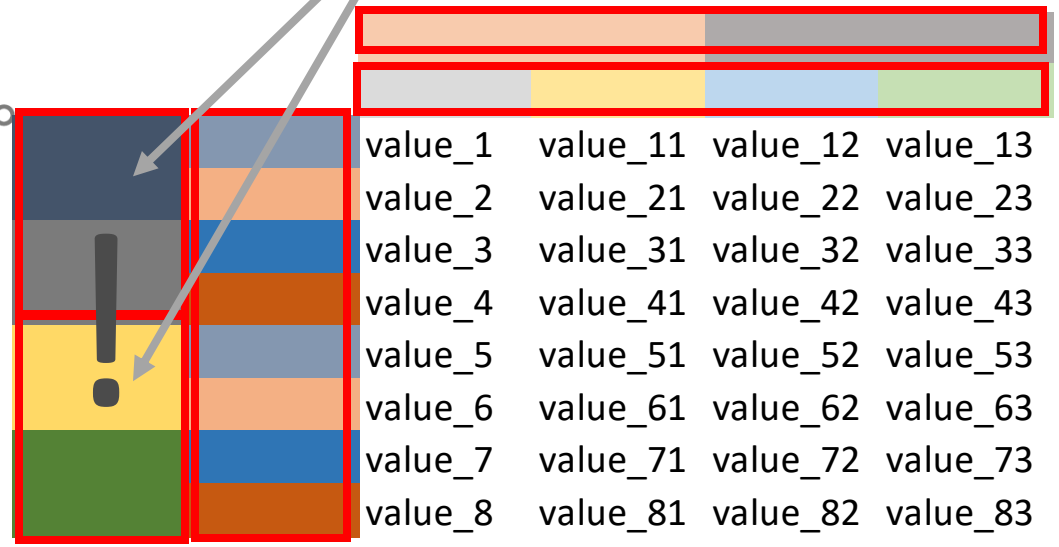
value_1	value_11	value_12	value_13
value_2	value_21	value_22	value_23
value_3	value_31	value_32	value_33
value_4	value_41	value_42	value_43
value_5	value_51	value_52	value_53
value_6	value_61	value_62	value_63
value_7	value_71	value_72	value_73
value_8	value_81	value_82	value_83

Číselník = výčet všech hodnot dimenze

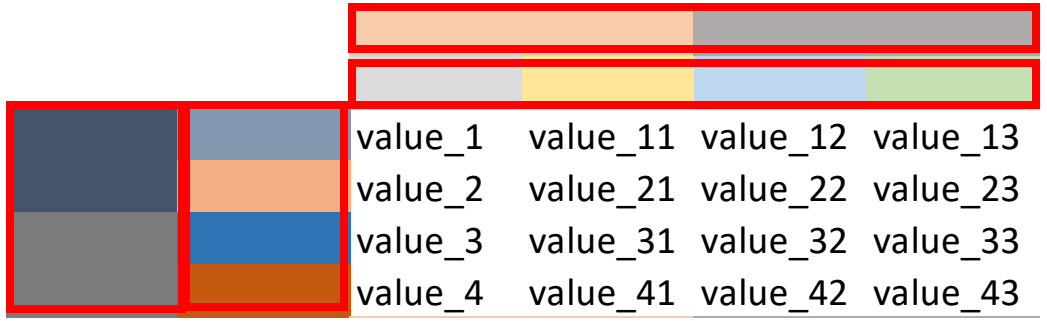
		value_1	value_11	value_12	value_13
		value_2	value_21	value_22	value_23
		value_3	value_31	value_32	value_33
		value_4	value_41	value_42	value_43
		value_5	value_51	value_52	value_53
		value_6	value_61	value_62	value_63
		value_7	value_71	value_72	value_73
		value_8	value_81	value_82	value_83



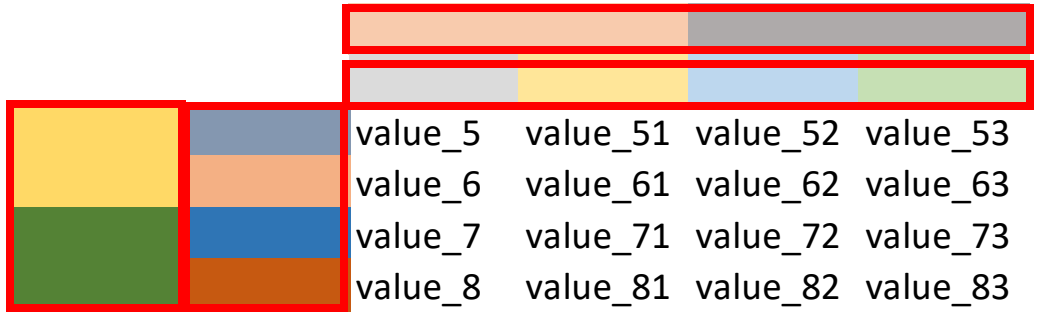
2 dimenze datové sady



value_1	value_11	value_12	value_13
value_2	value_21	value_22	value_23
value_3	value_31	value_32	value_33
value_4	value_41	value_42	value_43
value_5	value_51	value_52	value_53
value_6	value_61	value_62	value_63
value_7	value_71	value_72	value_73
value_8	value_81	value_82	value_83



value_1	value_11	value_12	value_13
value_2	value_21	value_22	value_23
value_3	value_31	value_32	value_33
value_4	value_41	value_42	value_43



value_5	value_51	value_52	value_53
value_6	value_61	value_62	value_63
value_7	value_71	value_72	value_73
value_8	value_81	value_82	value_83

