# Příloha B – Funkční a technická specifikace

**Investiční akce**
**„Software pro monitoring a analýzu zpráv v oblasti RO a JB s modulem pro systém včasného varování“**

**Zhotovitel** je označení účastníka zadávacího řízení nabízející své řešení požadovaného komplexního Software pro monitoring a analýzu zpráv v oblasti RO a JB s modulem pro systém včasného varování.

# 1 Požadavky na rozsah Díla

Cílem plnění veřejné zakázky je vytvoření software pro monitoring a analýzu zpráv v oblasti RO (radiační ochrany) a JB (jaderné bezpečnosti) s modulem pro systém včasného varování. Součástí software je analytický modul, který predikuje relevanci a sentiment (míru obav) pořízených zpráv.

## 1.1 Popis stávajícího stavu

V současné době SÚRO vlastní databázi mediálních zpráv (dále též příspěvků), týkající se problematiky RO a JB (dále „databáze SÚRO“). Tyto mediální zprávy jsou pořízeny na základě monitoringu médií a sociálních sítí, prováděných třetí stranou (historicky společnosti Dataweps a Monitora, v současnosti společností Monitora). Příspěvky jsou identifikovány na základě klíčových slov a rovněž jsou jim přiřazena témata z oblastí RO a JB (jeden příspěvek může být zařazen do více témat). Seznam klíčových slov, jejich kombinací a odpovídající témata jsou uvedena v **Apendix 1** – klíčová slova a tématické okruhy. V **Apendix 2** – Atributy v databázi SÚRO jsou hlavní atributy poskytované třetí stranou. Kromě příspěvků identifikovaných klíčovými slovy jsou v databázi komunikační vlákna vztažená k danému příspěvku. Logika zařazení příspěvku je popsána **v Apendix 3**. Příspěvky stažené třetí stranou jsou přístupné pomocí webové služby.

Příspěvky jsou systematicky ukládány v databázi SÚRO od března 2019. Data nejsou systematicky čištěna, jedná se o kopii a archivaci dat poskytovaných historicky třetí stranou prostřednictvím webové služby. Databáze je uložena v cloudovém prostředí CESNET.

Kromě databáze příspěvků k problematice JB a RO může SÚRO pro projekt poskytnout databázi příspěvků týkajících se problematiky COVID. V této databázi jsou mediální příspěvky sbírané v období, kdy probíhala pandemie COVID a obsahuje milióny příspěvků na základě výběru podle specifikovaných klíčových slov. Tato databáze může být případně využita jako podpůrná pro konstrukci modelů sentimentu (viz kapitola 2.2).

SÚRO má dále k dispozici základní software pro dynamické stahování příspěvků z databáze Monitora (dále „SOFTMON“). SOFTMON v pravidelných intervalech 1 hodina stahuje příspěvky z databáze Monitory a aktualizuje databázi SÚRO. Jednou denně je pak provedena finální uzávěrka dat za předchozí den. Nad databází SÚRO bude k datu zahájení projektu připraven základní reportovací nástroj ve webovém prostředí (nezávazně se předpokládá Google Data Studio), které zprostředkovává online aktuální informaci týkající se příspěvků JB a RO na základě databáze SÚRO.

V rámci pilotní přípravy na projekt SÚRO připravilo pilotní modely hodnocení relevance příspěvků a sentimentu (měření míry obav, které příspěvek vzbuzuje). Realizátor projektu dostane tyto pilotní výstupy k dispozici pro jejich zohlednění (nezávazné) ve vlastních budovaných modelech. Výstupy pilotních modelů relevance a sentimentu budou ukládány v databázi SÚRO pro každý příspěvek, který byl pořízen po datu implementace modelů (předpokládá se zhruba od července 2022).

Realizátor projektu dostane na začátku projektu od SÚRO k dispozici kompletní dokumentaci k pilotním modelům, včetně doporučené business specifikace výsledného díla. Realizátor projektu nebude zavázán se dodanou dokumentací řídit, nicméně tato dokumentace může sloužit jako základní pro porovnání přínosu realizátora na základě jím navržených modelů a specifikace výsledného díla.

## 1.2 Základní funkcionality dodávaného díla

Dodané dílo je softwarová aplikace spojující následující základní funkcionality:

i. Automatický import mediálních informací (příspěvků) z databáze třetí strany do interní databáze SÚRO, včetně vyčištění dat (tj. například odstranění duplicit, upozornění na porušení číselníků atp.). Databáze třetí strany je denně aktualizována a obsahuje příspěvky vztažené k problematice RO a JB, identifikované na základě předdefinovaných klíčových slov. Kromě příspěvků s identifikovanými klíčovými slovy obsahuje 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.[1]
   Interní databáze SÚRO již obsahuje historické záznamy v definované struktuře, stejně jako automatické stahování dat. Zhotovitel bude navazovat na tento již existující systém a může ho ve svém řešení plně nebo částečně využít.
ii. Identifikaci míry relevance pořízených příspěvků k tématům jaderné bezpečnosti a radiační ochrany. Příspěvky dále budou ohodnoceny mírou sentimentu (mírou obav, který daný příspěvek vzbuzuje u běžné populace). Modely relevance i sentimentu (předpokládají se vhodně zvolené modelovací metody z oblasti matematické statistiky a strojového učení) budou nově zkonstruovány v rámci projektu a mohou využít dílčích zkušeností z pilotních modelů, jejichž dokumentace bude předána realizátorovi v okamžiku zahájení projektu.
iii. Umožnění exportu dat z interní databáze na základě definovaných filtrů a výběrových podmínek. Zhotovitel navrhne business specifikaci a po jejím odsouhlasení Objednatelem realizuje v rámci softwarového řešení.
iv. Reportovací vrstvu nad interní databází, umožňující sledovat trendy a anomálie v časových řadách nasbíraných příspěvků prostřednictvím vhodného grafického rozhraní. Zhotovitel navrhne business specifikaci a po jejím odsouhlasení Objednatelem realizuje v rámci softwarového řešení.
v. Odlišení typů přístupu podle rolí uživatelů definovaných SÚRO. Zhotovitel navrhne business specifikaci a po jejím odsouhlasení Objednatelem realizuje v rámci softwarového řešení.
vi. Systém včasného varování, tj. komunikační vrstvu umožňující zasílání informací o anomáliích v datech identifikovaných systémem. Systém musí identifikovat náhlé změny v trendech časových řad, odlehlá pozorování a jiné anomálie identifikované na základě statistické analýzy podkladových dat a informovat o nich pomocí vhodných komunikačních nástrojů (reportovací vrstva, sms, email, atp.). Zhotovitel navrhne business specifikaci a po jejím odsouhlasení Objednatelem realizuje v rámci softwarového řešení.

Zcela zásadní podmínkou pro vytvoření software je vytvoření modelů relevance a sentimentu a jejich integrace v dodaném softwarovém řešení. Model relevance bude odhadovat, jak moc souvisí daný příspěvek s tématikou radiační ochrany a jaderné bezpečnosti. Model sentimentu bude odhadovat, jak velké obavy může příspěvek vzbuzovat v běžné (neodborné) veřejnosti. Metodika výstavby modelů bude specifikována Zhotovitelem. V rámci projektu se předpokládá sběr dat nutných pro

[1] Přístup k databázi, její strukturu a potřebná práva přístupu zajistí SÚRO.

výstavbu modelů, a to minimálně v rozsahu specifikovaným smlouvou, ohodnocených z hlediska relevance speciálně proškolenými pracovníky (proškolení pracovníků a hodnocení příspěvků je na straně Zhotovitele). Z takto určených relevantních příspěvků budou ohodnoceny všechny relevantní příspěvky běžnou populací z hlediska sentimentu (výběrové šetření je součástí dodávaného řešení). Zhotovitel může navrhnout i jinou logiku konstrukce modelů, vždy však s garancí konstrukce minimálně uvedeného schématu.

Součástí zakázky je implementace software, otestování software, proškolení uživatelů, provoz a zajištění servisních služeb po dobu projektu a celková detailní dokumentace k systému a modelům.

