VZ vyhledávání a indexace

Český rozhlas hodlá provádět automatizovanou transkripci všech svých pořadů dle dostupných záznamů kontinuálního vysílání od roku 2003. Transkripci následně potřebuje analyzovat a indexovat pro rychlé a přehledné vyhledávání i s návazností na konkrétní čas v rámci audia.

1. Popis situace

Český rozhlas (ČRo) hodlá zpřístupnit některé obsahy vysílání v textové podobě pro snadné hledání i sledování obsahu živého vysílání v textové podobě. Pro tento účel je vyvíjí Analytiku mluveného slova (AMS) a Analytiku vysílání (AV). Obě analytiky pracují souběžně při zpracování živého vysílání, AV identifikuje zvuky, jejich začátky a konce. AMS identifikuje řeč, provádí transkripci, diarizaci a identifikaci mluvčích s tím, že promluvy přiřazuje k hlasovým profilům - identitám anonymním i identifikovaným. Tím obě analytiky vytvářejí archiv vysílání ihned z vysílání dostupné online se zpožděním v řádu nižších minut. Přičemž AMS živého vysílání bude probíhat dvoufázově. V první fázi půjde o rychlost, v druhé, po odvysílání celé epizody ve vysílání, bude provedena AMS hloubková s důrazem na přesnost. AMS archivu bude probíhat průběžně od aktuálního vysílání do minulosti. Zpracování živého vysílání bude přímo navazovat na zpracování archivu.

ČRo hledá dodavatele řešení analýzy, indexace, vyhledávání a výstupů vyhledávání pro zpracování archivu i přepisů živého vysílání.

Předpokládáme, že s dalším vývojem AMS a AV dojde ke zpřesňování obou analytik a bude proto nutné indexovat již zpracované vysílání opakovaně.

Použité zkratky a výrazy:
ČRo - Český rozhlas
AMS - Analytika mluveného slova, zjednodušeně transkripce vysílání
AV - Analytika vysílání, zjednodušeně identifikace zvuků ve vysílání
SERP - Search engine result page, stránka s výsledky hledání
HW - hardware
SW - software
FE - Front end, vizuálně obslužný prostor v prohlížeči uživatele
Podcast - audio vytvořené primárně pro internet, neprošlo vysíláním
Blok úseků - vysílání se dělí na úseky - např. písnička, promluva atd., sloučené do bloku vytvářejí uchopitelný celek (např. reportáž ve zpravodajství je kombinace promluvy moderátora, redaktora, respondenta v reportáži)

1.1 Aktuální stav

1.1.1 Aktuální systém

Pro vyhledávání používá ČRo systém FAST (Microsoft).

1.1.2. Statistiky

Počet hledání měsíčně: max 800 000x
Denní hledání: 6 000 - 40 000 x

- víkendy a svátky 6K-20K denně
- pracovní dny 15K-40K denně

50 nejfrekventovanějších výrazů tvoří 6,35 % všech dotazů.

1.1.3 Výstupy ze systému AV a AMS

Dostupné skrze API ve standardizované podobě.
Je možné ji do určité míry uzpůsobit na straně ČRo dle potřeb dodavatele.

1.2 Předpokládaný stav

Po zavedení automatizované transkripce a identifikace mluvčích bude ČRo zpracovávat 24/7 všechny stanice živého vysílání (ca 24 stanic). Odhadem může tvořit mluvené slovo až 50 % z vysílání denně. Tedy půjde o zpracování max. 300 hodin transkripce vysílání denně.
Součástí budou zpracování budou navíc i on-demand audia, která neprošla vysíláním. Zde může jít o 2-6 hodin denně v horizontu roku 2022.

Na základě analýzy používaných výrazů a na základě nového, uživatelsky přístupnějšího systému se omezí počet výrazů, které se opakují. předpokladem ale je, že se vyhledávání může dotknout hranice 100 000 odeslaných dotazů denně v příštích letech (2021). Nicméně systém není nutné ladit na tuto hodnotu od začátku, kdy maxima tvoří 40K a minima 6K denně.

