

Projekt s názvem: **Centrum pro podporu obyvatelstva pro případ skutečného
nebo domnělého vzniku mimořádných jaderných a
radiačních událostí**
s identifikačním kódem: **VJ01010116:**

Název předkládaného výsledku:
**Specializovaná databáze mediálních informací shrnujících
příspěvky ze sociálních sítí a výstupy tradičních médií**

Typ výsledku dle UV č. 837/2017	Evidenční číslo (příjemce)	Rok vzniku
(V _{souhrn}) Souhrnná výzkumná zpráva	13/2023.	2023
ISBN-ISSN	Webový odkaz na výsledek	Kde a kdy publikováno
-	bude vyvěšena na: https://www.suro.cz/cz/vyzkum/vysledky/	Po odsouhlasení aplikačním garantem SÚJB

Anotace k výsledku:

Souhrnná zpráva představuje databázi mediálních informací, shrnující příspěvky ze sociálních sítí a výstupy tradičních médií. Jsou zde popsány zdroje dat pro tuto databázi, jejich automatický sběr a následné automatické nebo poloautomatické zpracování. Odkazuje na způsob, jakým jsou data zpracovávána a ukládána pro interní potřeby SÚRO, a shrnuje smysl a cíle, pro které je databáze budována v rámci projektu

Manažer: Jaroslava Merunková

Řešitelský tým:

Jiří Hůlka, Michal Jankovec, Ivana Fojtíková

OBSAH

1	SOUHRN	3
2	ÚVOD	4
3	DATA A JEJICH SBĚR	5
3.1	ZDROJE PŘÍSPĚVKŮ	5
3.2	IDENTIFIKACE PŘÍSPĚVKŮ.....	6
3.3	TÉMATA A KLÍČOVÁ SLOVA	7
3.4	DODAVATELE MEDIÁLNÍHO OBSAHU	8
3.5	SEZNAM POŽADOVANÝCH ATRIBUTŮ OD DODAVATELE DAT	9
4	PŘEVOD PŘÍSPĚVKŮ DO DATABÁZE A JEJICH ZPRACOVÁNÍ.....	10
4.1	FREKVENCE SBĚRU DAT	11
4.2	DEDUPLIKACE DAT	11
4.3	STRUKTURA DATABÁZE	11
4.4	GOOGLE DATA STUDIO	12
4.5	PIPELINE DIAGRAM	12
5	ZÁVĚR	14

1 SOUHRN

Tato výzkumná zpráva vznikla po 3. roce řešení výzkumného úkolu v rámci projektu Centrum pro podporu obyvatelstva pro případ skutečného nebo domnělého vzniku mimořádných jaderných a radiačních událostí. Popisuje význam a cíle budování databáze, která si vytýčila ambice shrnovat veškeré obsahy českého mediálního prostředí, týkající se tématu jaderné bezpečnosti a radiační ochrany. Zahrnuje mediální příspěvky ze sociálních sítí i výstupy tradičních médií. V rámci projektu je jejím hlavním cílem stát se podkladem pro výstup projektu Výstavba a udržování modelů autodetekce změn a predikce šíření zpráv v období krizového vývoje a potenciál příspěvků generovat vznik obav, stresu a panického chování.

Ve zprávě jsou popsány zdroje dat pro předmětnou databázi, jejich automatický sběr a následné automatické nebo poloautomatické zpracování. Odkazuje na způsob, jakým jsou data zpracovávána a ukládána pro interní potřeby SÚRO, a shrnuje smysl a cíle, pro které je databáze budována v rámci projektu. Zpráva bude sloužit po ukončení projektu k údržbě databáze, jež bude provozována a dále rozvíjena na SÚRO s tím, že výsledky analýz budou k dispozici Státnímu úřadu pro jadernou bezpečnost jako aplikačnímu garantu pro řízení komunikace.

