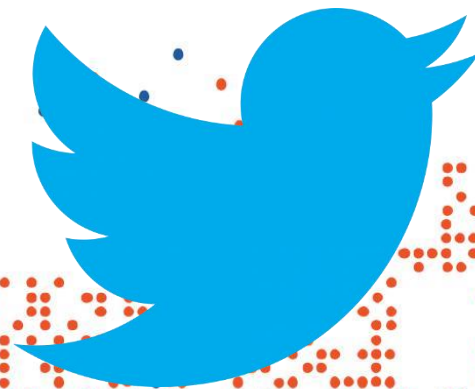




Archivace českých účtů sociální sítě Twitter

Martin Rehtorik

Schéma prezentace

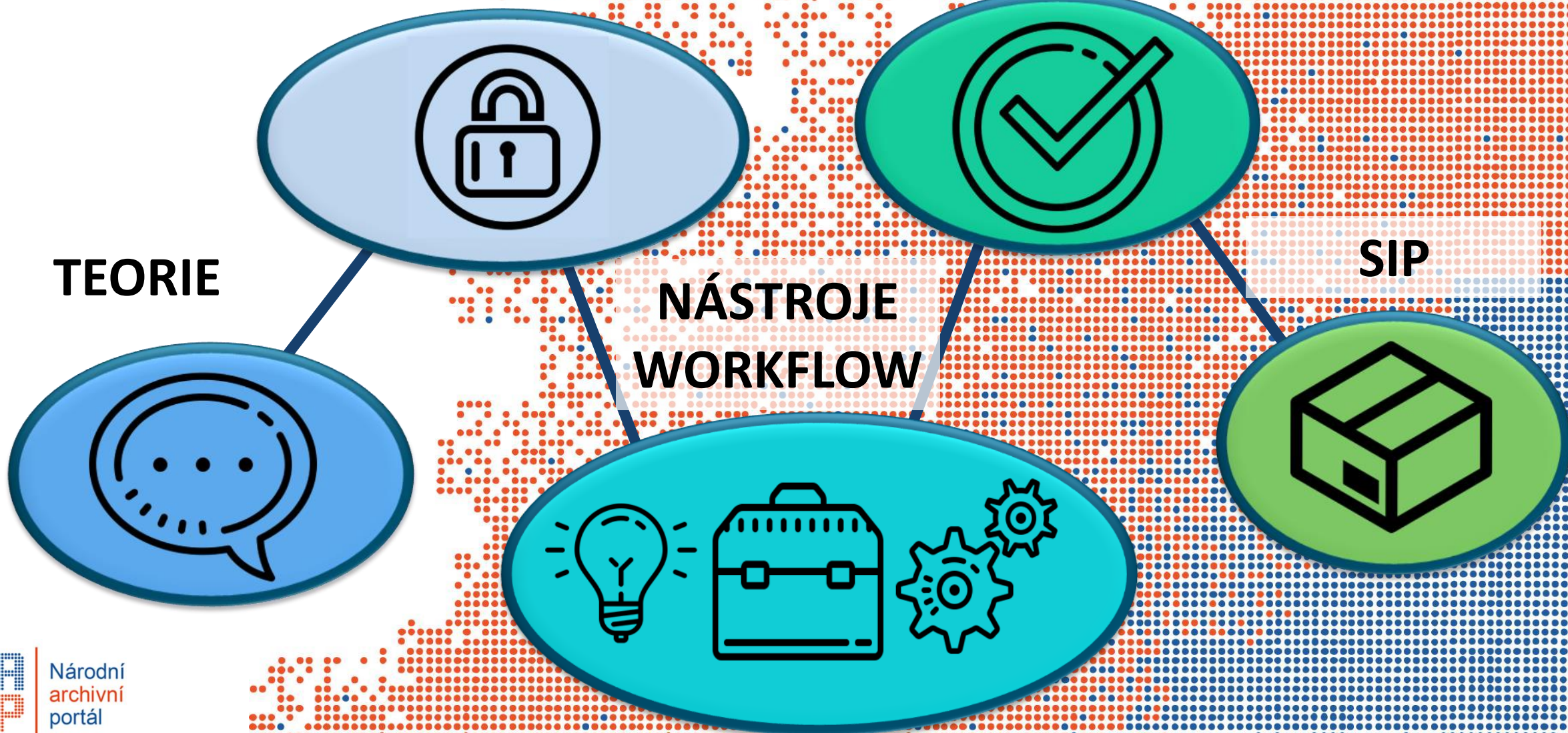
OMEZENÍ A LIMITY

VÝSLEDEK

TEORIE

NÁSTROJE
WORKFLOW

SIP

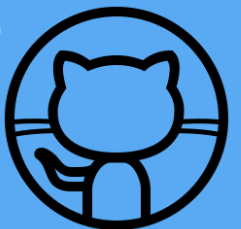




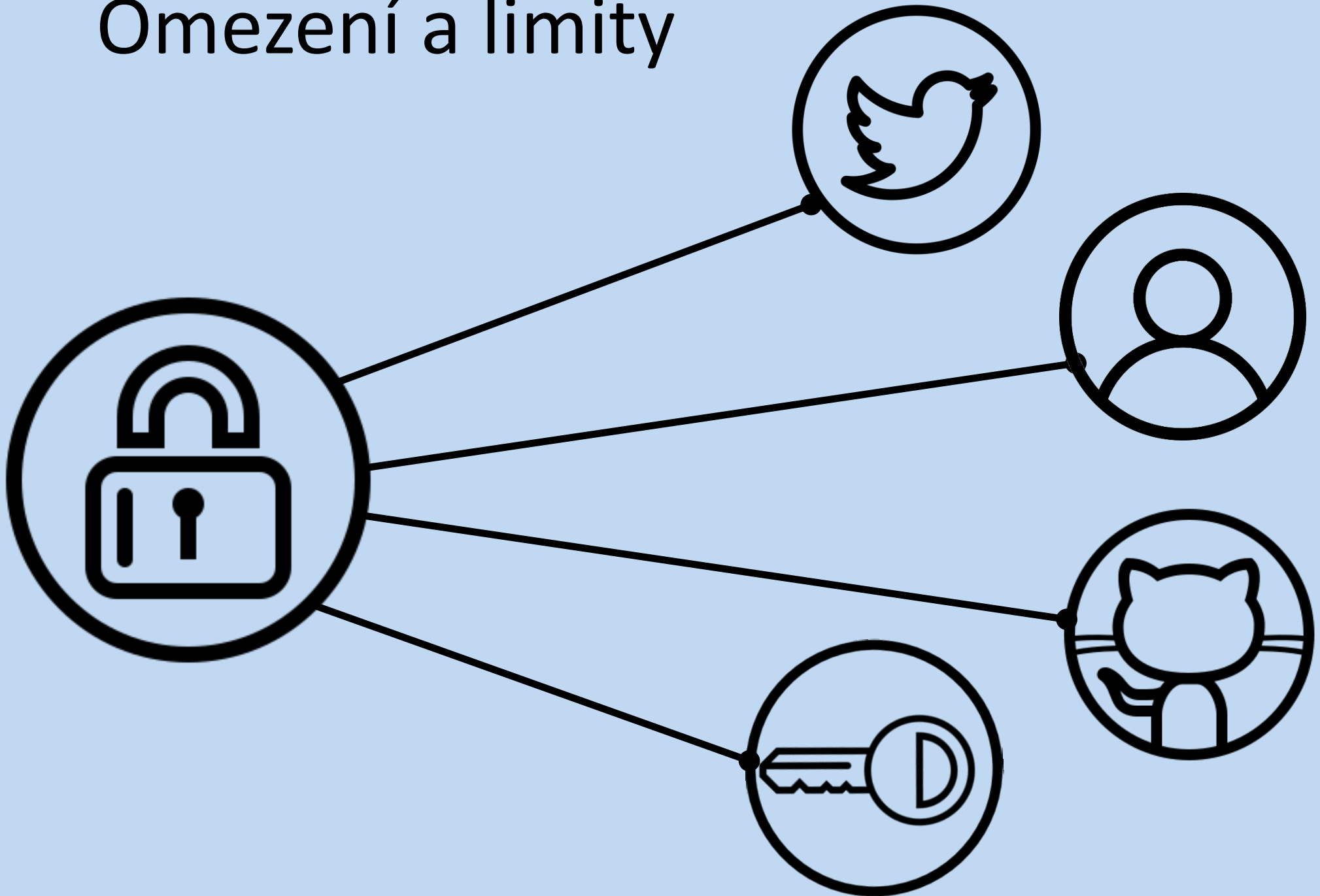