## 1.3 Předpokládané plnění

**Plnění se předpokládá postupně od zahájení projektu, včetně údržby systému po dobu 4 let:**

1. Importovací model propojující databázi třetí strany a automatické stahování dat. Lze využít a aktualizovat stávající systém. Předpokládá se udržování importů a kontrola kvality vstupních dat po dobu trvání projektu. (termín: udržování stávajícího systému okamžitě po převzetí Zhotovitelem, případné úpravy do konce roku 2022)
2. Udržování a optimalizace interního datového modelu a udržování aktuálního propojení s databází třetí strany včetně potřebných aktualizací datových struktur na základě případných změn ve formátech a číselnících (termín: 2022-2023)
3. Specifikace metodiky sběru dat pro konstrukci modelů (výběrové vzorky) (termín: 2022)
4. Specifikace metodiky konstrukce modelů (popis konstrukce modelů včetně měření jejich přesnosti a predikční síly) (termín: 2022)
5. Pořízení výběru vzorků příspěvků nutných pro výstavbu modelu relevance (trénovací množina) (termín: 2022)
6. Kódování příspěvků trénovací množiny pro model relevance proškolenými pracovníky (termín: 2022)
7. Kódování příspěvků trénovací množiny, které byly označeny jako relevantní, a to z hlediska jejich sentimentu běžnou populací na základě výběrového šetření (termín: 2022)
8. Vlastní tvorba modelů relevance a sentimentu (termín: 2023, 1. pololetí)
9. Seznámení se s aktuální podobou reportovacího nástroje SÚRO a vytvoření vlastního návrhu grafického rozhraní (interface) webové aplikace nebo reportovací vrstvy umožňující exporty a prohlížení dat včetně nezbytných statistik. Systém lze použít k exportům dat z databáze SÚRO. Lze použít a aktualizovat stávající reportovací vrstvu SÚRO. (termín: Pilotní verze aplikace – 2022, celkové řešení – konec 2023)
10. Pořízení výběru vzorků příspěvků nutných pro testování a rekalibraci modelu relevance (testovací množina) (termín: 1. pololetí 2023)
11. Kódování příspěvků testovací množiny pro rekalibraci modelu relevance proškolenými pracovníky (termín: 1. pololetí 2023)
12. Kódování příspěvků testovací množiny, které byly označeny jako relevantní, a to z hlediska jejich sentimentu běžnou populací (termín: 1. pololetí 2023)
13. Testování a rekalibrace modelů relevance a sentimentu (termín: konec 2023)

14. Vytvoření systému včasného varování, tj. identifikaci anomálií v datech na základě statistické analýzy a zajištění komunikační vrstvy software, která umožní zasílání varování prostřednictvím specifikovaných komunikačních kanálů. (termín: pilot 1. pololetí 2023, finální řešení – konec 2023)
15. Otestování aplikace nebo reportovací vrstvy včetně systému včasného varování (termín: posledních 6 měsíců projektu)
16. Spuštění ostrého provozu aplikace (dva měsíce před ukončením projektu)
17. Udržování systému v letech 2024+.

Jednotlivé odstavce uvedené v tomto seznamu jsou podrobně rozepsány v následující kapitole.

# 2 Definované oblasti – detailní požadavky na plnění zakázky

## 2.1 Sběr dat mediálních příspěvků a tvorba podkladové databáze (navazuje na existující aktivity SÚRO)

### 2.1.1 Odstavce podle kapitoly 1.3

1. Importovací model propojující databázi třetí strany a automatické stahování dat. Lze využít a aktualizovat stávající systém. Předpokládá se udržování importů a kontrola kvality vstupních dat po dobu trvání projektu. (termín: udržování stávajícího systému okamžitě po převzetí Zhotovitelem, případné úpravy do konce roku 2022)
2. Udržování a optimalizace interního datového modelu a udržování aktuálního propojení s databází třetí strany včetně potřebných aktualizací datových struktur na základě případných změn ve formátech a číselnících (termín: 2022-2023)

### 2.1.2 Podrobnější popis

Zhotovitel se seznámí s aktuálním stavem (strukturou a architekturou) existující databáze příspěvků v externím cloudovém prostředí CESNET. Příspěvkem (hlavním) se rozumí mediální sdělení obsahující klíčová slova a jejich kombinace určené tematicky oblastí radiační ochrany (RO) a jaderné bezpečnosti (JB). Příspěvkem se také rozumí sdělení v komunikačním vlákně navazujícím na příspěvek hlavní pro ty příspěvky, kde je komunikační vlákno součástí vstupní databáze třetí strany. Identifikace příspěvků a jejich obsah, komunikační vlákna, mediální typ atp. budou i nadále poskytovány třetí stranou v dané datové struktuře pomocí webové služby (viz kapitola 5)[2]. V případě, že dojde v průběhu projektu ke změně zmíněné třetí strany případně i změny struktury dat, Zhotovitel musí zohlednit nové skutečnosti tak, aby byla zachována kontinuita dat v maximální možné míře.

Zhotovitel naváže na existující databázi příspěvků SÚRO. Předpokládá se proto využití stávající datové struktury (viz kapitola 7). V případě, že Zhotovitel navrhne jinou optimálnější strukturu, musí být součástí dodávky převedení historických dat do nové navržené struktury (která musí být odsouhlasena SÚRO).

Využití databáze se předpokládá zejména pro

- Konstrukci modelů relevance a sentimentu (viz odstavec ii, kapitola 1.2)
- Reportovací systém a identifikaci impulsů pro systém včasného varování (viz odstavec vi, kapitola 1.2).

Zhotovitel bude zodpovědný zejména za

- Automatické denní stahování s využitím externí webové služby a doplňování databáze příspěvků po dobu trvání projektu.
- Rozšíření datového modelu o výstupy obou modelů (hodnocení relevance a sentimentu příspěvků) a o datové struktury nutné pro efektivní výpočet modelů a jejich monitoring. Rozšíření datového modelu o atributy nutné pro systém včasného varování (viz 1.3, odstavec 12).

[2] Funkčnost a dostupnost webové služby bude zajištěna SÚRO prostřednictvím třetí strany Monitora.

### 2.1.3 Minimální požadavky

Účastník řízení popíše předpokládané databázové řešení a uvede svůj stupeň expertní znalosti (školení, specializovaná vysoká škola atp.) Uvede možnosti řešení automatizovaného stahování dat, včetně odhadovaných časových náročností (časový snímek).

1) Prokázaná znalost SQL
2) Prokázaná znalost principů API
3) Prokázaná znalost práce v cloudových prostředích
4) Zhotovitel se zaváže k pravidelné online (minimálně denní) aktualizaci databáze.

### 2.1.4 Kritéria hodnocení

1) Deklarace akceptace přístupu k datům třetí strany prostřednictvím API
2) Deklarace akceptace aktuální struktury databáze SÚRO (lze navrhnout optimálnější)
3) Garance aktualizace databáze do 24 hodin po aktualizaci informace v databázi třetí strany, včetně řešení aktualizace hodnocení příspěvků (likes, retweets atp.)

## 2.2 Metodika konstrukce modelů predikce relevance a sentimentu včetně metodiky sběru dat nutných pro konstrukci modelů (výběr příspěvků pro jejich kódování)

1.

### 2.2.1 Odstavce podle kapitoly 1.3

1) Specifikace metodiky sběru dat pro konstrukci modelů (výběrové vzorky) (termín: 2022)
2) Specifikace metodiky konstrukce modelů (popis konstrukce modelů včetně měření jejich přesnosti a predikční síly) (termín: 2022)

### 2.2.2 Podrobnější popis

Ke konstrukci modelů relevance, tj. ohodnocení každého příspěvku (případně vlákna příspěvků) pravděpodobností toho, že je daný příspěvek relevantní k tématice RO nebo JB je potřeba využít historické příspěvky databáze SÚRO. U vybraných příspěvků je pak potřeba identifikovat expertně, zdali je příspěvek relevantní. U relevantních příspěvků je potom potřeby na základě výběrového šetření v populaci ČR identifikovat, nakolik příspěvky vzbuzují obavy. V následující kapitole Modely popíšeme přesněji požadovaný výstup modelů, v navazující kapitole Kódovací a výběrové postupy popíšeme přesněji požadavky na metody výběru a hodnocení příspěvků.

Zhotovitel může pro konstrukci modelu sentimentu využít databázi COVID. Ověření její vhodnosti i pro příspěvky RO a JB je však nutné ověřit a kvantifikovat odhad vhodnosti jejího použití.

#### *2.2.2.1 Modely*

Zhotovitel vytvoří/popíše metodiků konstrukce modelů predikce relevance a sentimentu.

a) Modely prvního stupně hodnotí míru relevance příspěvku (pravděpodobnost, že se příspěvek týká jaderné a radiační problematiky, bez rozlišení témat). Hodnocení relevance (závisle proměnná modelů) probíhá na základě objektivního posouzení, tj. nikoliv jak je příspěvek vnímán běžnou populací, ale populací vědomě posuzující relevanci na základě odborné znalosti (profesionální kódování – viz dále v kapitole Kódovací a výběrové postupy).

b) Modely druhého stupně hodnotí míru obav (sentiment), které relevantní příspěvky mohou vzbuzovat u běžné populace. Závisle proměnná je tedy získána výběrovým šetřením na relevantních příspěvcích (viz dále). Klasifikaci míry obav navrhne Zhotovitel.
Pokud Zhotovitel bude chtít v rámci konstrukce modelu sentimentu využít i databázi COVID musí specifikovat způsob a velikost výběru příspěvků z této databáze a navrhnout metodu ověření platnosti modelu i pro příspěvky JB a RO.