2 ÚVOD

Jedním z výstupů projektu IMPAKT je softwarová aplikace pro monitoring a analýzu zpráv v oblasti radiační ochrany a jaderné bezpečnosti spojující následující základní funkcionality:

- Provádění automatického importu mediálních informací (příspěvků) z databáze třetí strany do interní databáze SÚRO zahrnující příspěvky vztahující se k problematice radiační ochrany a jaderné bezpečnosti, identifikované na základě předdefinovaných slov.vyčištění dat (tj. například odstranění duplicit, upozornění na porušení číselníků atp.). Databáze třetí strany je denně aktualizována a obsahuje příspěvky vztažené k problematice radiační ochrany a jaderné bezpečnosti, identifikované na základě předdefinovaných klíčových slov. Kromě příspěvků s identifikovanými klíčovými slovy obsahuje navíc komunikační vlákna vztažená k těmto příspěvkům. Předpokládá tedy, že software bude automaticky stahovat identifikované příspěvky z databáze třetí strany ve struktuře optimalizované pro další použití v rámci projektu. Interní databáze SÚRO již obsahuje historické záznamy v definované struktuře, stejně jako automatické stahování dat. Nový systém bude navazovat na tento již existující systém a může ho ve svém řešení plně nebo částečně využít.
- Identifikace míry relevance pořízených příspěvků k tématům jaderné bezpečnosti a radiační ochrany a ohodnocení příspěvků podle míry sentimentu (obav, které daný příspěvek vzbuzuje u běžné populace) podle zkonstruovaných modelů relevance.
- Umožnění exportu dat z interní databáze na základě definovaných filtrů a výběrových podmínek.
- Příprava reportovací vrstvy nad interní databází, umožňující sledovat trendy a anomálie v časových řadách nasbíraných příspěvků prostřednictvím vhodného grafického rozhraní.
- Zajištění systému včasného varování, tj. komunikační vrstvy umožňující zasílání informací o anomáliích v datech identifikovaných systémem.

Zcela zásadním předpokladem pro vytvoření celého softwaru pro monitoring a analýzu zpráv v oblasti radiační ochrany a jaderné bezpečnosti je tak existence zabezpečené

databáze shrnující mediální příspěvky z daného oboru splňující definovaná kritéria a umožňující pravidelnou aktualizaci, ideálně automatizovanou, případně alespoň poloautomatizovanou.

Vytvořená databáze bude sloužit následujícím hlavním cílům:

- Skladování historických příspěvků
Z důvodu omezení na straně dodavatelů dat (hlavně omezení na množství a časovou dostupnost příspěvků) tato databáze umožňuje skladovat příspěvky za dlouhé časové období. K těmto historickým datům je umožněn přístup skrze rozhraní databáze.
- Zdroj dat pro systém včasného varování o nežádoucích jevech v mediálním prostředí.
- Zdroj dat pro vývoj modelů relevance a sentimentu.
- Zdroj dat pro sledování dlouhodobých trendů rozvoje mediálního prostředí a interakce na sociálních sítích.

3 DATA A JEJICH SBĚR

Systematické budování databáze mediálních příspěvků týkajících se oblastí vztahujících se k jaderné problematice a radiační ochraně bylo započato na SÚRO v roce 2020 v rámci institucionální podpory. Budovaná databáze je dále rozvíjena a rozšiřována o aktuální příspěvky a články.

3.1 Zdroje příspěvků

Každý zachycený příspěvek pochází z nějakého zdroje, neboli média. Tyto zdroje jsou pak dále rozděleny do několika kategorií, neboli typů zdroje:

- tisk,
- televize,
- rozhlas,
- online,
- sociální sítě,

- ostatní sociální média.

Do kategorie online patří zpravodajské a jiné informační portály a weby. Kategorie sociální sítě je definována výčtem: facebook.com, instagram.com, twitter.com, youtube.com, tiktok.com. Do kategorie ostatní sociální média pak patří diskuzní fóra, diskuze na zpravodajských webech atd (tj. online jsou články, ostatní sociální média jsou diskuze mimo sociální sítě).