Archivace Twitteru



- Sociální síť Twitter umožňuje stahování nejen tweetů, ale i mediálního obsahu
- Data lze získat nezávisle na zájmu/pomoci původce
- přístup k datům pomocí API v jazyce REST nebo Python
- Téma řešeno na více konferencích
- Prezentováno více nástrojů
- Většina z nich zveřejněna na GitHub

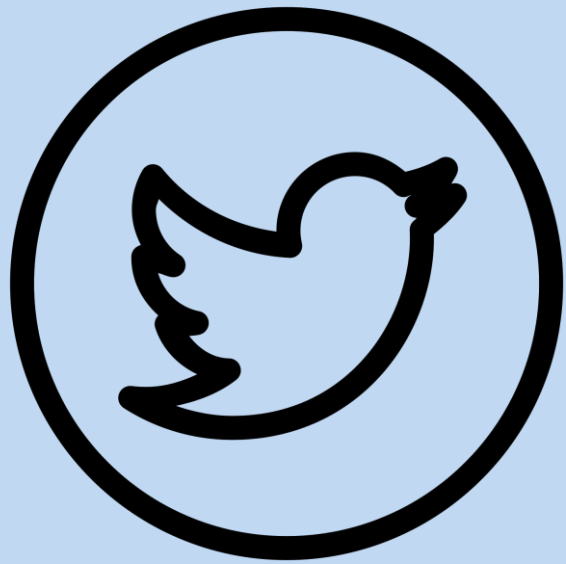