Zhotovitel popíše možné modelovací přístupy a popíše předpokládaný proces modelování. Uvede minimální uvažované velikosti modelových vzorků pro oba stupně modelů (pro model relevance je minimální požadavek kódovaných příspěvků stanovený smlouvou, pro model sentimentu se předpokládá hodnocení všech příspěvků hodnocených jako relevantní případně rozšířených o výběr z databáze COVID.

Zhotovitel popíše výstupy modelů a způsoby měření jejich kvality. Popíše metodiku měření stability modelů a metodiku monitoringu modelů. Popíše metodiku a škálu hodnocení obav. Uvede předpokládanou[3] „životnost modelu“, tj. časový horizont, ve kterém budou modely splňovat minimální požadavky na jejich kvalitu bez nutnosti jejich kalibrace nebo nového vývoje.

### 2.2.3 Kódovací a výběrové postupy

Metodika konstrukce modelů bude předpokládat kódování statisticky relevantních vzorků ať již z pohledu relevance, tak z pohledu míry obav (sentimentu). Z tohoto pohledu je kritická správná statistická metodika vlastních výběrů z databáze příspěvků. Výběry pro modely sentimentů budou provedeny z databáze relevantních příspěvků případně doplněny o výběr z databáze COVID.

Ve stupni jedna, tedy pro modely relevance se jedná o kódování odborné, tj. kódování proškolenými pracovníky schopnými posoudit relevanci daných příspěvků s tématikou RO nebo JB. Ve stupni dva se jedná o organizaci reprezentativního výběrového šetření na populaci 18-65 let na relevantních příspěvcích s požadavky na hodnocení sentimentu.

Zhotovitel popíše možné přístupy týkající se kódovacích technik pro stupeň jedna, včetně způsobu kontroly kódování. Dále popíše doporučenou a v ceně zvažovanou metodu výběrového šetření ve stupni dva, předpokládaný rozsah dotazníku, odhadovanou návratnost a navýšení.

Zhotovitel popíše možné výběrové přístupy (algoritmy) pro účely získání vypovídajícího datového vzorku pro účely konstrukce modelů relevance a sentimentu.

Zhotovitel popíše způsoby kontroly kódovacích prací a výběrového šetření včetně jejich rozsahu.

Zhotovitel může kódovací práce a výběrová šetření řešit prostřednictvím odborně garantovaného subdodavatele[4].

[3] Zadavatel si je vědom, že bez konkrétních dat nelze požadované charakteristiky garantovat. Předpokládá se proto, že účastník řízení uvede údaje na základě svých zkušeností s podobnými typy modelovacích projektů.
[4] Garanci za cenu nese Zhotovitel.

### 2.2.4 Minimální požadavky

- Teoretická i praktická znalost metod matematické statistiky a strojového učení (prokázaná odbornost RNDr., Mgr. nebo ing. v příslušném oboru, ideálně PhD.)
- Praxe v oblasti datových analýz minimálně 5 let doložená seznamem úspěšně realizovaných projektů z oblasti datových analýz
- Garance kvality metodiky kódovacích a výběrových prací doložená úspěšně dokončenými projekty za posledních 5 let.
- Garance kvality celkového pojetí metodiky odborným sociologickým pracovníkem (vzdělání FSV UK nebo jiného odpovídajícího vysokoškolského zaměření)
- Garance výběrových postupů matematickým statistikem (vzdělání v oboru matematická statistika MFF UK nebo jiného odpovídajícího vysokoškolského zaměření).

### 2.2.5 Kritéria hodnocení

#### *2.2.5.1 Kvalita modelů*

- Předpokládané smluvně garantované hodnocení kvality modelu (např. Gini koeficient, míra chybného zařazení, Kolmogorov-Smirnov statistika nebo podobné).
- Metodika monitoringu modelů.
- Přístup ke konstrukci modelů a jejich adekvátnost k řešení problematiky odhadu měr relevance a sentimentu. Jasný popis postupu včetně zvažovaných matematických přístupů a formulí. Jasná specifikace výstupních proměnných modelů (měřítka míry relevance a míry sentimentu).

#### *2.2.5.2 Kvalita výběrů a kódovacích prací*

- Výběrové postupy uvedené v kapitole 2.2.3 v rozsahu, aby byly replikovatelné třetí stranou.
- Reprezentativita výběrů (z hlediska typu příspěvků pro model relevance, z hlediska struktury populace pro modely sentimentu). Detailnější specifikace a jednoznačnost popisu budou preferovány.
- Existence metodiky kódování a míra jejího detailu (jednoznačnost postupu kódování z hlediska mediálně vědního a sociologického a kvalita těchto postupů).
- Rozsah a kvalita navržených kontrol dodržení metodik kódování.

## 2.3 Kódovací práce a výběrové šetření

### 2.3.1 Odstavce podle kapitoly 1.3

3) Pořízení výběru vzorků příspěvků nutných pro výstavbu modelu relevance (trénovací množina) (termín: 2022)
4) Kódování příspěvků trénovací množiny pro model relevance proškolenými pracovníky (termín: 2022)

5) Kódování příspěvků trénovací množiny, které byly označeny jako relevantní, a to z hlediska jejich sentimentu běžnou populací na základě výběrového šetření (termín: 2022)

### 2.3.2 Podrobnější popis

Na základě metodiky vypracované na základě požadavku kapitoly 2.2, odstavec 3 se v další fázi musí provést vlastní kódovací práce pro model relevance a výběrové šetření pro model sentimentu (míry obav).

Zhotovitel uvede, zdali bude práce provádět vlastními silami nebo prostřednictvím subdodavatele.

Zhotovitel rozepíše podrobně cenu jednotlivých fází kódovacího procesu (cena za výběr z databáze, cena za kódovací práce, cena za kontrolu kódovacích prací atp.).

Zhotovitel rozepíše cenu za výběrové šetření pro model sentimentu (míry obav) v závislosti na zvolené metodě včetně kontrolních prací.

V případě, že Zhotovitel zvolí metodiku konstrukce modelu sentimentu (míry obav) i na základě databáze COVID, uvede zvlášť cenu za výběrové a kontrolní práce na tomto modelu.

### 2.3.3 Minimální požadavky

- Praxe v kódovacích pracích minimálně 5 let.
- Praxe v projektech výběrových šetření minimálně 5 let.
- V případě, že účastník bude na kódovací práce a vlastní výběrové šetření najímat třetí stranu, musí doložit její kvalitu a praxi včetně jejího vyjádření k obsahu a ceně dodávaných služeb.

### 2.3.4 Kritéria hodnocení

- Reálnost provedení kódovacích prací a výběrového šetření na základě navržených metodik.
- Stupeň smluvní garance dodržení navržených postupů včetně případné penalizace za nedodržení kvality.

## 2.4 Vlastní tvorba modelů

### 2.4.1 Odstavce podle kapitoly 1.3

6) Vlastní tvorba modelů relevance a sentimentu (termín: 2023, 1. pololetí)

### 2.4.2 Podrobnější popis

Na základě metodiky dle 2.2 se předpokládá provedení vlastní konstrukce modelů (relevance a sentimentu) včetně detailní dokumentace. Předpokládá se doložení kvality modelů a jejich stability. Předpokládá se konstrukce alternativních modelů a výběr optimálního modelu z množiny alternativních modelů. Metodika vytvořená v 2.2. může být zpřesněna nebo upravena na základě reálných dat.

Účastník řízení uvede předpokládané modelovací prostředí (modelovací systém, programovací jazyk), ve kterých bude provádět vlastní modelování a analýzu dat. Doloží, že v daném prostředí je možné provést veškeré modelovací práce podle návrhu modelovací metodiky. V případě, že není jasné, že uvedené modelovací prostředí neobsahuje potřebné prostředky, popíše, jakým způsobem je doplní.

### 2.4.3 Minimální požadavky

- Hluboká odborná znalost matematické statistiky a datového inženýrství, včetně metod umělé inteligence a strojového učení zajištěná odborným členem (datovým garantem) řešitelského týmu (předpokládá se znalost na úrovni absolventů statistických oborů a datového inženýrství na matematicko-fyzikální fakultě UK, FI ČVUT nebo VŠE, ideálně doplněná stupněm PhD.)
- Praxe pracovníků týmu v oblasti datových analýz minimálně 3 roky, úspěšně řešené odborné projekty v oblasti datových analýz ať již vědecky publikovaných nebo v rámci reálných aplikací v praxi.
- Znalost modelovacího prostředí.

### 2.4.4 Kritéria hodnocení

- Adekvátnost zvoleného modelovacího prostředí pro optimální konstrukci požadovaných modelů.
- Reálnost dodržení navržených kritérií kvality modelů.
- Stupeň smluvní garance dodržení kritérií kvality modelů včetně případné penalizace za nedodržení kvality.