Nehomogenní tabulka

Homogenní = atomické tabulky

Těžba dřeva, 1970–2019

Dřeviny	2013	2014	2015	2016	2017	2018	2019
	tis. m^3 b. k.						
Těžba dřeva celkem	15,331	15,476	16,163	17,617	19,387	25,689	32,586
v tom:							
jehličnaté	13,229	13,472	14,385	15,924	17,735	24,213	31,313
z toho:							
smrk	10,667	10,984	12,230	13,986	15,775	22,412	29,350
jedle	119	117	107	115	116	136	157
borovice	1,879	1,805	1,558	1,368	1,363	1,127	1,288
modřín	532	523	462	424	457	522	505
listnaté	2,102	2,004	1,778	1,693	1,652	1,476	1,273
z toho:							
topol osika	485	448	410	391	353	305	264
buk	949	897	763	747	721	654	567
lípa	74	73	66	54	52	42	37
topolosika	88	91	76	60	64	48	42
Nahodilá těžba	4,248	4,527	8,153	9,399	11,743	23,013	30,945
v tom:							
živelní	2,277	2,455	4,388	2,636	4,345	8,378	5,879
exhalační	22	19	28	29	20	19	20
hmyzová	1,052	1,133	2,309	4,420	5,853	13,059	22,780
ostatní	897	920	1,428	2,314	1,525	1,557	2,266

Pozn.: Těžba dřeva zahrnuje hmotu hroubí i část nehroubí (většina nehroubí – těžební zbytky – nezapočteno), která byla přijata jako hotový sortiment
Zdroj: ČSÚ

Těžba dřeva - dřeviny

Dřeviny	2013	2014	2015	2016	2017	2018	2019
	tis. m^3 b. k.						
smrk	10,667	10,984	12,230	13,986	15,775	22,412	29,350
jedle	119	117	107	115	116	136	157
borovice	1,879	1,805	1,558	1,368	1,363	1,127	1,288
modřín	532	523	462	424	457	522	505
ostatní jehličnaté	32	43	29	21	24	16	13
dub	485	448	410	391	353	305	264
buk	949	897	763	747	721	654	567
lípa	74	73	66	54	52	42	37
topol osika	88	91	76	60	64	48	42
ostatní listnaté	506	495	463	441	462	427	363

Nahodilá těžba

Nahodilá těžba	2013	2014	2015	2016	2017	2018	2019
	tis. m^3 b. k.						
živelní	2,277	2,455	4,388	2,636	4,345	8,378	5,879
exhalační	22	19	28	29	20	19	20
hmyzová	1,052	1,133	2,309	4,420	5,853	13,059	22,780
ostatní	897	920	1,428	2,314	1,525	1,557	2,266

DS1

DS2

Metadata

Evropská unie
Evropský sociální fond
Operační program Zaměstnanost

NERP

cenia

Číselník = výčet všech hodnot dimenze

DSC3

Dřeviny

smrk	jehličnaté
jedle	jehličnaté
borovice	jehličnaté
modřín	jehličnaté
dub	listnaté
buk	listnaté
lípa	listnaté
topol osika	listnaté
javor	listnaté
jasan	listnaté

Hierarchický číselník

DSC1

Těžba dřeva - dřeviny

Dřeviny	2013	2014	2015	2016	2017	2018	2019
	tis. m^3 b. k.						
smrk	10,667	10,984	12,230	13,986	15,775	22,412	29,350
jedle	119	117	107	115	116	136	157
borovice	1,879	1,805	1,558	1,368	1,363	1,127	1,288
modřín	532	523	462	424	457	522	505
ostatní jehličnaté	32	43	[illegible]	31	24	16	13
dub	485	448	410	391	353	305	264
buk	949	897	763	747	721	654	567
lípa	74	73	66	54	52	42	37
topol osika	88	91	76	60	64	48	42
ostatní listnaté	506	495	463	441	462	427	363

DS1

Nahodilá těžba

živelní	
exhalační	
hmyzová	
ostatní	

DSC2

Nahodilá těžba

Nahodilá těžba	2013	2014	2015	2016	2017	2018	2019
	tis. m^3 b. k.						
živelní	2,277	2,455	[illegible]	2,636	4,345	8,378	5,879
exhalační	22	19	[illegible]	29	20	19	20
hmyzová	1,052	1,133	2,309	4,420	5,853	13,059	22,780
ostatní	897	920	1,428	2,314	1,525	1,557	2,266