3.2 Identifikace příspěvků

Vyhledávání příspěvků, které jsou zahrnuty v databázi, je založeno na dvou principech:

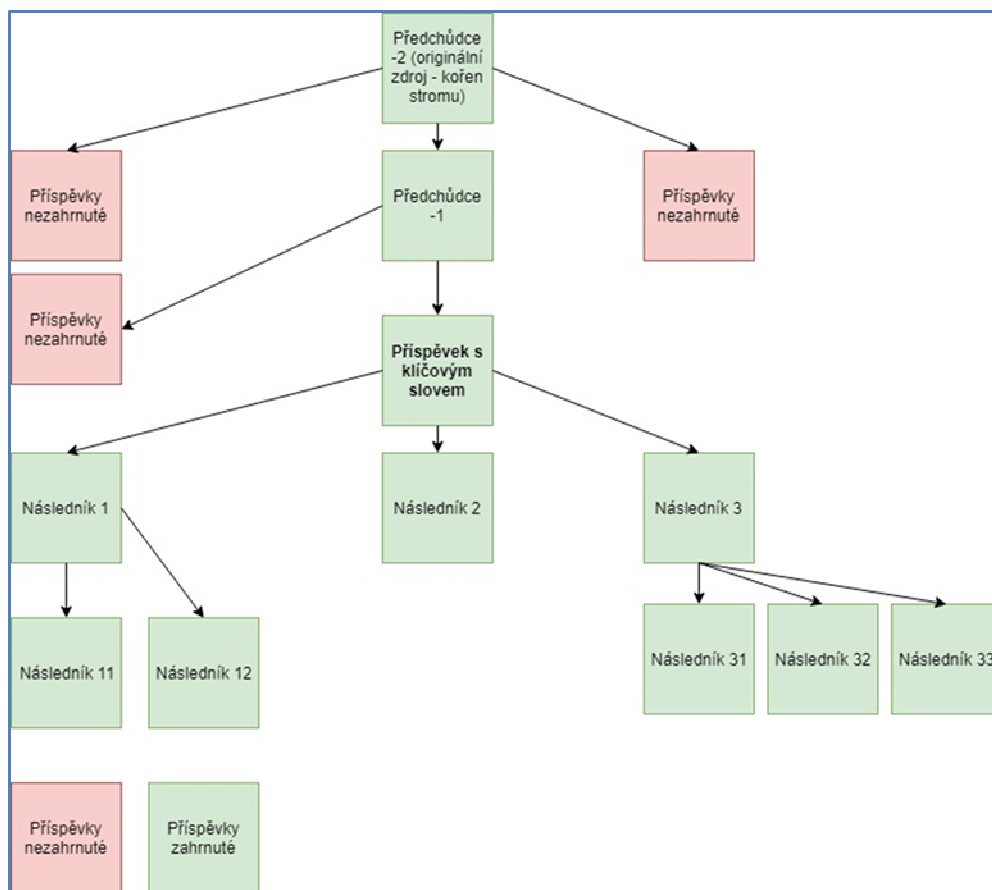
- Prvotní identifikace - vyhledávání podle klíčových slov

Prohledává se všechnen obsah v monitorovaných zdrojích, do databáze se ukládají pouze příspěvky obsahující specifikovaná klíčová slova/slovní spojení (Konkrétní specifikace klíčových slov je k dispozici v podkapitole 2.2).

- Druhotná identifikace – přidání příspěvků na základě vlákna

Do databáze se ukládají i příspěvky, které se nacházejí ve stejném vlákně jako příspěvky identifikované na základě klíčových slov (viz bod 1). Příspěvky náležejí do stejného vlákna, pokud mezi nimi existuje logická návaznost (např. vlákno komentářů na Facebooku nebo Twitteru, navzájem navazující komentáře na zpravodajských portálech atd.).