2. Popis zakázky

2.1 Co je úkolem dodavatele

2.1.1. Dodavatel dodává návrh - architekturu řešení a použité technologie, techniky a postupy s ohledem na další možný vývoj v rozsahu max 10 stran A4.

2.1.2. Dodavatel vyvíjí a dodává systém pro automatizované zpracování výstupů z AMS a AV (text - transkripce, osoby, pořady vysílání…) takové, které umožňuje uživatelům rychlé a snadno parametrizovatelné hledání v obsahu vysílání v rámci projektu mujROZHLAS.

2.1.2.a. Systém zpracovává výstupy z AMS živého vysílání (rychlý, “real-time” přepis) s návazností na timecode.

2.1.2.b. Systém zpracovává výstupy z hloubkové AMS (přesný a důkladnější výstup z AMS dostupný po ukončení epizody odvysílaného pořadu) s návazností na timecode.

2.1.2.c. Systém zpracovává výsledky AMS archivu ČRo, které jsou dodávány postupně od roku 2019 až do roku 2003 s návazností na timecode.

2.1.2.d. Systém zpracovává speciální výsledky z AMS, např podcasty a speciální projekty

2.1.2.e. Systém opětovně zpracovává upravené výsledky z AMS s návazností timecode

2.1.2.f. Systém upravuje výsledky na SERP podle dosavadních hledání, tedy systém by měl umět reagovat na prováděná hledání tak, aby poskytoval vyhledávané / uživateli proklikávané údaje.

2.1.3. Dodavatel vyvíjí a dodává API pro výstupy na SERP či jiných stránkách, na kterých bude probíhat zadávání hledání i jeho výsledky.

2.1.4. Doporučuje a zavádí technologie a techniky, které poskytují “očekávané” výsledky např. pomocí techniky collaborative filtering, similarity search či jiných.

2.1.5. Dodavatel instaluje a spravuje systémy včetně HW pro indexaci a vyhledávání dle bodu 3 níže.

2.1.6. Dodavatel poskytuje servis a podporu, včetně logování informací o zpracování do protokolů pro následnou detekci případných problémů.

2.1.7. Dodavatel dodává a spravuje slovníky, které budou součástí dodávaného řešení, viz 2.4.2.5.

2.1.8. ČRo uvítá ale nevyžaduje, pokud dodavatel již využívá techniky pro vytváření sémantických map a následné vytváření témat. Tedy na základě významových frází provede analytiku, detekci a identifikaci témat, extrahuje jednotlivé entity a na základě jejich sumarizace provede sémantickou analýzu.

2.2 K čemu ČRo hodlá systém využívat

2.2.1. Automatizované zpracování výstupů z AMS a ručních úprav
- Analytika a indexace obsahů ad.
- Štítkování obsahu pro snadnější kategorizaci a dohledání ad.

2.2.2. Vyhledávání
- Parametrizované vyhledávání v celém archivu.
  - Očekávaný počet hledání do 100 000 denně v roce 2021, systém musí být připravený na případné navýšení zátěže (škálovatelnost výkonu).
- Vyhledávání v právě probíhajícím vysílání (AMS živého vysílání)
  - Všechno živé vysílání stanic ČRo bude online analyzováno - prováděna transkripce, diarizace a identifikace - se zpožděním do 10 a 10-30 vteřin.

2.2.3. Zobrazení výsledků hledání
- dle práv uživatelů (práva řeší ČRo) - systém bude velmi využíván i pro interní účely nebo speciální přístupy - “akademické” a pak běžné public - běžní návštěvníci

2.3 Předmět zpracování systémem

Předmětem zpracování jsou transkripce, jména osob, názvy pořadů, seriálů, epizod, stanic atd.

Předměty ke zpracování jsou

2.3.1. Výstupy z AMS archivu kontinuálního vysílání, dodávané průběžně
- Archiv se bude zpracovávat postupně od roku 2019 do roku 2003.
- Jde o strojové přepisy.
- Některé budou ručně upravovány, po úpravě by měl text projít opětovnou analýzou / indexací spuštěnou ručně uložením / uzavřením.
- Přepisy budou dodávány postupně v letech 2019-2023.