## 2.5 Pilotní návrh na reportovací systém a systém včasného varování

### 2.5.1 Odstavce podle kapitoly 1.3

9. Seznámení se s aktuální podobou reportovacího nástroje SÚRO a vytvoření vlastního návrhu grafického rozhraní (interface) webové aplikace nebo reportovací vrstvy umožňující exporty a prohlížení dat včetně nezbytných statistik. Systém lze použít k exportům dat z databáze SÚRO. Lze použít a aktualizovat stávající reportovací vrstvu SÚRO. (termín: Pilotní verze aplikace – 2022, celkové řešení – konec 2023)

### 2.5.2 Podrobnější popis

V této fázi projektu se předpokládají práce na budování reportingového systému umožňujícího sledování statistik vývoje příspěvků v jednotlivých tématech RO a JB.

Systém bude vybudován jako webová aplikace – reportovací systém nad podkladovou databází příspěvků (viz výše). Lze využít i existující reportovací systémy (typu Microsoft PowerBI, Google Data Studio), vždy však s ohledem na celkovou funkcionalitu systému, včetně systému včasného varování, rozlišení přístupových práv atp. V případě využití komerčních platforem musí být ceny za případné licence součástí cenové nabídky. Zhotovitel uvede výhody a nevýhody zvoleného přístupu (vlastní webová aplikace oproti komerčnímu reportovacímu systému).

Systém umožní exportovat výstupy z podkladové databáze. Exporty lze provádět na základě specifikace filtrů zadaných prostřednictvím webové aplikace. Filtry budou navrženy Zhotovitelem a odsouhlaseny Objednatelem v rámci projektu. Filtrace musí být možná minimálně na základě datumu, zdroje dat, domény, typu média, typu příspěvku atp. Výhodou je implementace operace typu „dril down“ v reportovací vrstvě systému.

Systém musí umožní sledovat výstupy zkonstruovaných modelů v navržených segmentech. Systém bude s navrženou časovou intenzitou informovat o významných změnách ve vývoji sledovaných metrik a bude schopen vhodnou formou (sms, email, …) informovat o těchto změnách – tzv. systém včasného varování.

Předpokládá se, že pilotní verze systému bude odsouhlasena Objednatelem a jeho připomínky budou zohledněny ve finálním řešení.

Systém umožní přistupovat k různému typu výstupů a exportů podle role uživatele[5].

Správa systému bude prováděna Objednatelem po předání Zhotovitelem, který ale musí zohlednit přístupová práva a standardní bezpečnostní pravidla pro práci s podkladovou databází a vlastní softwarovou (reportovací aplikací).

Účastník řízení uvede orientačně předpokládané sledované charakteristiky (metriky) reportovacího systému a třídící charakteristiky (filtry a dimenze). Dále popíše způsoby reportování výstupu modelů a jejich charakteristik (tabulkové výstupy, grafické výstupy). Uvede možnosti uživatele pro práci s reportem (úprava dimenzí, exporty, atp.). Součástí navrhovaného řešení bude frekvenci updatu reportů a podkladových dat včetně způsobu řešení dynamicky se měnících charakteristik (např. hodnocení příspěvků sociálních sítí).

### 2.5.3 Minimální požadavky

1) Zkušenost s návrhy reportovacího systému doložená úspěšnými historickými realizacemi projektů členů týmu.

[5] Specifikaci rolí dodá Objednatel nejpozději v první polovině projektu.

2) Zkušenost s tvorbou webových aplikací nebo využíváním komerčních reportovacích systémů doložená úspěšnými historickými realizacemi projektů členů týmu.

### 2.5.4 Kritéria hodnocení

1) Rozsah navržených metrik a typů výstupních reportů (podstatná je informační hodnota návrhu s ohledem na potřeby sledování vývoje mediálních aspektů JB a RO).
2) Kvalita řešitelského týmu – ohodnocení garance dodávky.

## 2.6 Kódovací práce a výběrové šetření pro účely kontrolního vzorku.

### 2.6.1 Odstavce podle kapitoly 1.3

10. Pořízení výběru vzorků příspěvků nutných pro testování a rekalibraci modelu relevance (testovací množina) (termín: 1. pololetí 2023)
11. Kódování příspěvků testovací množiny pro rekalibraci modelu relevance proškolenými pracovníky (termín: 1. pololetí 2023)
12. Kódování příspěvků testovací množiny, které byly označeny jako relevantní, a to z hlediska jejich sentimentu běžnou populací (termín: 1. pololetí 2023)

### 2.6.2 Podrobnější popis

Předpokládá se provedení kódovacích prací stejně jako v kapitole 2.3 ale pouze pro kontrolní účely. Minimální požadavek na kódování z hlediska relevance je 1000 náhodně vybraných příspěvků. Výběrové šetření bude probíhat opět pouze na relevantních příspěvcích. V případě, že Zhotovitel navrhne jinou metodiku, musí prokázat, že splňuje minimálně uvedené potřeby pro hodnocení kvality modelů.

### 2.6.3 Minimální požadavky

Viz 2.3.3

### 2.6.4 Kritéria hodnocení

Viz 2.3.4

## 2.7 Kalibrace modelů na základě kontrolního vzorku

### 2.7.1 Odstavce podle kapitoly 1.3

13. Testování a rekalibrace modelů relevance a sentimentu (termín: konec projektu)

### 2.7.2 Podrobnější popis

Na základě testovacích dat pořízených podle 2.6 Zhotovitel provede hodnocení modelů vytvořených podle 2.4. Zhodnotí se změny ve vypovídajících parametrech modelů a případně se navrhne jejich rekalibrace.

Účastník řízení popíše přístup k hodnocení změn modelů mezi trénovací a testovací množinou. Uvede vztah k pravidelnému monitoringu podle jím navržené metodiky z požadavku 2.2.

### 2.7.3 Minimální požadavky

Viz 2.4.3

### 2.7.4 Kritéria hodnocení

Viz 2.4.4

## 2.8 Reportovací systém a systém včasného varování (finální verze)

### 2.8.1 Odstavce podle kapitoly 1.3

14. Vytvoření systému včasného varování, tj. identifikaci anomálií v datech na základě statistické analýzy a zajištění komunikační vrstvy software, která umožní zasílání varování prostřednictvím specifikovaných komunikačních kanálů. (termín: pilot 1. pololetí 2023, finální řešení – konec projektu)
15. Otestování aplikace nebo reportovací vrstvy včetně systému včasného varování a přístupových práv (termín: posledních 6 měsíců projektu)
16. Spuštění ostrého provozu aplikace (dva měsíce před ukončením projektu)

### 2.8.2 Podrobnější popis

Tento požadavek uzavírá celý projekt z hlediska všech jeho funkcionalit. Jedná se dokončení reportovací aplikace (ať již vlastní webové nebo komerční) a její uvedení do reálného provozu a její akceptaci SÚRO. Aplikace musí splňovat všechny požadavky podle specifikace schválené v rámci 2.5.

Aplikace musí projít otestováním na straně SÚRO. Zhotovitel popíše navrhovaný proces testování aplikace včetně systému pro podporu zachycení požadavků na opravy (change requests).

Předpokládá se, že změny oproti schválenému zadání podle 2.5. budou tzv. change requests. Zhotovitel uvede cenu za manday pro práce nad rámec schváleného zadání.

### 2.8.3 Minimální požadavky

Viz 2.5.3

### 2.8.4 Kritéria hodnocení

Viz 2.5.4. a dále cena za manday služeb („manday“ = 8 pracovních hodin) za požadavky nad rámec schválené specifikace vzniklých v průběhu projektu.

## 2.9 Údržba systému po ukončení projektu (2024+)

### 2.9.1 Odstavce podle kapitoly 1.3

17. Podmínky udržování systému v letech 2024+

### 2.9.2 Podrobnější popis

Předpokládá se, že systém bude budován postupně v rámci projektu. Pilotní verze bude dokončena v roce 2022 a finální systém začne být testován 18 měsíců od zahájení projektu. Ostrý provoz projektu se předpokládá nejpozději 24 měsíců od zahájení projektu. Po celou dobu projektu musí být udržována a aktualizovaná databáze příspěvků, jak je popsáno v kapitole 2.1.

Zhotovitel dále udržuje databázi a chod systému po dobu 4 let od zahájení projektu (za zahájení projektu se chápe podpis smlouvy).

### 2.9.3 Minimální požadavky

Funkčnost systému podle etap projektu popsaných v tomto Technickém dodatku. Plná funkčnost systému minimálně od třetího roku projektu, tj. podpisu smlouvy. Reakce na připomínky k funkčnosti systému podle této Technické specifikace a specifikací odsouhlasených v rámci projektu podle kapitoly 3.5.

### 2.9.4 Kritéria hodnocení

Splnění minimálních požadavků. Nabídky Zhotovitele nad rámec těchto minimálních požadavků z hlediska termínů dodání budou zohledněny (viz Výzva k podání nabídek a zadávací podmínky).

# 3 JINÉ POŽADAVKY A UPŘESNĚNÍ ZADÁNÍ

## 3.1 Jiné požadavky na dodávané řešení

Součástí dodávky „Software pro monitoring a analýzu zpráv v oblasti RO a JB s modulem pro systém včasného varování se rozumí komplex návrhu, dodání, instalace a implementace veškerého potřebného softwarového vybavení na všech úrovních, školení, dokumentace, ověřovací a produktivní provoz. Vše tak, aby byla zajištěna úplná a bezchybná funkcionalita systému a podpora procesů Objednatele dle požadavků zadávací dokumentace a platné legislativy.