Ukládají se všechny příspěvky v daném vlákně. Pro některé zdroje je pro vlákno zachycena návaznost příspěvků (stromová struktura vlákna), pro některé zdroje se ukládá pouze seznam všech příspěvků ve vlákně bez zachycení návaznosti jednotlivých příspěvků. Od března 2022 se struktura zachovává pro následující zdroje: Facebook.com, Twitter.com, Youtube.com, TiTok.com. (viz Obr. 1).



Obr. 1: Schéma zahrnování příspěvků.

3.3 Témata a klíčová slova

Do databáze jsou ukládány příspěvky pro 9 identifikovaných témat souvisejících s radiační ochranou a jadernou bezpečností. Pro každé téma jsou definována klíčová slova:

1) téma - lékařská:

(záření or ozáření or ozařování or dávka or radiace or nukleární or radioaktivita or radioaktivní) and (RTG or rentgen or terapie or diagnostika or vyšetření or léčba or medicína or rezonance or CT or (mutace and buňka) or (genetický and buňka) or rakovina or nemocnice or pacient)) or (proton and (terapie or léčba)) or mamograf

2) téma - příroda:

(záření or ozáření or ozařování or (dávka and (záření or ozáření or radioaktivní)) or radiace or nukleární or radioaktivita or radioaktivní) and (letadlo or houby or

potravin y or kosmické or (př írodní and pozadí) or radionuklid or izotop or ((radioaktivní or radioaktivita) and (potravin y or jídlo)))

3) téma - havárie:

(radioaktivní or radioaktivita or ionizující or nukleární or jaderný or radiační) and (havárie or porucha or nehoda or ovzduší or mrak or kontaminace or únik or jod or cesium or spad or tragédie)

4) téma - medializované havárie a nehody:

(Černobyl or černobylská or Fukušima or Fukushima or Temelín or Dukovany or Goiânia or (Jaslovské and Bohunice)) and (havárie or nehoda or dávka or radiace or ozáření or záření or tragédie or únik)

5) téma - korporátní témata:

SÚJB OR „Státní úřad pro jadernou bezpečnost“ OR SÚRO OR „Státní ústav radiační ochrany“ OR Drábová

6) téma - radon:

radon or radonový

7) téma - úložiště jaderného odpadu

(radioaktivní or radioaktivita or ionizující or jaderný or radiační) and (odpad or úložiště)

8) téma - české jaderné elektrárny

(jaderná OR elektrárna) AND (Temelín OR Dukovany)

9) téma – ostatní

radioaktivní or radioaktivita or ionizující or jonizující or nukleární or jaderný or radiační or izotop

3.4 Dodavatelé mediálního obsahu

Budování databáze navazovalo na aktivity ve druhé polovině roku 2019 týkající se sledování vývoje mediálního obrazu, kdy byla zahájena spolupráce s firmou Monitora a s firmou Dataweeps. Oba dodavatelé dat byli vybráni na základě široké rešerše na

našem trhu: firma Monitora pro šíři pokrývaných typů médií (TV, rozhlas, tištěná média, online média včetně diskusí pod příspěvky) a ochotu umožnit přístup k datovým zdrojům přes rozhraní API a přizpůsobit své výstupy našim požadavkům při budování databáze, firma Dataweps byla tehdy jediným dodavatelem obsahu sociálních médií na trhu.

Dodavateli obsahu byly do března 2021 společnosti Monitora a Dataweps, od tohoto měsíce dále pouze společnost Monitora, která převzala monitorování sociálních sítí od firmy DataWeps.