DS2

Číselníky jsou společné pro všechny DS

Číselníky

Celkem 

Nedefinováno 

Ostatní xyz 

Číselníky

Nahodilá těžba	1970	1980	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000
celkem	3,226	7,060	9,822	6,939	5,400	8,100	9,282	7,855	6,855	5,527	3,840	3,736	3,288
živelní	-	-	8,701	5,309	2,803	3,935	4,355	2,766	4,058	3,982	2,592	2,743	2,388
exhalační	-	-	289	290	286	323	276	303	192	212	171	126	78
hmyzová	-	-	178	257	837	2,097	2,211	2,376	1,145	546	408	327	320
ostatní	-	-	654	1,083	1,474	1,745	2,440	2,410	1,460	787	669	540	502

Zdroj: ČSÚ

Nahodilá těžba	1970	1980	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000
~~celkem~~	~~3,226~~	~~7,060~~	~~9,822~~	~~6,939~~	~~5,400~~	~~8,100~~	~~9,282~~	~~7,855~~	~~6,855~~	~~5,527~~	~~3,840~~	~~3,736~~	~~3,288~~
živelní			8,701	5,309	2,803	3,935	4,355	2,766	4,058	3,982	2,592	2,743	2,388
exhalační			289	290	286	323	276	303	192	212	171	126	78
hmyzová			178	257	837	2,097	2,211	2,376	1,145	546	408	327	320
ostatní nahodilá těžba			654	1,083	1,474	1,745	2,440	2,410	1,460	787	669	540	502
nedefinováno	3,226	7,060	0	0	0	0	0	0	0	0	0	0	0
Celkem	**3,226**	**7,060**	**9,822**	**6,939**	**5,400**	**8,100**	**9,282**	**7,855**	**6,855**	**5,527**	**3,840**	**3,736**	**3,288**

Výpočet v BI

cenia

Faktové tabulky

- Formát dat (celé číslo, desetinné číslo, text, datum, …)
- Zaokrouhlení
- Vhodná agregace
- Poznámky

Datové formáty

Typické strojově zpracovatelné datové formáty:

- csv + json
- json
- xml

Formát csv

```
1   C-ROKY,C-GEO-KRAJE,HODNOTA,C-JEDNOTKY,C-GEO-TYP-UZEMI,C-AUTOR
2   2018,Hlavní město Praha,0,ha,Národní parky,AOPK
3   2018,Středočeský kraj,0,ha,Národní parky,AOPK
4   2018,Jihočeský kraj,34073,ha,Národní parky,AOPK
5   2018,Plzeňský kraj,34511,ha,Národní parky,AOPK
6   2018,Karlovarský kraj,0,ha,Národní parky,AOPK
7   2018,Ústecký kraj,7926,ha,Národní parky,AOPK
8   2018,Liberecký kraj,11649,ha,Národní parky,AOPK
9   2018,Královéhradecký kraj,24671,ha,Národní parky,AOPK
10  2018,Pardubický kraj,0,ha,Národní parky,AOPK
11  2018,Kraj Vysočina,0,ha,Národní parky,AOPK
12  2018,Jihomoravský kraj,6274,ha,Národní parky,AOPK
13  2018,Olomoucký kraj,0,ha,Národní parky,AOPK
14  2018,Zlínský kraj,0,ha,Národní parky,AOPK
```

Strojově čitelný formát

Rozloha chráněných území dle krajů			
	2018 Národní parky ha	2018 Chráněné krajinné oblasti ha	2018 Národní pří ha
Hlavní město Praha	0	518	145
Středočeský kraj	0	109,785	615
Jihočeský kraj	34,073	164,034	1,335
Plzeňský kraj	34,511	96 876	231
Karlovarský kraj	0	5	
Ústecký kraj	7,926	1	
Liberecký kraj	11,649	9	
Královéhradecký kraj	24,671	7	
Pardubický kraj	0	3	

```
1  C-ROKY,C-GEO-KRAJE,HODNOTA,C-JEDNOTKY,C-GEO-TYP-UZEMI,C-AUTOR
2  2018,Hlavní město Praha,0,ha,Národní parky,AOPK
3  2018,Středočeský kraj,0,ha,Národní parky,AOPK
4  2018,Jihočeský kraj,34073,ha,Národní parky,AOPK
5  2018,Plzeňský kraj,34511,ha,Národní parky,AOPK
6  2018,Karlovarský kraj,0,ha,Národní parky,AOPK
7  2018,Ústecký kraj,7926,ha,Národní parky,AOPK
```

File Origin: 65001: Unicode (UTF-8) | Delimiter: Comma | Data Type Detection: Based on first 200 rows

C-ROKY	C-GEO-KRAJE	HODNOTA	C-JEDNOTKY	C-GEO-TYP-UZEMI	C-AUTOR
2018	Hlavní město Praha	0	ha	Národní parky	AOPK
2018	Středočeský kraj	0	ha	Národní parky	AOPK
2018	Jihočeský kraj	34073	ha	Národní parky	AOPK
2018	Plzeňský kraj	34511	ha	Národní parky	AOPK
2018	Karlovarský kraj	0	ha	Národní parky	AOPK
2018	Ústecký kraj	7926	ha	Národní parky	AOPK
2018	Liberecký kraj	11649	ha	Národní parky	AOPK
2018	Královéhradecký kraj	24671	ha	Národní parky	AOPK