2.3.2. Výstupy z AMS živého vysílání
- Výstupy z online AMS
- Výstupy z hloubkové AMS

2.3.3. Výstupy z AMS obsahů vyvíjených pro mujROZHLAS (Podcasty apod.)

2.3.4. Vyhledávání
- Zpracování historie hledání pro nabízení preferovaných hledaných frází
- Zpracovávání výběru uživatelů, pokud uživatelé hledají nějakou fráti a ze SERP volí až třetí položku, měla by se tato položka přesunout na první místo

2.4 Systém - jeho funkce

2.4.1 Analytika

Pod analytikou si představujeme přípravu zpracování a procházení textů a položek u epizod - sloučení a položek k epizodě pořadu / seriálu a jejich váhy (titulek, příp. anotace, autoři, mluvčí…), vyhledání významových frází / klíčových slov atd.
Tedy jde o analytiku:

2.4.1.1. výstupů z AMS archivu, dodávaných postupně do r. 2023,
2.4.1.2. real-time výstupy AMS ze živého vysílání,
2.4.1.3. výstupy hloubkové AMS po ukončení pořadu v živém vysílání,
2.4.1.4. výstupy ze speciálních projektů mimo vysílání (podcast apod.).

2.4.2 Indexace, klíčová slova

2.4.2.1. indexace klíčových slov či významových frází
2.4.2.2. indexace osob
   a. autor promluvy (mluvčí)
   b. autoři / tvůrci obsahu (např. režisér, scénárista, účinkující…)
   c. zmíněné osoby (např. pokud se v promluvě objeví jméno)
2.4.2.3. štítkování obsahu (automatická klíčová slova)
   a. automatické štítkování by mělo probíhat podle předem daných pravidel a mělo by být možné upravit štítkování podle potřeb ČRo
   b. Pravidla
      i. Váha
         1. Slova by měla být vybírána dle váhy mezi obecnými výrazy, osobami, názvy ad., kterou nadefinuje ČRo s dodavatelem.
         2. Např. Jména jsou důležitější než obecné výrazy, tedy spíše uvést Miloš Zeman nežli čučkař.
      ii. Zakázaná klíčová slova
         1. ČRo vyžaduje online přístupný systém pro zavádění výrazů, které se nemají nabídce analytiky objevit.
         2. Např.: ČTK, BBC, Český rozhlas, tiskový mluvčí ad. by se neměly ve výsledku objevit.
      iii. Automatická záměna / sloučení
         1. ČRo vyžaduje online přístupný systém pro zavádění výrazů, které se mohou v textech objevit, ale v nabídce by měly být formulovány jinak.
         2. Např. centrální banka jako Česká národní banka, Spojené státy jako USA
      iv. Nenabízet následné výrazy po určitém slovním spojení
         1. ČRo vyžaduje editovatelný systém (stačí ze strany dodavatele) pro skrývání jmen za konkrétními nadefinovanými výrazy.
         2. Např. z věty tiskový mluvčí Karel Vomáčka, tiskový mluvčí Jana Polívková systém nenabídne žádné jméno po výrazu tiskový mluvčí. ČRo tedy vytvoří seznam výrazů po kterých obvykle následují jména a ty vyloučí z výsledků analýzy.
      v. Vystihnutí obsahu před nabízením kl. slov
         1. ČRo přivítá takovou analytiku, která správně vystihuje téma obsahu.
         2. Např. Pokud osobnost řekne, že “s XY by na pivo nešla”, což se může objevit i v titulku, tak nenabízet kl. slovo pivo, o tom text není.

vi. Označení kl. slov
    1. Na výstupu uživatele ČRo požaduje označovat kl. slova příznakem, která nemá ČRo zavedená ve svém slovníku.
    2. Např. ČRo používá klíčové slovo Donald Trump namísto Donald F. Trump.
vii. Slučování klíčových slov
    1. V textu může být více označení pro to samé, ČRo vyžaduje na výstupu používané kl. slovo v ČRo
    2. Např. Výraz "Národní úřad pro kybernetickou a informační bezpečnost" má zkratku NÚKIB a jeho zápis může být různý, ČRo ale preferuje jen jednu variantu a tu by mělo mít možnost nastavit jako defaultní pro zobrazení ve výsledcích hledání.