Na základě předchozích zkušeností byly v rámci projektu IMPAKT navrženy metodické změny při přípravě nových databázových struktur. Hlavní metodickým vylepšením, které se v databázi promítlo, je skutečnost, že se neukládají pouze příspěvky identifikované definovanými klíčovými slovy, ale celé vlákno příspěvků, které se ke klíčovanému příspěvku vztahují. K tomuto metodickému kroku bylo přistoupeno zejména z toho důvodu, že singulární pohled, tj. pohled pouze na jednotlivý příspěvek vytržený z kontextu znemožňuje hlubší analýzu nejen významu, ale i dopadu navázané komunikace na sociálních sítích a obecně v mediálním prostoru. Vlákna lze přesně definovat zejména v internetových příspěvcích a diskusích pod nimi, nicméně z textu těchto příspěvků lze případně v konkrétních situacích odhalit i vztahy k neinternetovým typům médií.

3.5 Seznam požadovaných atributů od dodavatele dat

Od dodavatele dat jsou kromě samotného textu příspěvku požadovány následující atributy

- ID příspěvku,
- autor (pokud je dostupný),
- text,
- ID otce ve vlákne (pokud existuje, nedefinovaná hodnota, pokud jde o kořen stromu),
- ID vlákna,

- identifikace, jestli se jedná o příspěvek s klíčovým slovem nebo následníka nebo předchůdce,
- zdroj (číselník zdrojů),
- odkaz na daný příspěvek (adresa),
- informace o uživatelské interakci (počet likes, views atd.), co nejvíce atributů, které jsou pro daný zdroj dostupné (s časovým rozlišením podle okamžiku sběru dat – viz Frekvence sběru dat),
- datum vytvoření příspěvku,
- datum stažení příspěvku do databáze,
- další atributy dostupné v systému dodavatele (specifikace).

4 PŘEVOD PŘÍSPĚVKŮ DO DATABÁZE A JEJICH ZPRACOVÁNÍ

Proces stahování dat a veškeré související výpočty jsou prováděny na virtuálním serveru v síti CESNET. Celý proces je realizován pomocí automatizovaných skriptů v programovacím jazyce Python. Pro uložení dat je využívána PostgreSQL databáze, jež je také provozována na virtuálním serveru.

Virtuální server je ve správě SÚRO, pro připojení za účelem administrace procesu stahování dat a správy databáze je využíván protokol SSH se zabezpečením pomocí soukromého klíče.

Seznam použitých technologií a verzí (aktuální stav):

- Virtuální server s operačním systémem Linux Debian, ver. 11
- Programovací jazyk Python, ver. 3.9.2
- Prostředí pro automatizaci Python skriptů Airflow, ver. 2.2.4
- Relační databáze PostgreSQL, ver. 13.7
- Reportovací systém Google Data Studio

4.1 Frekvence sběru dat

Příspěvky identifikované dle klíčových slov jsou prostřednictvím API přístupu poskytovatele dat stahovány automaticky každou hodinu tak, aby v databázi byla vždy aktuální data.

Příspěvky z vláken jsou stahovány jednou denně, vždy pro určitý den 7 dní zpět; důvodem tohoto zpětného stahování je snaha zachytit celé vlákno (po 7 dnech je drtivá většina vláken již uzavřená).

Přesný rozvrh stahování (tento rozvrh může být potenciálně upravován dle potřeby):

U příspěvků zachycených dle klíčových slov probíhá stahování příspěvků pro daný den každou celou hodinu v rozmezí od 8:00 do půlnoci, a dále v 7:00 následujícího dne.

Pro vlákna proběhne každý den v 5:00 stažení příspěvků z vláken pro den 7 dní před aktuálním datem.

4.2 Deduplikace dat

Dodavatel dat Monitora provádí tzv. deduplikaci příspěvků pro téma Ostatní. Principem deduplikace je z tématu Ostatní vyřadit příspěvky, které zároveň patří do některého ze zbývajících 8 témat. Deduplikace je pro příspěvky z daného dne provedena v čase 6:55 dne následujícího. Tomuto odpovídá rozvrh stahování dat popsany výše, kdy se pro daný den naposledy stahují data v čase 7:00 dne následujícího; tímto je zajištěno, že data uložená v databázi jsou již po deduplikaci.