Omezení a limity





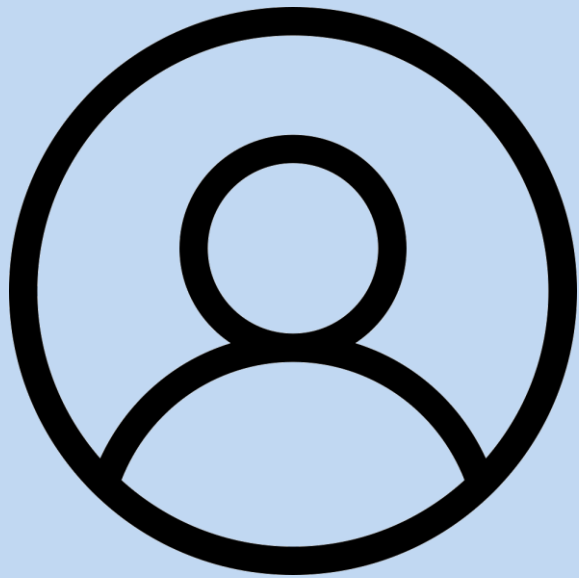
Omezení a limity



- Twitter chrání svá data
- 1 projekt, 3 nástroje
- Zpětně lze stáhnout max 3250 tweetů
- Zpětně lze stáhnout max 2500 mediálních příloh (pouze obrázky)
- Nástroje limitovány počty dotazů pomocí API na 90K na 1 přístup