2.4.2.4. Opětovná analytika a indexace výběru vysílání při:
- a. novém zpracování transkripce
- b. opravě transkripce
- c. novém klíčovém slovu (třeba czexit)
- d. štítky (klíčová slova)
    - i. téma zahrnuje skupinu klíčových slov , v podstatě jde o kategorii pod níž se slučují klíčová slova
- e. hloubkové analýze AMS

2.4.2.5. Slovníky
- a. zařizuje dodavatel a jsou součástí nabídkové ceny
- b. aktualizace slovníků, tedy zavádění / úprava / vylučování výrazů provádí v průběhu zakázky ČRo a to na na základě doporučení systému nebo zcela ručně z interních podnětů

2.4.3 Hledání a výsledky

2.4.3.1. Hledání
- a. Obecné hledání
- b. Parametrizované hledání
- c. Vyhledávání s možnou filtrací jak pro vyhledávání tak pro výsledky hledání
    - i. obsahy / osoby / štítky viz FE systému
- d. Vyhledávání i v obsazích, které nemají povolenou veřejnou transkripci nebo již není možné je veřejně spustit
- e. Personalizace výsledků

2.4.3.2. Oblasti vyhledávání / parametry
- a. Fulltext
- b. Hledání výrazů
- c. Hledání štítků
- d. Hledání osob o nichž se mluví (štítek)
- e. Hledání osob jež mluví ve vysílání i s jejich promluvami ve vysílání
- f. Hledání pořadů / epizod / bloků vysílání
- g. Hledání stanic, dnů vysílání

2.4.3.3. Výsledky hledání
- a. Slučování výsledků (jedna epizoda)
- b. Řazení výsledků
    - i. ruční nastavení
    - ii. na základě strojového učení
- c. Nastavení mixu výsledků
    - i. ruční nastavení
    - ii. na základě strojového učení

2.4.3.4. Kategorizace indexovaného obsahu dle přístupových práv uživatelů (uživatel výsledek vyhledá, ale nebude pro něj dostupný v plném rozsahu / uživateli se nepřístupné výsledky nezobrazí) - autentizaci a práva řeší ČRo.

2.4.3.5. Práva
   a. Výstup vyhledávání dle práv uživatele (práva řeší ČRo)

2.4.3.6. Rozšiřitelnost
   a. Rozšíření na další projekty ČRo (iROZHLAS, ROZHLAS.cz) - spíše jde o otevřenou možnost rozšířit systém o další obsahy mimo projekt mujROZHLAS.

2.4.3.7. Učení (strojové učení)
   a. Systém musí reflektovat hledané a vybírané výsledky tak, aby zpětnou vazbou měnil příští výsledky na SERP dle chování uživatelů.

2.4.4 API

Dodavatel nadefinuje parametry.

- Rozhraní pro získávání výsledků hledání podle
  - zadaných požadavků a parametrů,
  - práv uživatele.
- REST API

2.4.5 Zpřesňování výsledků

Dodavatel v dokumentaci popíše, jak hodlá dále zpracovávat provedená hledání pomocí machine learningu či jiných metod pro zpřesňování příštích výsledků.

2.4.6 Zachování hledaných výrazů

ČRo bude uchovávat u účtů použité hledané výrazy. Dodavatel a ČRo musí najít způsob správného předání této informace, aby je ČRo mohl zavést k účtu uživatele.

2.5 FE systému

ČRo definuje a vyvíjí FE.
Níže je popis FE pro informaci dodavateli, s jakými parametry na FE ČRo počítá.

2.5.1 Zadání hledaného výrazu

- Input s našeptávačem
  - Zadávání klávesnicí
  - Zadávání hlasem na mobilu (od ledna je možné i v HTML prostředí přepnout klávesnici na hlasové zadávání a telefon převede sám hlas na text)

Výsledkem zadání je vždy textový výraz / fráze s případnými parametry.