Validace

Csv formát

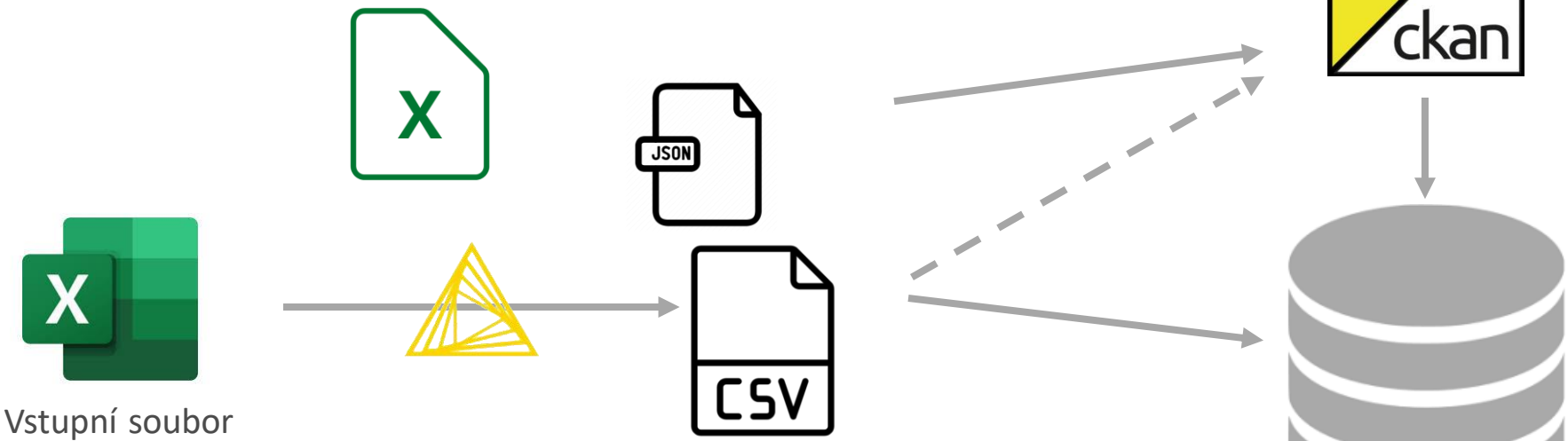
Základní pravidla pro distribuce datových sad ve formátu csv:

- ✓ utf-8 kódování
- ✓ oddělovač hodnot, nebo řetězců = čárka
- ✓ textové řetězce (stringy) v uvozovkách
- ✓ desetinná tečka v číselných hodnotách
- ✓ jednotky zvlášť ve sloupci
- ✓ prázdná hodnota jako prázdná (ne „null“, „n/a“ apod.)
- ✓ pojmenování hlaviček konzistentní napříč soubory
- ☒ prázdné řádky a sloupce

Standardizace

- Metodika
- Národní a mezinárodní standardy
- Otevřená data

Názvosloví



Vstupní soubor (Zdrojový excel od poskytovatele)