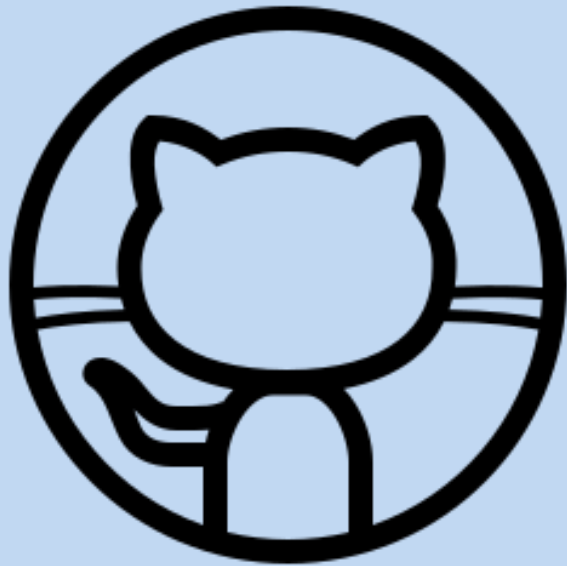
Omezení a limity



- Limity může nastavit i „vlastník“ účtu
- Limitem bude především počet tweetů/retweeetů
- Lze se setkat limitem 1 rok zpětně, např. [PREZIDENTmluvčí](#)
- Stahují se i retweety, u AB jsme se dostali pouze do 4/2021 ačkoliv bylo staženo 3250 příspěvků
- Blokace stažení mediální obsahu
- Např. [hasici_cr](#) retweetují krajské jednotky



Limity nástrojů na GitHub

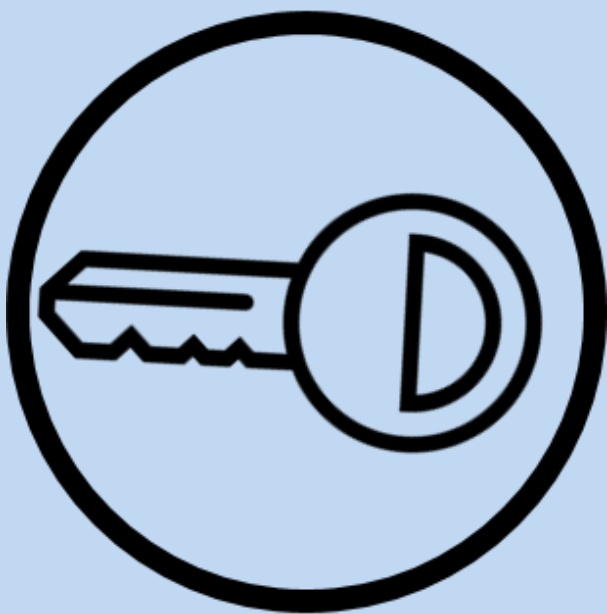


- Ne vše, co je zveřejněno je opravdu použitelné
- Jsou to často prototypy, nebyly ani vyzkoušeny, bývají zastaralé a nefunkční (změny v API atp.)
- chybí dokumentace, manuál, wiki etc.
- Nutné postupovat krok za krokem, dohledávat komponenty atp.
- Riziko při doplňování přímo do kódu, kdy mohou být vyžadovány ‘ ’ nebo “ ” a jejich záměna není možná



Účet - Developer premium

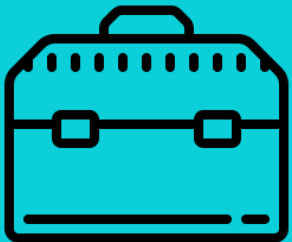
- Přístup k úplnému archivu Twitteru
- Možnost stahovat diskuse pod tweety



Total Requests PER MONTH ⓘ	Month-to-month PRICE PER MONTH ⓘ
Up to 500	\$149.00
Up to 1000	\$289.00
Up to 2,500	\$699.00
Up to 5,000	\$1,299.00
Up to 10,000	\$2,499.00

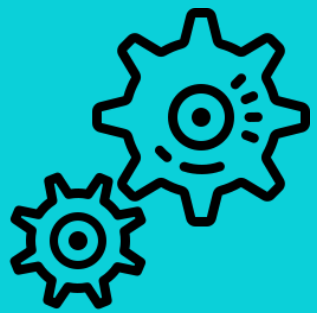
Last 30 days
500
Both
1024 characters
Premium
60 requests/min
URLs, Polls, Profile Geo
2
5M
10 requests/sec

PRAVIDELNĚ PLATIT



Nástroje

- Využíváme finské nástroje z dílny [Digitalia XAMK](#) a [Miguela Martinez-Alvareze](#)
 - všechny jsou v Pythonu 
 - Všechny nástroje jsme si upravili/opravili nebo dodělali
1. Stahuje veškeré tweety zpětně do SQLite DB
 2. Stahuje pouze nové tweety (aktualizuje) do SQLite DB (lhůta je nastavitelná)
 3. Stahuje dostupný mediální všechny dostupné obrázky
 4. Stahuje pouze nové obrázky/fotky