2.5.2 Parametrizované hledání

Kategorie

- Pohádky
- Detektivky a krimi
- Povídky
- Četba
- …

Čas

- Bez omezení (od nejnovějších)
- Termín od do
- 24 hodin (dnes a včera od 0:00)

Délka
- nastavit délku
- do 15 minut
- do 30 minut
- do 1 hodiny

Stanice / projekty
- mujROZHLAS
- Radiožurnál
- Plus
- Dvojka
- Vltava
- Wave
- …

Pořady
- Našeptávač
- Podcasty (audia pro internet)

Osoby
- Našeptávač

Štítky / klíčová slova
- Našeptávač

Práva
- Pouze výsledky s možností přehrání, defaultně zaškrtnuté)

2.5.3 Filtrování a řazení výsledků

Filtry
- Osoby
  - Konkrétní osoba / osoby
- Zprávy
- Datum - rozmezí
- Premiéry / poslední reprízy
- Štítky
- Subkategorie
  - Děti
  - Teenageři
- Délka epizody

Řazení
- Podle data vydání
- Podle oblíbenosti
- Personalizované

2.5.4 Výsledky hledání

Na výsledné stránce by měl být mix hledání dle nastavení kombinace ručního nastavení a doporučení na základě strojového učení. Váhy, kterou by měl určit dodavatel ze svých zkušeností s vyhledáváním uživateli.

Obsah výsledků hledání na SERP

- Možnosti
    - Zobrazit počet výsledků
    - Zobrazit / skrýt i výsledky bez možností přehrání
- Objekty (vše co nevede přímo ke konkrétní audiostopě):
    - Pořady
        - do autoplaylistu
        - do mujROZHLAS playlistu
        - hlídací pes
    - Projekty
    - Osoby
    - Stanice
    - Štítky
- Konkrétní úsek audia v epizodě / bloku
    - Obsahující hledaný výraz
    - Odkazující na konkrétní úsek, nebo blok úseků v audiu
        - Kliknutím na takovýto výsledek přenese uživatele do audia na konkrétní časový úsek, řeší ČRo
    - Je možné přehrát daný úsek přímo na SERP
    - Položky
        - Název epizody
        - Název pořadu
        - Datum vysílání
        - Délka epizody
        - Díl epizody
        - Odkaz na všechny díly v případě, že jde o pokračování
        - Část textu s hledaným výrazem
        - Hlídat nové epizody pořadu
        - vložit do autoplaylistu
            - tuto epizody
            - tento pořad
        - Vložit do mujROZHLAS playlistu
            - tuto epizodu
            - tento pořad
        - Hlídací pes
            - tohoto pořadu
- Sloučené úseky v rámci epizody
    - Pokud je hledaný výraz v rámci jedné epizody, na SERP bude vypsána tato jedna epizoda s možností rozbalení výsledků a výběru toho požadovaného s kliknutím do epizody a přehráváním, nebo přehráváním přímo na stránce hledání.

2.6 Budoucnost

Níže uvedené je pouze pro informaci o dalších vývoji v příštích letech. Níže uvedené tedy není v rámci této zakázky vyžadováno, dodavatel to tedy nezahrnuje do cenové nabídky :

1. ČRo bude mít zájem na vytváření automatických shrnutí z epizod a to nikoliv ve formě vybraných vět, ale souvislého textu do ca 500 znaků. V druhém, paralelním kroku, se chce ČRo podílet na vývoji text-to-speech systému, který by umožňoval našim posluchačům / uživatelům shrnutí vybraných pořadů (personalizované / customizované nastavení). S tím souvisí zavedení sémantické analýzy a mapování dle významových frází.

2. Rádi bychom slučovali epizody nebo bloky úseků vybraných pořadů do témat, tedy provádění kroků automatická detekce a identifikace témat, extrakce entit, sumarizace, sémantická analýza. Např. Na téma se vážou obsahy dle významu nikoliv jen použití klíčového slova. Např. téma Okupace 1939 sdružuje pouze obsahy k tomuto tématu a nikoliv všechny epizody / úseky vysílání Téma se váže dle významu téma je všeobecnější, které zahrnuje štítky, např. vesmír, zdraví a umožňuje nabízet obsahy na HP jako sekci)

3. Hardwarové požadavky

Dodavatel může využít VMware Českého rozhlasu v rámci cloudu u O2 specifikovaného níže a dle výchozích požadavků. V takovém případě nacení HW dle přiložené tabulky náklady po dobu trvání smlouvy a cenu připočte k ceně celkové. V případě využití VMware dodavatel do ceny promítne i jeho správu, kterou bude vykonávat - ČRo nebude vykonávat správu na VMware.
Pokud dodavatel nevyužije VMware ČRO, zahrne své náklady na HW a správu po dobu trvání smlouvy do celkové ceny.