KNIME transformace

Atomizované datové sady

Databázové záznamy

poskytovatel_soubor_DS_rok_datum platnosti

csu_lesy_tezba-celkem_2019_20201007

csu_lesy_nahodila-tezba_2019_20201007

Organizace dalších seminářů

- 2 hod/ 4 hod
- Hybridní účast
- Monitorovací listy

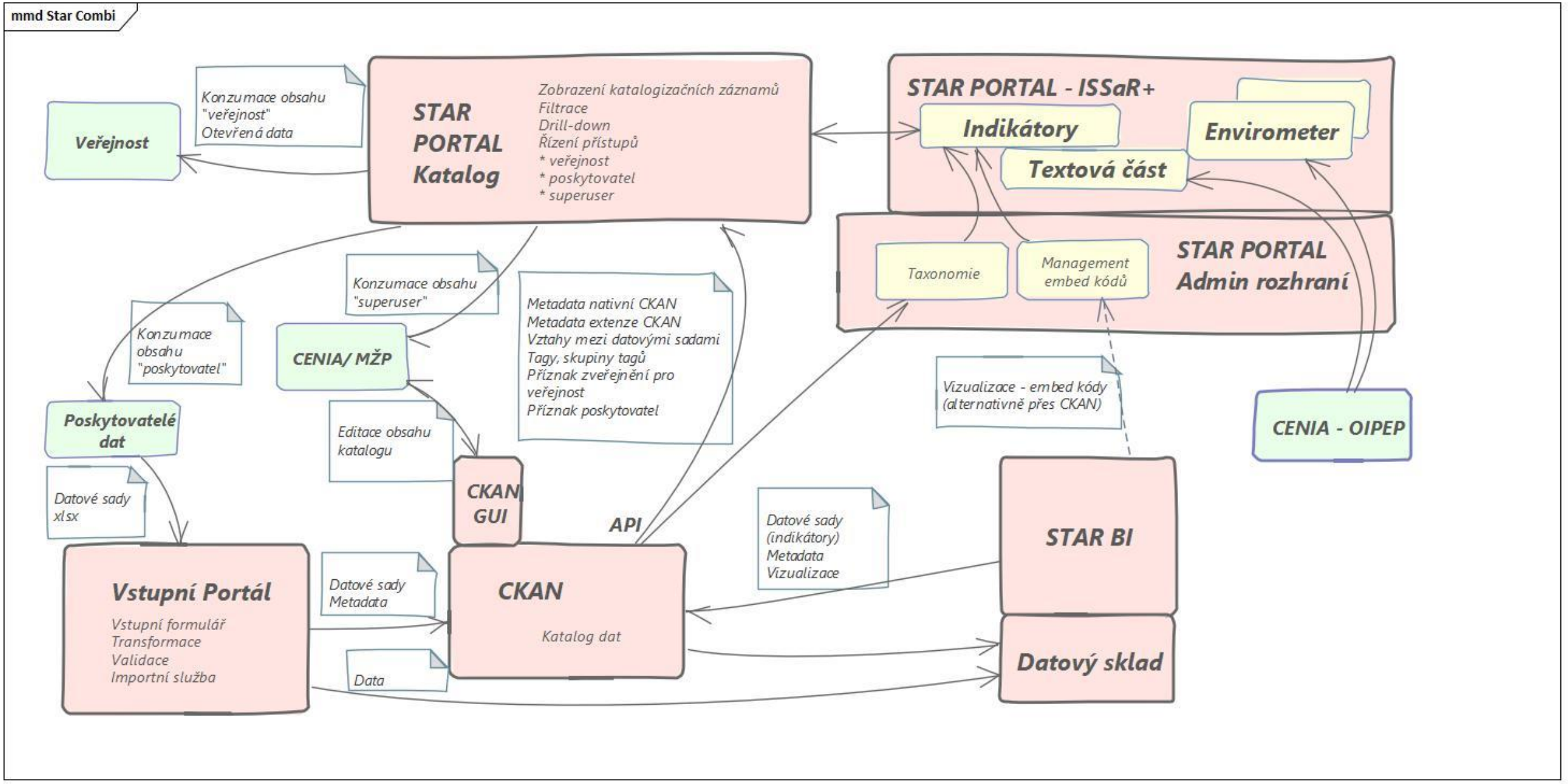
Témata na další semináře:

- Otevřená data
- Datové formáty a šablony
- Datová kvalita
- Proces akvizice dat a jeho optimalizace
- Procesní modelování
- Klasifikace dat a katalogizace
- Datová analytika


