**Příloha č. 5 – Technická specifikace – systém fulltextového vyhledávání, evidenční systém a digitální knihovna Kramerius**

**Technický popis stávajícího řešení**

**Vyhledávací systém**

využívající open source platformu Elastic Search, verze 7 <https://www.elastic.co/enterprise-search>

Přílohou je tabulka integrovaných datových zdrojů.

Celkově index vyhledávání ČRo obsahuje přibližně 25 M objektů.

**Datové zdroje vyhledávacího systému ČRo:**

## Intranet

Aktuální velikost dat 120 GB.

Očekávaný přírůstek do 500 tisíc objektů/rok.

|  |  |
| --- | --- |
| **Název zdroje** | **Počet dokumentů v tisících** |
| archivCTK | 4 955 |
| archivKose | 2 384 |
| archivStudia | 7 762 |
| archivStudiaCD | 399 |
| archivStudiaPr | 7 |
| archivTovek | 621 |
| Knihovna | 91 |
| Hudebniny | 151 |
| intranet | 19 |
| intranetSoubory | 7 |
| telefonniSeznam | 2 |

## Extranet

Aktuální velikost dat 140 GB.

Očekávaný přírůstek do 300 tisíc objektů/rok.

|  |  |
| --- | --- |
| **Název zdroje** | **Počet dokumentů v tisících** |
| drupal-nodes | 2 031 |
| drupal-files | 893 |
| drupal-terms | 28 |
|  |  |
| irozhlas-authors | 3 |
| irozhlas-pages | 1 |
| irozhlas-sections | 1 |
| irozhlas-tags | 2 |
| irozhlas | 3 137 |
|  |  |
| mr-episodes | 613 |
| mr-fulltexts | 262 |
| mr-keywords | 8 |
| mr-livestream | 1 |
| mr-persons | 2 |
| mr-serials | 6 |
| mr-shows | 2 |
| mr-stations | 1 |
| mr-topics | 1 |
|  |  |
| transcription-index | 1 073 |
|  |  |
| Hudebniny | 151 |
| Knihovna | 90 |

**Evidenční systém**

Data aplikací Knihovna a Hudebniny jsou uložena v relační databázi Oracle 11g

Současné řešení využívá databázový Framework Relief, verze 4.

Framework je open source dostupný na <https://gitlab.com/inovatika/aplikator>

**Digitální knihovna Kramerius**

Digitální knihovna ČRo využívá open source publikační nástroj Kramerius, včetně výrobního systému ProArc.

Kramerius: <https://github.com/ceskaexpedice/kramerius/wiki>

ProArc: <https://github.com/proarc/proarc/wiki>

**Technická specifikace softwarové a konfigurační úpravy**

Správa ES prostředí pro externí a interní vyhledávací systém, včetně integračních komponent aplikací a zdrojů na období 4 let.

1. **Předmětem je povinnost poskytovatele poskytovat objednateli služby:**

* **pro systém fulltextového vyhledávání** (dále také jen „vyhledávací systém“) nasazený na
* externí systémy objednatele:

Zpravodajství, iRadio, (dále také jen „externí systémy“)

a

* interní systémy objednatele:

Telefonní seznam, Intranet, Knihovna, Hudebniny, Osobnosti, Archiv zpráv (dále také jen „interní systémy“).

* **pro evidenční systém** (dále také jen "evidenční systém") nasazený na:
* agendu Knihovna a Hudebniny;
* **pro digitální knihovnu Kramerius** (dále také jen "digitální knihovna ") nasazený na:
* digitalizované dokumenty (monografie, periodika, výstřižky);
* e born dokumenty;

1. **Služby poskytované poskytovatelem budou zahrnovat následující činnosti:**

Systémovouá podporu a dohled vyhledávacího systému 24x7 (hotline poskytovaný na celé řešení IS, proaktivní monitoring, pravidelné vzdálené kontroly systému). Monitoring bude zahrnovat provádění následujících činností:

* diagnostiku provozu;
* údržbu vazeb vyhledávacího systému na jednotlivé interní a externí systémy;
* indexaci dat;
* služby instalace a zprovoznění upgrade interních a externích systémů a začlenění nových funkcí vyhledávacího systému po odsouhlasení odpovědnou osobou objednatele;
* provádění softwarových a konfiguračních úprav vyhledávacího systému, provádění systémové optimalizace, vše v maximálním rozsahu 5 MD/měsíc;
* ústní a písemné konzultace k novým funkčnostem vyhledávacího systému dle požadavků objednatele;

**Technická specifikace – Rozvoj indexace obsahu a vyhledávání včetně poskytování podpory**

Indexace a vyhledávání pro internetové pilíře ČRo jsou vybudovány na platformě ElasticSearch. Základem celého řešení je Elasticsearch verze 7.5 cluster (ES můj rozhlas) instalovaný na 3 serverech v síti ČRo spojených do jednoho clusteru. Backend ElasticSearch je doplněn o frontend aplikaci hledani.rozhlas.cz, která komunikuje s backendem přes API, které je současně vystaveno pro využití externích aplikací jako web a mobilní aplikace mujRozhlas. Předmětem zakázky je rozvoj backendu, API a frontendu vyhledávání dle potřeb ČRo (vyznačeno zeleně na obrázku níže).