- Výchozí stav VMware / HW
- Počet požadavků za sekundu
- Rychlost odezvy
- Objem vstupních dat

3.1 Specifikace VMware

VMware vSphere HA + VMware vCloud Director minimálně verze 5.

3.2 Tabulka cen

Parametr	**Množství**	**Jednotka/čas**	**Jednotková cena (Kč bez DPH)**
vCPU (1x2.6 GHz)	1	ks/měsíc	230,-
RAM	1	GB/měsíc	75,-
velikost systémového disku	1	TB/měsíc	3000,-

4. Use case vyhledávání

Typy nestrukturovaného vyhledávání (uživatel vše zapíše do vyhledávacího políčka bez nastavení omezujících filtrů). Způsoby vyhledávání mohou být téměř nespočetné, nicméně zde uvedené příklady by měly vést k nalezení hledaného výrazu.

1. Hledání mluvčího a tématu v určitém období
    a. **Situace**: O rozšíření Temelína mluvil Miloš Zeman už koncem prvního desetiletí.
    b. Předpokládaný dotaz: Temelín Zeman 2000 - 2010

    c. Předpokládaný výsledek: mix odkazů na přesný časový údaj audia s projevem Zemana hovořícím o Temelínu, nebo zprávy
2. Hledání pořadu na určité stanici
    a. **Situace**: Byl to pořad myslím s Čechem někde na Dvojce
    b. Předpokládaný dotaz: Čech Dvojka
    c. Předpokládaný výsledek: pořady / epizody, ve kterých je Čech autorem, hostem.
3. Hledání epizody na téma
    a. **Situace**: Včera jsme slyšel někde v radiu, jak se bavili o schizofrenii
    b. Předpokládaný dotaz: schizofrenie / schizofrenie včera
    c. Předpokládaný výsledek: epizody v nichž téma (klíčové slovo) bylo schizofrenie, řazené od nejnovějších
4. Hledání pořadu
    a. **Situace 1**: Chci poslední Meteor.
    b. Předpokládaný dotaz: poslední meteor / meteor
    c. Předpokládaný výsledek: odkaz do pořadu a výpis posledních epizod pořadu od nejnovějších
    d. **Situace 2**: Hledám pořad Jak to vidí.
    e. Předpokládaný dotaz: Jak to vidí / Jak to vidíte
    f. Předpokládaný výsledek: Odkaz na pořad ČRo Dvojka "Jak to vidí…" a výpis posledních epizod od nejnovější
5. Hledání osoby / moderátora
    a. **Situace**: Kovařík, ten uvádí nějaký pořad.
    b. Předpokládaný dotaz: Kovařík / pořad
    c. Předpokládaný výsledek: pořady / epizody, ve kterých je Kovařík autorem, hostem.
6. Hledání podle délky a kategorie
    a. **Situace 1**: Potřebuji nějakou četbu na hodinu do letadla
    b. Předpokládaný dotaz: četba hodina / četba 60 minut
    c. **Situace 2**: Pohádku pro malé děti do 30 minut
    d. Předpokládaný dotaz: pohádka 30 minut
    e. Předpokládaný výsledek: epizody požadovaných žánrů v požadované délce.
7. Hledání konkrétního dílu
    a. **Situace**: Ve třetím díle četby Husáka to myslím bylo.
    b. Předpokládaný dotaz: Husák třetí díl
    c. Předpokládaný výsledek: seriál četby Husák, výpis epizod, třetí jako první
8. Hledání v právě vysílaném
    a. **Situace**: Slyšel jsem před pár minutami někde v radiu jak mluvili o spalničkách.
    b. Předpokládaný dotaz: spalničky / spalničky vysílání
    c. Předpokládaný výsledek: aktuálně vysílaný pořad v němž se mluví o spalničkách