GRACIAS
THANK
YOU
ARIGATO
SHUKURIA
BOLZIN
MERCI
BIYAN
TASHAKKUR ATU
GRAZIE
MEHRBANI

Backup

Architektura



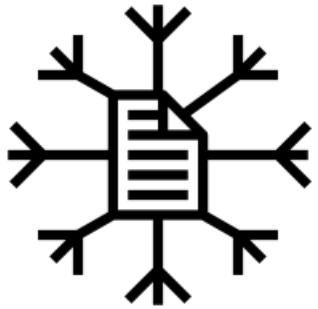
Schema datové sady

Schema musí být definováno pro každou datovou sadu před jejím prvním příjmem

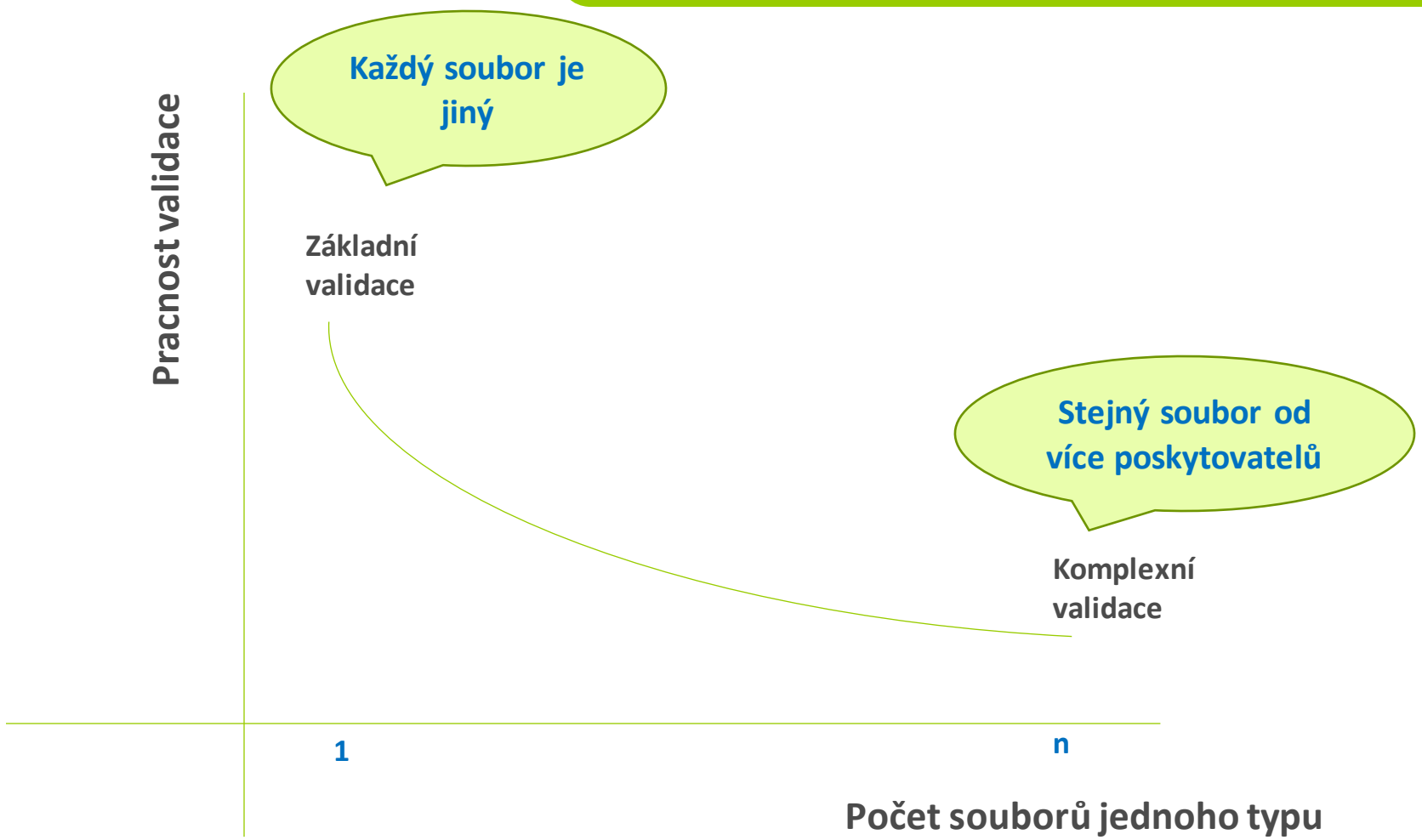
Schema obsahuje všechny parametry datové sady, které jsou potřebné pro její strojové zpracování

Ze **Schema** datové sady se odvozuje např.:

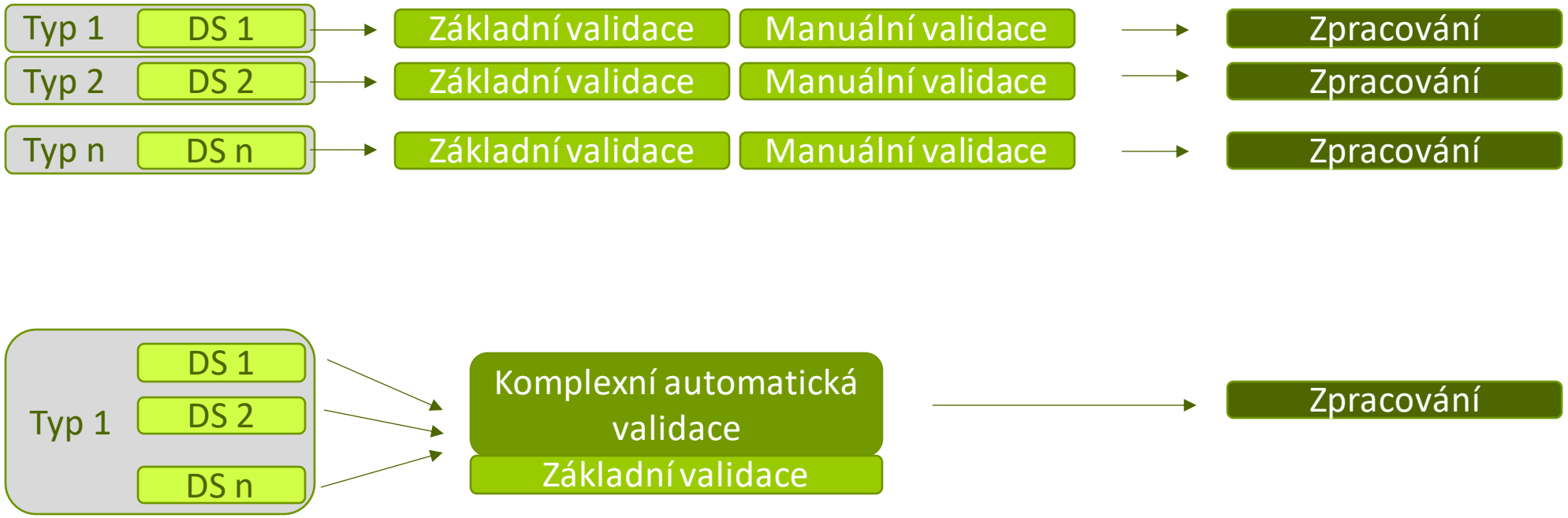
- Transformační skript
- Validační skript
- Importní skript