Dodávka systému „Software pro monitoring a analýzu zpráv v oblasti RO a JB s modulem pro systém včasného varování“ musí zahrnovat zejména veškeré:

a) **Dodávku všech propojení a interface na okolní informační systémy**. Systém musí být napojen na další informační systémy a služby (nezahrnuje garanci dostupnosti okolních informačních systémů)
   a. Systém je provozován v prostředí CESNET
   b. Systém je propojen na databázi původních příspěvků třetí strany. Aktualizace je dynamická minimálně jednou denně.

b) **Veškeré práce spojené s realizací dodávky „Software pro monitoring a analýzu zpráv v oblasti RO a JB s modulem pro systém včasného varování“**

   - Návrh architektury, tj. vyprojektování aplikační, systémové a komunikační infrastruktury
   - Integrace s interními aplikacemi a externími informačními systémy
   - Zajištění ověřovacího provozu včetně odpovídající odborné technické podpory
   - Nasazení systému do ostrého provozu
   - Řízení změn (change requests)
   - Instalace dodávaných softwarových řešení, vč. **realizace testovacího prostředí** pro testování a uvolňování změn v aplikačním software (patche, updaty, upgrady, verze) a provozním prostředí do užití,
   - Všechny typy **školení** potřebné pro práci s dodávaným systémem.

c) **Kompletní dokumentaci** v elektronické podobě.
d) **Zajištění provozní podpory** a rozvoje Systému po dobu trvání projektu.
e) **Předání provozu systému Objednateli** po ukončení projektu, resp. pokračující provozní podpora v případě, že se Zhotovitel a Objednatel dohodnou na podmínkách pokračování provozní podpory.

## 3.2 Umístění infrastruktury řešení

- Správa systému bude prováděna Objednatelem po ukončení projektu a předání Zhotovitelem, který ale musí zohlednit přístupová práva a standardní bezpečnostní pravidla pro práci s podkladovou databází a vlastní softwarovou (reportovací aplikací)[6].
- Databáze příspěvků bude umístěna na CESNET a stávající stav k začátku projektu zajistí Objednatel.
- Metodika a model budou předány Objednateli v elektronické podobě na cloud CESNET nebo jiné úložiště podle specifikace Objednatele.
- Veškeré kódy aplikace a systému budou umístěny a spouštěny z cloudového prostředí CESNET.
- Bezpečnost CESNET a backup databáze je na straně Objednatele.
- Backup a archivace kódů, modelů, dat pro modely a dokumentace je na straně Zhotovitele.

## 3.3 Popis stávajícího prostředí Objednatele

**A. Popis prostředí SÚRO pro vytváření internet/intranet aplikací***

2. Na webových serverech SÚRO běží OS Debian GNU/Linux
3. Webový engine je založen na Apache 2
4. Vývojové prostředí zajišťují produkty PHP/MySQL(MariaDB) a PostgreSQL
5. SÚRO nedisponuje vývojovým serverem pro tvorbu a nasazování nových aplikací
6. Aplikace jsou v rámci webových serverů nasazovány ve spojení s AD SÚRO, každá aplikace by měla být na AD napojená a komunikovat s AD pomocí šifrovaného spojení (není povolené SSLv3), AD slouží primárně pro autentizaci uživatelů, není využívána pro aplikační role (oprávnění)
7. Aplikace nejsou uživatelům přístupné přímo, ale jsou schované za proxy serverem Apache na serveru INTRANET, řídící url pro webové aplikace je https://intranet.suro.local/aplikace/nazev_aplikace
8. Intranetový portál SÚRO se řídí centrálním systémem autentizace uživatelů právě přes AD
9. SÚRO má vlastního správce internetových/intranetových aplikací
10. Pro „Enterprise“ aplikace ve třívrstvé architektuře tenký klient/aplikační server/databázový server je k dispozici technologie aplikačního serveru WildFly v. 19 a databázový systém Oracle 18c, preferované webové prohlížeče jsou FF60 a vyšší a EDGE

[6] Objednatel a Zhotovitel se mohou po skončení projektu dohodnout na servisních službách Objednatele pro účely udržování a kalibrace systému.

11. Aplikace v prostředí Oracle musejí splňovat požadavek na svou architekturu tak, aby plně podporovaly replikaci dat na úrovni Oracle, tzv. Oracle Streams Replication v rámci Oracle Database Enterprise Edition, prostředí je plně licencováno
12. Webové aplikace musí splňovat podmínky HTTPS-Only, tedy musí pracovat výhradně nad protokolem https
13. Webové aplikace musejí být vyvíjeny bezpečným způsobem, musí splňovat základní bezpečnostní rámce minimálně podle metodiky OWASP TOP 10 a před nasazením musejí být Zhotovitelem otestovány na základní zranitelnosti podle OWASP Testing guide v.4 (zejména cross-site-scripting, sql injekce, xml injekce clickjacking, únik přes chybová hlášení, a podobně)

* Užívá se také název „webových aplikací“ nebo „aplikace webového typu“

**B. Popis prostředí SÚRO pro vytváření aplikací jiného než webového typu**

1. SÚRO provozuje v desktopovém prostředí pouze operační systém Microsoft Windows 10 Enterprise a Pro
2. Aplikace pro desktopové prostředí musejí být „win32“ kompatibilní
3. Pro serverové části „win32“ aplikací je možné využívat operační systém Windows Server 2012 64-bit a Windows Server 2019 64-bit
4. Aplikace mohu být lokální, určené k instalaci do operačního systému Windows 10 nebo určené k instalaci na server do operačního systému Windows Server 2012/2019 64-bit
5. Aplikace mohou být typu klient/server, kdy je použito dvou vrstvé technologie (tlustý klient/databáze) za tento typ aplikace NENÍ považováno prostředí, kdy aplikace využívá pro svou funkci sdílených prostředků na serveru Windows Server 2012/2019 64-bit (CIFS SHARE)
6. Pro aplikace typu klient/server se požaduje plnohodnotný tlustý klient pro klientské počítače nezávislý na dalším software a pro svou distribuci a aktualizaci včetně bezpečnostních aktualizací je vyžadován samoinstalační balíček pro hromadnou distribuci pomocí AD GPO nebo řešení přírůstové aktualizace ze síťového úložiště, které je součástí řešení aplikace
7. V případě instalace aplikací v prostředí MS Windows je možné využívat databázové prostředí Oracle 18c, jiná databáze není k dispozici,
8. SÚRO nedisponuje vývojovým prostředím pro aplikace „win32“
9. SÚRO nedisponuje prostředím pro webové aplikace na platformě Microsoft, v případě, že je taková aplikace na systémy SÚRO nasazována, musí být součástí servisní smlouva s definovanými úkony systémového správce technologie Microsoft

**C. Přístupy třetích stran do interních systémů SÚRO**

- Do interních systémů SÚRO je možné přistupovat pouze přes VPN spojení
- VPN spojení zajišťuje bezpečnostní řešení Checkpoint Mobile
- K přístupu je možné používat pouze šifrované protokoly jmenovitě definované v servisním či instalačním dokumentu (projektu).
- Přístupové údaje jsou vydávány na konkrétní osobu oprávněnou provádět servisní úkony na základě smluvního ujednání, osoba oprávněná k VPN přístupu musí projít bezpečnostním školením, které zajišťuje IKT SÚRO.
- Osoba oprávněná k VPN přístupu je povinna poskytnout emailový kontakt a číslo mobilního telefonu, na který se zasílají autentizační údaje.
- Osoba oprávněná k VPN přístupu odpovídá za přidělené přístupové údaje, v případě ztráty či kompromitace vydaných přístupových údajů je povinna neprodleně tuto skutečnost oznámit smluvnímu partnerovi SÚRO. Tato oznamovací povinnost platí i v případě ukončení smlouvy o servisních činnostech nebo v případě, že oprávněná osoba z jakéhokoliv důvodu servisní činnost přestala vykonávat.
- K serverům CESNET a databázi SÚRO lze přistupovat na základě přidělení přístupových práv Objednatelem na základě podpisu příslušných dokumentů Objednatele garantujících jeho bezpečnostní pravidla.

## 3.4 Požadavky na školení a dokumentaci

**Plán školení**

Zhotovitel nejméně 30 dnů před plánovaným zahájením školení předloží Objednateli k projednání plán školení vymezující obsah školení, termíny školení, místa a způsob provedení školení.

**Podklady pro školení**

Zhotovitel vytvoří podklady pro školení uživatelů ve formě prezentace školených vlastností a funkcionalit systému. Podklady pro školení budou strukturovány podle dílčích modulů reflektujících skupiny samostatně prováděných činností.