Workflow sklízení



1. Archivář navrhne účet vhodný ke sklizni



2. Stažení tweetů zpětně



3. Stažení mediálního obsah zpětně (obrázky)



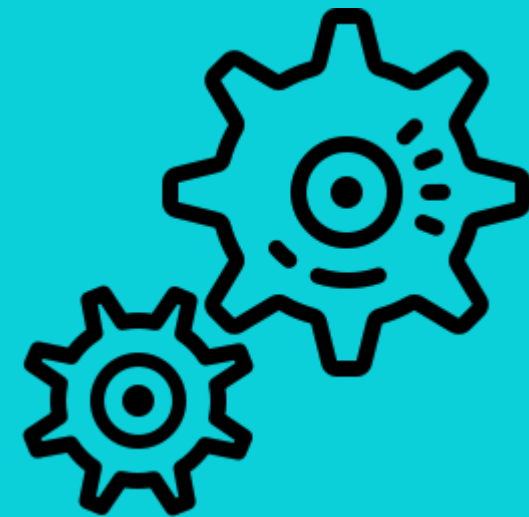
4. Doplnění seznamu pro pravidelnou sklizeň



5. Provádění pravidelné sklizně



Workflow archivace



1. Export tweetů do ? + mediální obsah v .zip



2. Doplnění metadat pro prohlížení



3. Vytvoření balíčku pro prohlížení archivovaného datasetu



4. Uložení do LTP systému digitální archiv



Výsledek

- Sklízíme účty našich původců + něco navíc...
- usoud_official, vnitro, Vit_Rakusan, CelnispravaCR, sfzp_cr, CHMUCHMI, narodniknihovna, MZeCr, mpsvcz, strakovka, P_Fiala, min_dopravy, RSD_oficialni, JanRydl, MatlRadek, DrazniU, CFilharmonie, PamatkovyUstav, Cinohra_ND, Opera_ND, Pamatnik_Lidice, ZdravkoOnline, ChytraKarantena... nyní asi 306 účtů, 660 tisíc tweetů, 195 tisíc mediálních příloh, datový objem cca 50GB



Podoba balíčku SIP pro uložení



- › Vytváření prohlížečího balíčku do DB badatelný je „práce navíc“ (CSV s tweety + CSV se seznamem obrázků, obrázky v .zip + metadata – createdb.sql, list.txt., queries.xml)
- › Balíček SIP (z volných souborů) může mít i jednodušší podobu např. jen CSV s tweety + obrázky; xml + obrázky; html+obrázky
- › Je žádoucí mít v SIP tweety ve více reprezentacích?

Výsledky/zjištění



- › Pro ÚŘO je zpětné stažení 3250 tweetů a 2500 obrázků ve většině případů dostatečné
- › Diskuse pod tweety lze získat jen s premium účtem
- › Pravidelná sklizeň je funkční a zcela dostatečná
- › Zpětná sklizeň není dostatečná pro vysoce aktivní uživatele a twitter maniaky
- › Např. na [stackoverflow](#) jsou možnosti jak zobrazovat průběžná tabulková data v podobě .html

- › Rekordman/ka?



[KarelHavlicek](#)

A jak dál?



1. Archivy využijí námi připravené nástroje
 - › Vhodné/nutné se shodnout na podobě AIP
 - › Rozhodujete, sklízíte, prezentujete sami
2. „Archiv českých účtů Twitteru“ spravovaný NAČR/NDA
 - › Archivy využívají infrastrukturu NAČR/NARP
 - › Archivy průběžně navrhuji účty do sklizně
 - › O sklizení rozhoduje NAČR
 - ✓ Jednotná struktura AIP
 - ✓ Prohlížecí infrastruktura součástí NARP



Archivace Twitteru

Děkuji za pozornost



Národní
archivní
portál



Národní archiv

Ikony: www.icons8.com

Archivy, knihovny, muzea v digitálním světě 2022