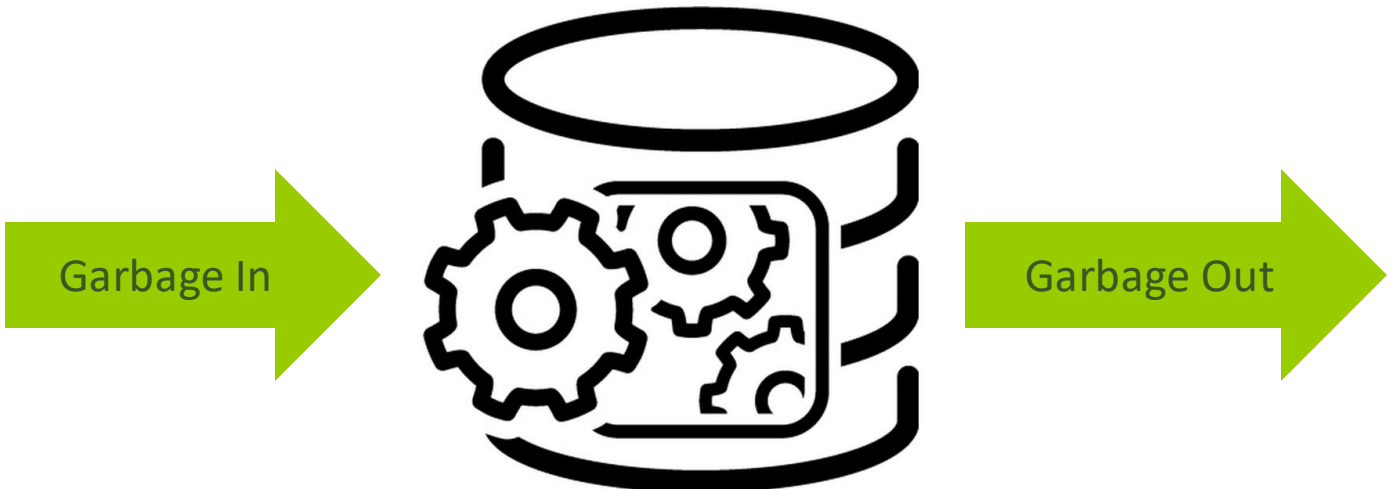
Komplexnost validace



Komplexnost validace



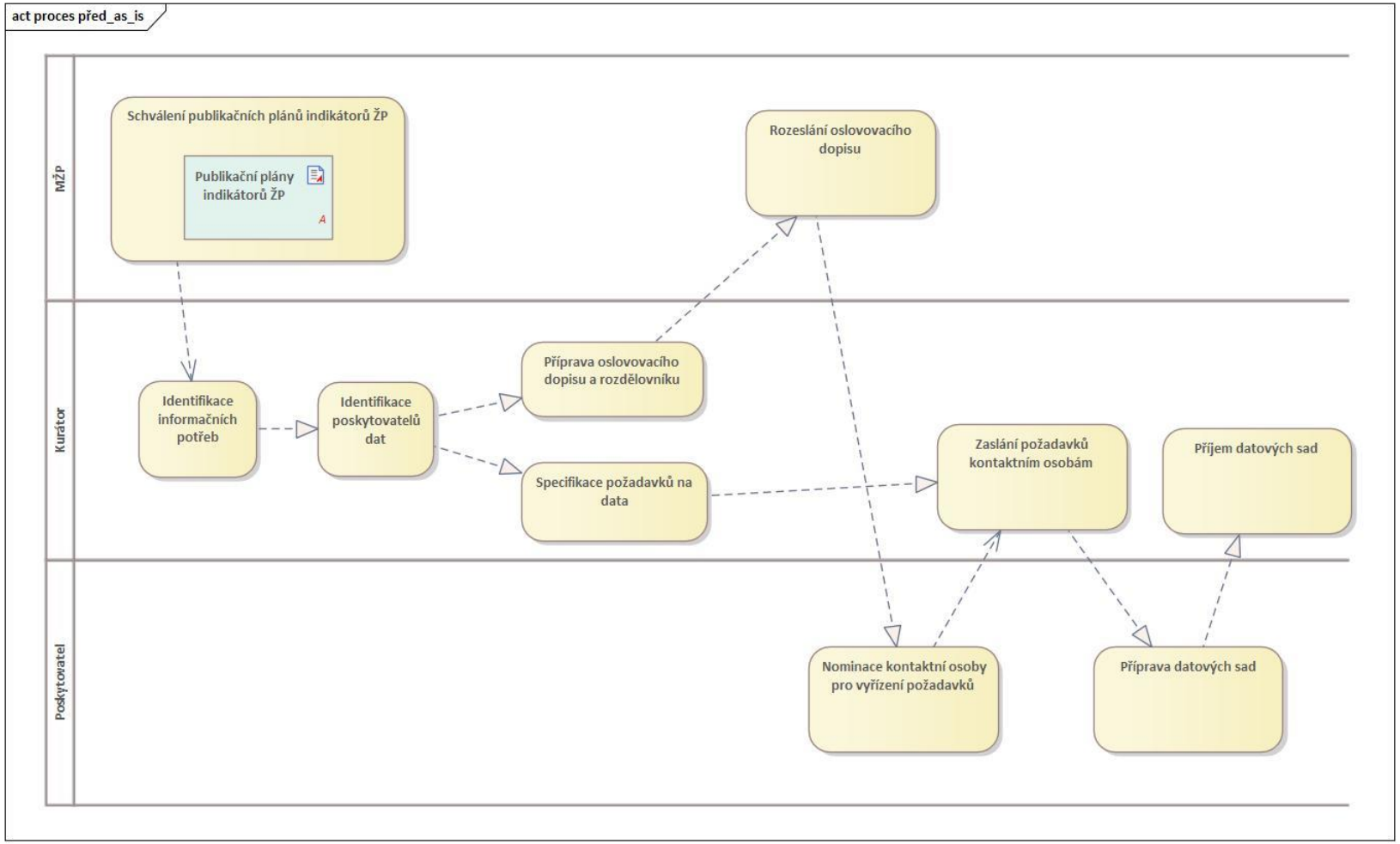
Kvalita