Duplicity napříč tématy mohou nastat pro jakákoli z 9 témat, deduplikace se ale provádí jen pro téma Ostatní. Smyslem tématu Ostatní je totiž zachytit ta témata, která nejsou pokryta žádným ze zbývajících témat, duplicity jsou proto nežádoucí.

4.3 Struktura databáze

Databázi lze rozdělit na dvě hlavní části: produkční tabulky (tabulky využívané v rámci procesu stahování dat) a tabulky pro tvorbu modelů:

Produkční tabulky sestávají ze tří vrstev: vrstvy původních dat, vrstvy zpracování dat a reportovací vrstvy.

Tabulky využívané pro tvorbu modelu zahrnují články a jejich kódování a tabulky s údaji popisujícím vstupní data pro modely (Exploratory data analysis).

4.4 Google data Studio

Tabulky z reportovací vrstvy databáze včetně tabulky s hodnotami systému včasného varování jsou zobrazovány v reportu v prostředí Google Data Studio. Pro načtení tabulek je využíván PostgreSQL konektor od společnosti Google, jež je v systému Google Data Studio dostupný. Aktualizace tabulek probíhá se stejnou frekvencí jako stahování dat, tedy jednou za hodinu.

V samotném reportu jsou tedy data zrcadlena z databáze a zobrazena v grafech, již nedochází téměř k žádným výpočtům nad daty, s několika výjimkami: agregace dle času (např. denní řadu počtu příspěvků zobrazit po týdnech), výpočet procent (skupinové sloupcové grafy).

4.5 Pipeline diagram

Celý proces stahování dat je znázorněn na Obr. 2.

Obdélníky znázorňují jednotlivé tabulky v databázi. Žlutě jsou vyznačené tabulky, jež slouží jako podklad pro grafy v reportovacím systému, zeleně je pak zvlášť vyznačena klíčová tabulka s podkladovými daty pro systém včasného varování (EWS), jež je součástí reportingu.