**Dokumentace**

V rámci plnění bude dodána zejména následující dokumentace – instalační dokumentace, uživatelská dokumentace, základní bezpečnostní dokumentace, dokumentace migrace dat, provozní dokumentace a školící dokumentace.

**Správa dokumentace**

Objednatel požaduje řízenou správu veškeré dokumentace k systému, a to zejména instalační, uživatelské, projektové, základní bezpečnostní a provozní. Zhotovitel povede centrální knihovnu těchto dokumentů s uvedením všech verzí a datem posledních změn v úložišti dokumentů poskytnutém Objednatelem.

**Aktualizace dokumentace**

S dodávkou aktualizace Systému je vždy též aktualizována dokumentace systému. Aktualizovaná dokumentace obsahuje nové verze celých dokumentů, nikoliv jen dokumenty zaznamenávající dílčí změnu.

## 3.5 Servis a konzultace v průběhu projektu

V rámci projektu Zhotovitel garantuje následující konzultační a servisní služby:

- Podpora a upgrade systému na aktuální verze systémů třetích stran využívaných v rámci projektu (Monitora, operační systém, databáze, programovací jazyk atp.) a to po celou dobu trvání projektu.
- Akceptování připomínek Objednatele k dodaným částem systému podle této Technické specifikace.
- Akceptování připomínek Objednatele k dodaným technickým a jiným specifikacím Díla vytvořených v rámci projektu. Objednatel se zavazuje připomínky vyjádřit nejpozději do 10 pracovních dnů od dodání příslušných specifikací, pokud není dohodnuto jinak. Nedodání připomínek v rámci 10 pracovních dnů ze strany Objednatele se považuje za akceptaci bez připomínek.
- Konzultační činnost a prezentace podle potřeb Objednatele v rozsahu až 10 MD měsíčně (MD =„manday“=8 pracovních hodin, nevyčerpané hodiny se nepřenášejí do dalšího měsíce). Zhotovitel uvede cenu za MD pro konzultační a prezentační služby nad rámec 10 MD. Do konzultační činnosti a prezentací není zahrnuto školení pracovníků Objednatele podle kapitoly Požadavky na školení a dokumentaci.
- Reakční doba na zahájení řešení případných nedostatků v systému v rámci této Technické specifikace a schválené dokumentace vytvořené v rámci projektu je 24 hodin od vznesení požadavku Objednatelem. Zhotovitel navrhne a udržuje systém pro reportovaní nedostatků systému a monitoring stavu jejich řešení. Doba řešení je limitována 5 pracovními dny pro nedostatky způsobené Zhotovitelem.

# 4 Apendix 1 – klíčová slova a tématické okruhy

**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 potraviny or kosmické or (přírodní and pozadí) or radionuklid or izotop or ((radioaktivní or radioaktivita) and (potraviny 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

# 5 Apendix 2 – Atributy webové služby

Seznam atributů poskytovaných třetí stranou:

- ID příspěvku
- Autor (pokud je dostupný)
- Text
- ID otce ve vlákně (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 Zhotovitele (specifikace)

# 6 Apendix 3

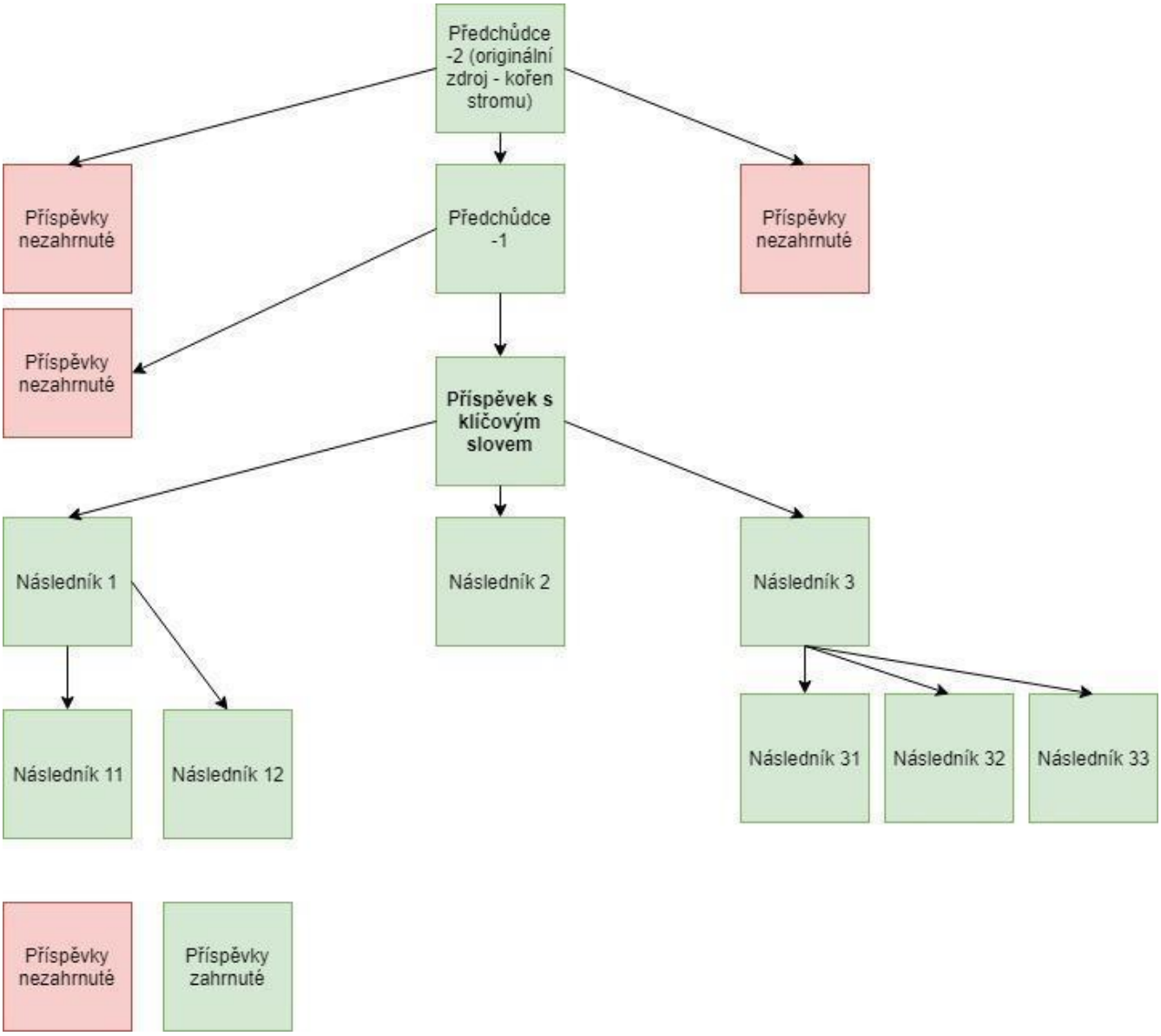
Vyhledávání příspěvků, které by měly být zahrnuty v databázi, je založeno na dvou principech[7]:

1 **Prvotní identifikace – vyhledávání podle klíčových slov**
Prohledává se všechen obsah v monitorovaných zdrojích, do databáze se ukládají pouze příspěvky obsahující specifikovaná klíčová slova / slovní spojení (Konkrétní specifikace klíčových slov je k dispozici v 4.5).

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 náleží do stejného vlákna, pokud mezi nimi existuje logická návaznost[8] (např. vlákno komentářů na Facebooku nebo Twitteru, navzájem navazující komentáře na zpravodajských portálech atd.).

Ukládají se všechny příspěvky navazující na příspěvek s klíčovým slovem a všichni bezprostřední otcové tohoto příspěvku (viz Obrázek 1: Schéma zahrnování příspěvků.)

[7] Specifikace vláken se může vyvíjet podle systému, který sbírá příspěvky (k datu vypsání výběrového řízení jde o systém Monitora)
[8] Přesná definice vlákna záleží na konkrétním zdroji.




*Obrázek 1: Schéma zahrnování příspěvků.*

## 6.1 Frekvence sběru dat

Sběr dat probíhá v různých obdobích po zveřejnění příspěvku tak, aby byla k dispozici informace o časovém průběhu interakce s internetovými uživateli (např. počet likes, počet shlédnutí atd.). Sběr dat probíhá v následujících časových rozestupech[9]:

- Co nejdříve po zveřejnění příspěvku s klíčovým slovem.
- 3 dny po zveřejnění příspěvku s klíčovým slovem (včetně nových následníků).
- 14 dnů po zveřejnění příspěvku s klíčovým slovem (včetně nových následníků).

Časové okno 14 dnů je vztaženo k příspěvku s klíčovým slovem. Pokud tedy jeho následníci již klíčové slovo neobsahují, sbírají se informace o nich pouze do časového limitu původního příspěvku s klíčovým slovem.

[9] - Ideálně mít možnost volit časové rozestupy jako jeden z parametrů.