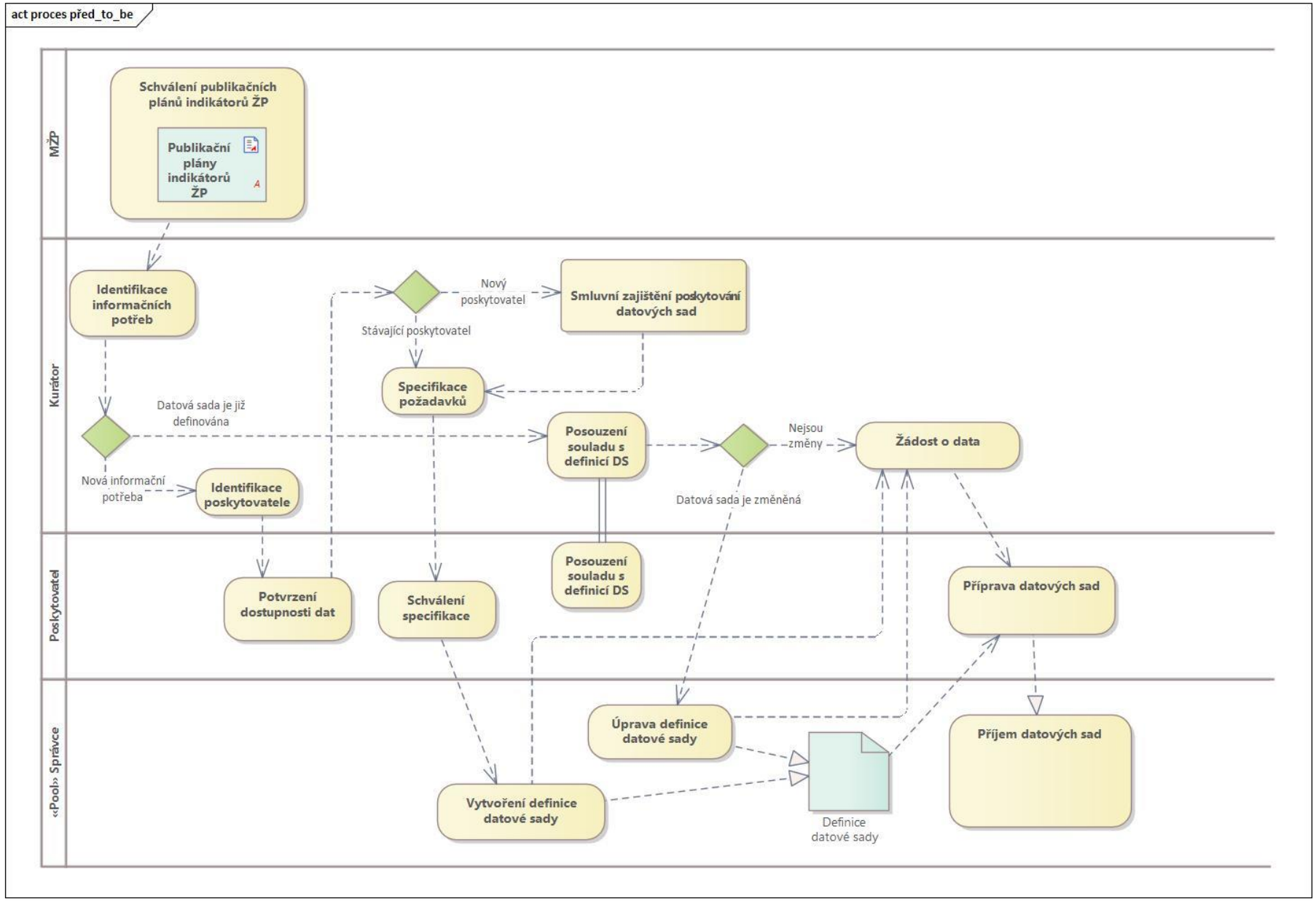
Čistota dat – standardizace, čištění, validace

Proces akvizice dat

Proces akvizice dat – AS IS



Proces akvizice dat – TO BE



cenia