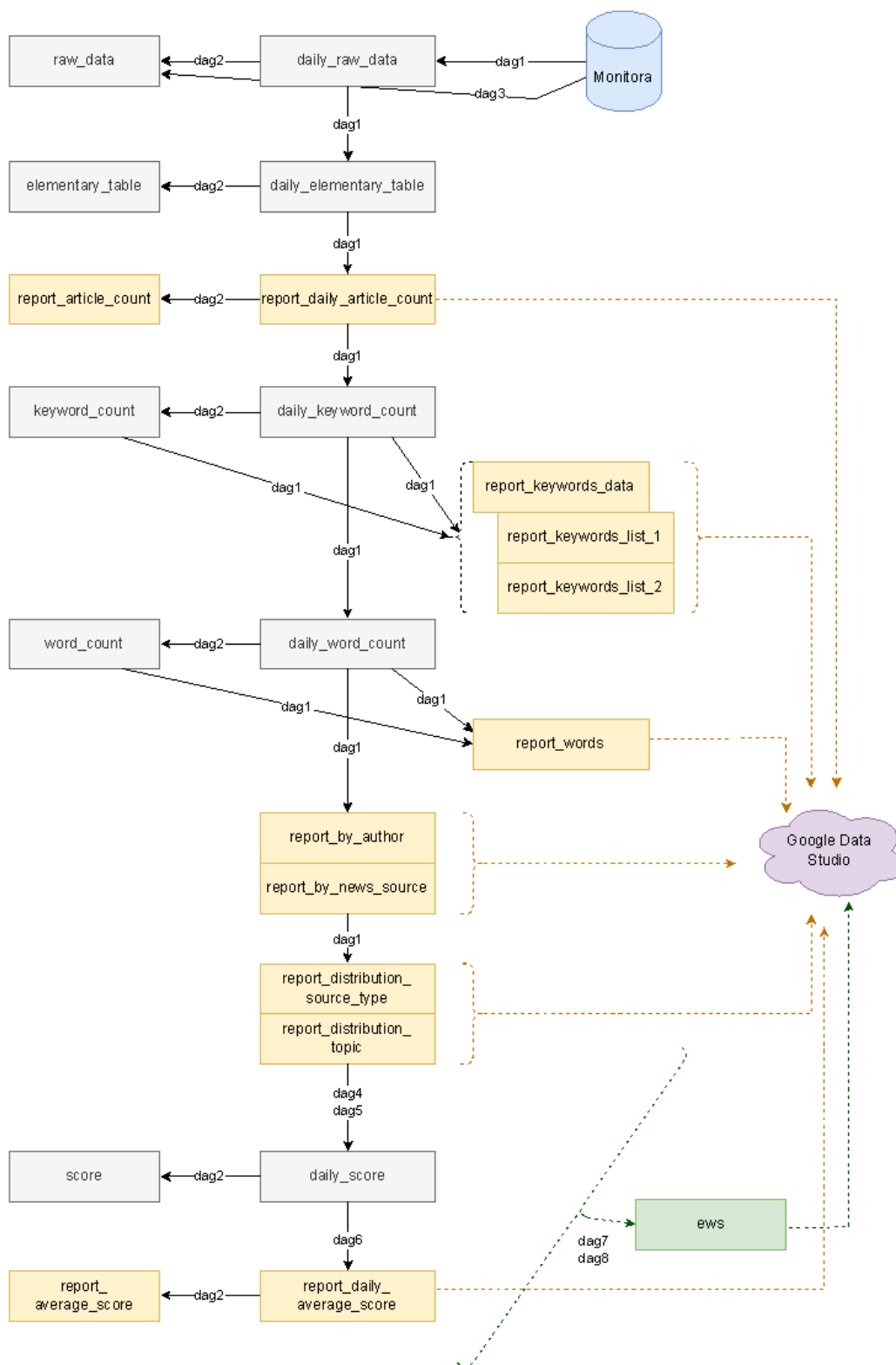
Šipky vyznačují toky dat; označení dag1,...,dag8 odpovídá identifikaci jednotlivých úloh v rámci automatického procesu stahování dat. Přehled úloh ukazuje Tabulka 1.

Tabulka 1: Úlohy řešené v rámci stahování dat

dag1	Hlavní část procesu stažení a zpracování dat a výpočet většiny tabulek pro report
dag2	Přenos dat z denních tabulek do hlavních tabulek
dag3	Stažení vláken
dag4	Výpočet skóre relevance pro nově stažené příspěvky
dag5	Výpočet skóre sentimentu pro nově stažené příspěvky
dag6	Výpočet tabulky průměrného skóre pro report
dag7	EWS: statistiky s hodinovou frekvencí

dag8

EWS: statistiky s denní, týdenní nebo měsíční frekvencí



Obr. 2.: Přehled datových toků při stažení a zpracování dat.

5 ZÁVĚR

Připravená databáze, jejíž geneze, popis, postupy vytváření, ukládání a úprav jsou popsány v této souhrnné zprávě, má primárně za cíl napomoci dosažení cílů v rámci tohoto projektu, a to sloužit jako:

- Zdroj dat pro systém včasného varování o nežádoucích jevech v mediálním prostředí.
- Zdroj dat pro vývoj modelů relevance a sentimentu.

Kromě toho nalezne využití i nad rámec projektu. Jedním z možných využití je “skladování” historických příspěvků z oblasti jaderné bezpečnosti a radiační ochrany. Z důvodu omezení na straně dodavatelů dat (hlavně omezení na množství a časovou dostupnost příspěvků) je tato databáze rezervoárem pro zabezpečené skladování mediálních obsahů za dlouhé časové období. K těmto historickým datům bude umožněn přístup skrze rozhraní databáze. Využití takových historických dat umožní sledovat dlouhodobé trendy rozvoje a kultivace mediálního prostředí a interakce na sociálních sítích v zájmových oblastech.