# 7 Apendix 4 Struktura databáze SÚRO

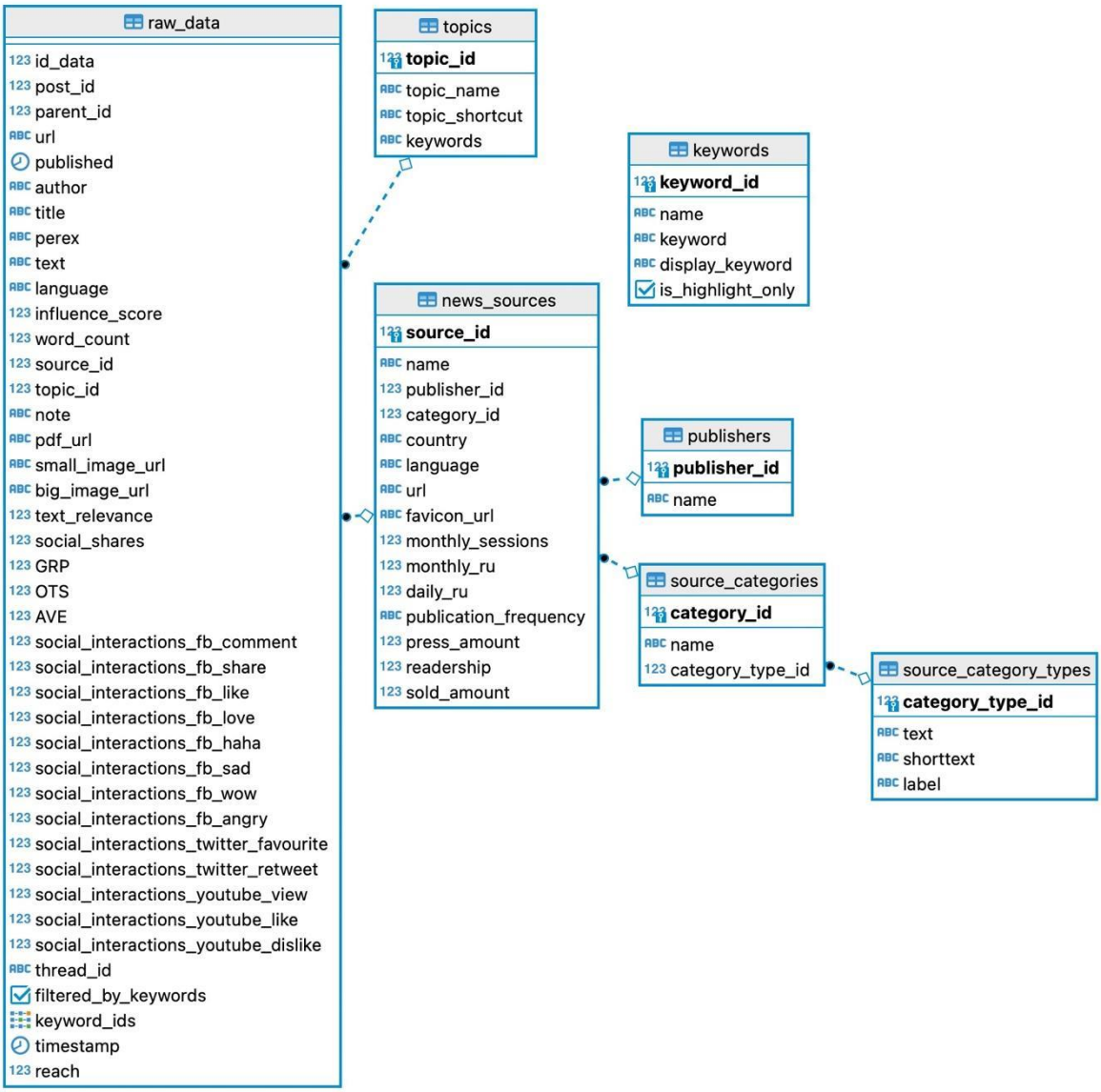
# Struktura databáze

*Platné k datu 21. 6. 2022*

## I. Schéma *production_suro*

### A. Vrstva původních dat

Tabulky sloužící k uložení dat v původní podobě. Struktura dat odpovídá formátu definovaném na straně dodavatele dat.




*Obrázek 1. schéma production_suro, vrstva původních dat*

`raw_data`

Tabulka sloužící k uložení veškerých stažených dat. Společná tabulka pro příspěvky zachycené dle klíčových slov i příspěvky stažené na základě příslušnosti k vláknu.

`topics`

Tabulka se seznamem tematických okruhů definovaných v rámci monitorování příspěvků.

`news_sources`

Seznam informačních zdrojů, jenž jsou monitorovány (webové stránky, sociální sítě, televizní stanice atd.). Každý příspěvek z tabulky `raw_data` přísluší jednoznačně jednomu informačnímu zdroji z tabulky `news_sources`.

`publishers`

Seznam vydavatelů mediálního obsahu. Některé informační zdroje z tabulky `news_sources` mají definovaného vydavatele v tabulce `publishers`.

`source_categories`

Seznam kategorií informačních zdrojů. Každý informační zdroj spadá právě do jedné kategorie.

`source_category_types`

Seznam typů kategorie zdroje. Každé kategorii zdroje odpovídá právě jeden typ kategorie zdroje.

`keywords`

Seznam klíčových slov, které jsou součástí definic jednotlivých témat (zachycených v tabulce `topics`).

## B. Vrstva zpracování dat

Tabulky pro uložení dat po úvodním základním zpracování.

`elementary_table`

Tabulka pro uložení příspěvků identifikovaných na základě klíčových slov. Před zápisem dat se provádí sloučení příspěvků dle témat: je-li jeden příspěvek zachycen vícekrát (dle klíčových slov z více témat najednou), dojde ke sloučení odpovídajících záznamů z tabulky `raw_data` do jednoho řádku v této tabulce.

`keyword_count`

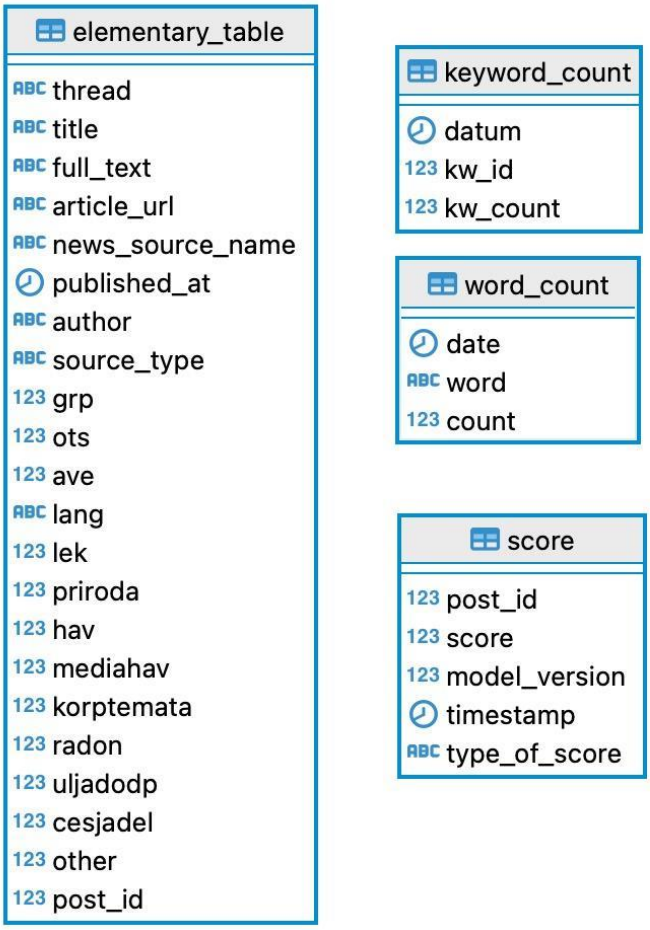
Seznam počtu výskytů klíčových slov v identifikovaných příspěvcích po dnech.

`word_count`

Seznam počtu výskytů jednotlivých slov v identifikovaných příspěvcích po dnech.

`score`

Hodnoty skóre relevance a sentimentu pro jednotlivé příspěvky. Obecná struktura tabulky umožňuje potenciální rozšíření o další druhy skóre i uložení více verzí skóre libovolného typu.




*Obrázek 2. schéma production_suro, vrstva zpracování dat*

## C. Reportovací vrstva

Tabulky sloužící k uložení dat ve struktuře, jež umožňuje prezentování dat v reportovacím systému dle specifikace.