**Zdroje dat pro indexaci**

Zdrojem dat jsou následující API která zprostředkovávají data z Rozhlas.cz a mujRozhlas.cz

SAPI - data z Rozhlas.cz

* <https://rozhlas.cz/sapi>
* Používané endpointy: nodes, files
* nahrazuje standardní OAI modul

iRAPI - data z iROZHLAS.cz

* <https://www.irozhlas.cz/api/v2>
* používané endpointy: articles-indexer, authors-indexer, pages-indexer, terms-indexer

RAPI - data z mujRozhlas

* <https://mujrozhlas.croapi.cz>
* Používané endpointy: topics, persons, keywords, stations, serials, shows, episodes-indexer

Dalším zdrojem dat je přepis mluveného slova (systém AMS), kde získáváme tyto typy obsahu:

* přepis živého vysílání - websocket aplikace (brutt přepis)
  + ws://mrcix01:8080/livestream/ingest/[nazevStanice]
* offline přepisy dat z disku (od 2003 do roku 2020) (přepis archivu)
  + NFS: pr-s-f-fs001.cro.cz/mrams /mnt/mrams01
* online data přes API (od roku 2020) (deep přepis živého vysílání)
  + https://api.mujrozhlas.cz/ams/stt/station/
* online indexace autorských dat přes API (přepis ondemand)
  + https://www.irozhlas.cz/admin/ams/indexer?page=0

2. **Indexace dat**

Indexace dat probíhá v následujících krocích:

* identifikace poslední aktualizace dat
* stažení nových/změněných dokumentů (epizod, pořadů …) pomocí REST API, websocket či přímo ze souborů
* analýza dokumentu, rozšíření o potřebná data (např. detaily pořadů, stanic, lidí)
* normalizace dokumentu
* zápis do ElastiSearch indexů
* detekce smazaných záznamů a jejich odmazání z indexu

Indexery jsou psány v jazyce JAVA vždy pro konkrétní typ dat. Jsou umístěny na jednom serveru v rámci ES clusteru a spouštěny periodicky pomocí cronu.

**Aplikace a výstupy pro vyhledávání**

Na všech ES serverech je nainstalován webový server Tomcat a je spojen přes loadbalancer. Všechny aplikace následně přistupují pouze přes rozhraní loadbalanceru. Všechny aplikace jsou napsány v jazyce JAVA a nasazeny pomocí war souborů.

Seznam aplikace/contextů:

|  |  |
| --- | --- |
| **Context** | **Účel** |
| /extranetAll | stará https://hledani.rozhlas.cz/stare/ |
| /hledani | UI stávajícího https://hledani.rozhlas.cz |
| /hledani-api | API https://hledani.rozhlas.cz/api |
| /hledani-beta | UI testovací verze nového hledání https://betasearch.rozhlas.cz |
| /hledani-beta-api | API https://betasearch.rozhlas.cz/api |
| /iRadio | UI starého audioarchivu https://hledani.rozhlas.cz/iradio/ |
| /irozhlas | API pro mobilní apku irozhlasu |
| /livestream | webcocket pro přijímání textové analýzy z živého vysílání |
| /mujrozhlas | API pro web/mobilní aplikaci mujrozhlas |
| /mujrozhlas\_stage | stage API pro web/mobilní aplikaci mujrozhlas |
| /searchstats | API pro zápis statistik |

**API pro hledání**

API slouží pro vyhledávání a je mezivrstvou mezi UI (mobilní aplikace, webového vyhledávání aj.) a Elasticsearchem. Jednotlivá API se liší zejména definicí indexů, ve kterých se hledá a definicí dotazů, jak se hledá. Komunikuje se s následujícími endpointy:

* /search – vrací seznam výsledků a agregací
* /aggregatesearch – vrací seznam výsledků sloučených dle typu obsahu (epizoda, pořad …) a agregací
* /suggest – našeptávač. Na základě minimálně 2 písmen jsou vráceny názvy odpovídajících objektů

2. **Webové UI rozhraní**

Webové rozhraní je dostupné na adrese<https://hledani.rozhlas.cz>. Jedná se o Angular aplikaci komunikující přes API s vyhledáváním Elasticsearch.

**Ukázky indexovaných dat (RAPI, SAPI)**

SAPI - https://portal.rozhlas.cz/sapi/nodes?dateFrom=2023-01-29T20:00:00+01:00&dateTo=2023-01-29T20:10:00+01:00

RAPI - <https://api.mujrozhlas.cz/episodes-indexer/12c67855-dcc7-4597-a45b-44623094147f>

**Odezva systému**

Získání výsledku dotazu nesmí přesáhnout 3 vteřiny.