*Obrázek 3. schéma production_suro, reportovací vrstva*

### D. Denní data

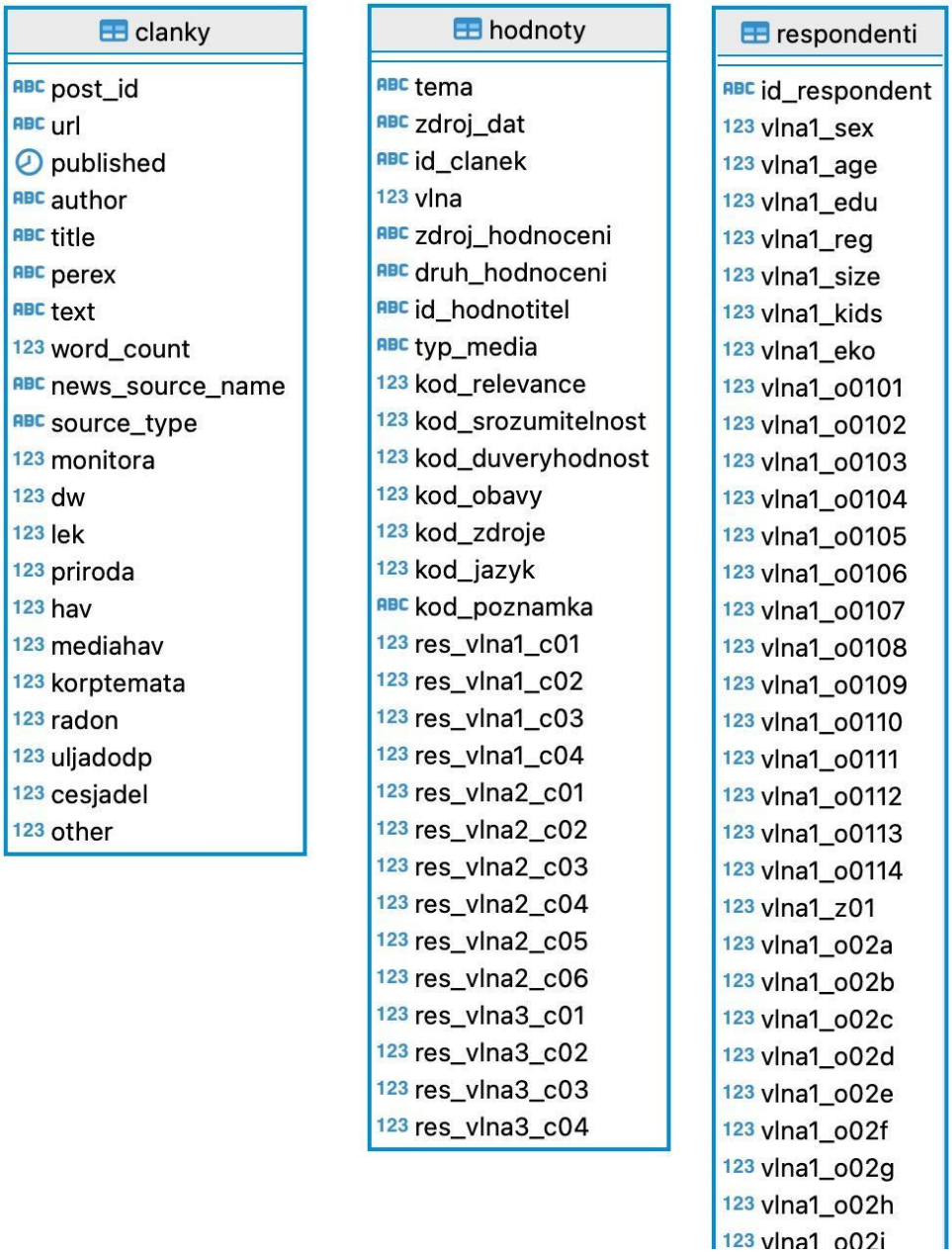
K většině tabulek popsaných v částech A. až C. existují tabulky s identickou strukturou, jež slouží k ukládání dat stažených průběžně během dne. Z těchto tabulek se data jednou denně přepisují do hlavních tabulek, v nichž se ukládá kompletní historie. Jméno denní tabulky odpovídá jménu hlavní tabulky s prefixem `daily_` (např. `raw_data` a `daily_raw_data`).

## II. Schéma *model2022*

Tabulky vytvořené v rámci projektu modelování relevance a sentimentu v první polovině roku 2022.

## A. Články a kódování

Tabulky, které slouží k uložení článků z náhodného výběru pro modelování, a dále k uložení hodnot kódování (kodéry i respondenty) a informačních údajů o respondentech.




*Obrázek 4. schéma model2022, články a kódování*

`clanky`

Seznam článků z náhodného výběru dat pro potřeby modelování. Data mají strukturu do velké míry odpovídající tabulce `elementary_table`: součástí je údaj o příslušnosti článku k jednotlivým tématům, plný text článku, údaj o autorovi článku atd.

`hodnoty`

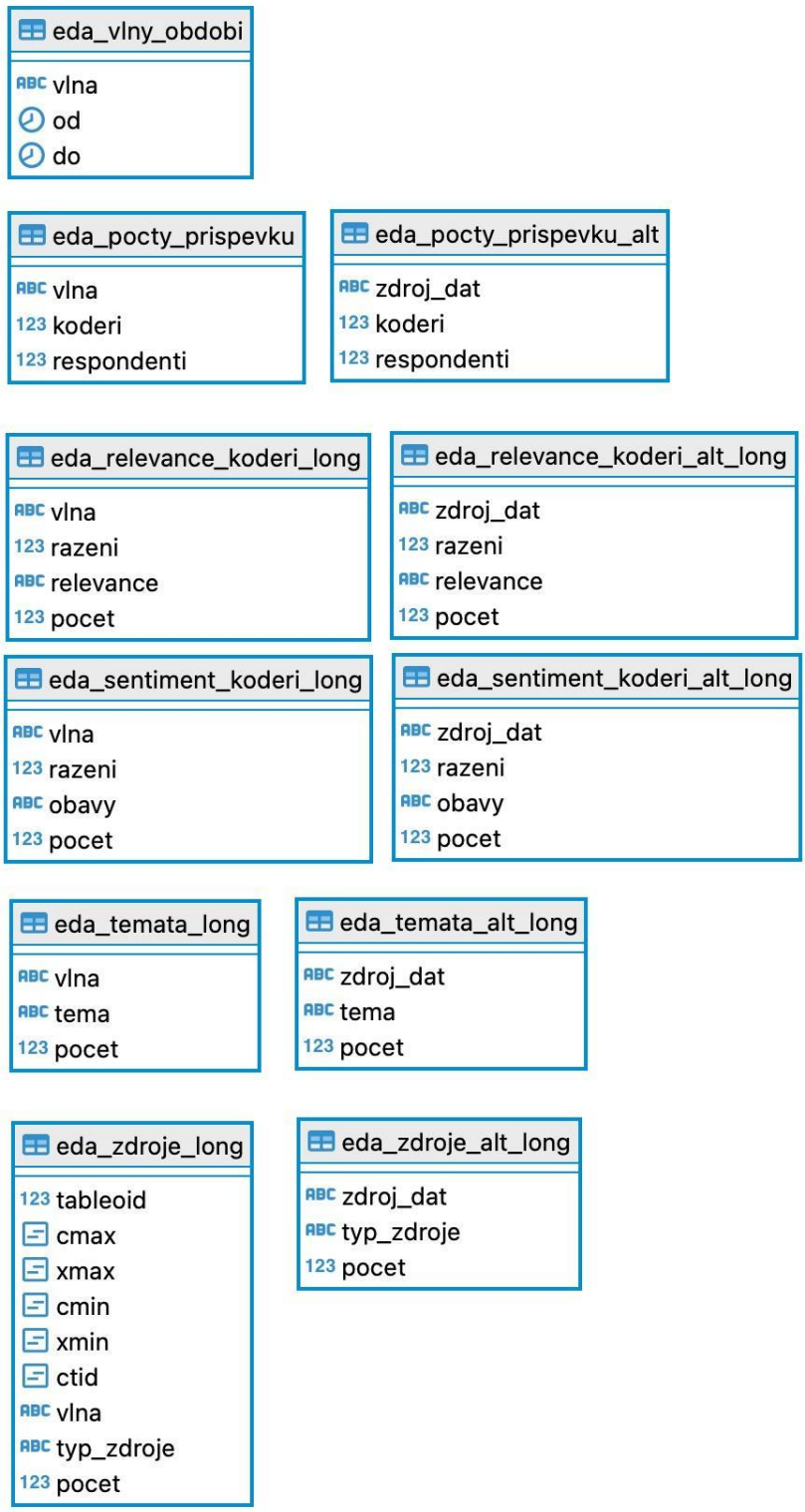
Seznam hodnot z kódování článků. Společná tabulka pro data od kodérů i respondentů.

`respondenti`

Informační údaje o jednotlivých respondentech.

## B. Exploratory data analysis

Tabulky popisující jednotlivé aspekty článků ve výběru (počty příspěvků dle zdroje, dle témat, dle hodnot kódování relevance a sentimentu atp.).



Účastník čestně prohlašuje, že nabízené plnění splňuje požadované technické podmínky uvedené výše.

Pavel Charamza

Digitally signed by Pavel Charamza
Date: 2022.08.05 10:10:59 +02'00'

.....

podpis