Vyhledávání na webových stránkách Českého rozhlasu

Základní informace o hledání

Český rozhlas centralizuje hledání ve svých internetových projektech na jedno místo, které je dostupné na adrese hledani.rozhlas.cz. Nástroj hledani.rozhlas.cz do sebe integruje články publikované na www.rozhlas.cz i www.irozhlas.cz a zároveň umožňuje hledání v publikovaných audio souborech, což jsou převážně sestříhané záznamy vybraných pořadů Českého rozhlasu.

Audio záznamy vybraných pořadů se publikují na webu Českého rozhlasu cca od roku 2006 a jejich nabídka postupně roste. Rozsah a množství publikovaných pořadů vždy záležel na personálních a technických kapacitách, a také samozřejmě na autorských právech a jiných právních omezeních.

Rozsah hledání

V současné chvíli má fulltextové hledání zaindexováno cca **747 000** článků a **497 000** audio záznamů. Index vyhledání sahá cca do roku 2000. **Jednoduše řečeno: nástroj hledani.rozhlas.cz průběžně každý den zaznamenává veškerý publikovaný obsah na internetových stránkách Českého rozhlasu, který je pak kompletně dohledatelný.**

Hledání umožňuje zapínání různých omezení tzv. filtrů. Tento nástroj uživateli zjednoduší zacílení dotazu nebo vybrání požadovaného okruhu. Může si tak zvolit, jestli chce hledat ve veškerém obsahu, audio záznamech nebo zpravodajství a nemusí být znalý, na který web Českého rozhlasu je samotný obsah publikován. Další možnosti jsou nastavení pro konkrétní stanici nebo pořad a lze také nastavit časové omezení. Tato omezení si volí buď návštěvník, nebo jsou v některých případech tzv. předpřipravena, pokud návštěvník např. zadá hledání na zpravodajském webu, má automaticky zaškrtnutou volbu zprávy, kterou si může samozřejmě dle svého rozhodnutí následně vypnout.

Nástroj hledani.rozhlas.cz prohledává vždy textový obsah, což např. u audiozáznamů je jejich titulek, klíčové slovo a přiřazený pořad včetně stanice a data. **Nevyhledává se v samotném audio souboru.**

Podle čeho se vybírají pořady, které jsou „překlápěné“ na web a kdo jejich výběr určuje?

Výběr pořadu „překlápěný“ na web probíhá na základě dohody mezi vedením stanice (šéfredaktorem) a webeditorem Nových médií. Poslechovost pořadů ve webovém audioarchivu průběžně měříme, pokud dlouhodobě klesne na úroveň jednotek nebo desítek poslechů týdně, zvažujeme nahrazení pořadu v audioarchivu jiným, posluchačsky atraktivnějším. To se ale děje spíše výjimečně, většina pořadů je dlouhodobě zavedených a má na webu stabilní počet posluchačů.

Je na webu ve vyhledávači umístěno v audiu celé vysílání každé stanice po hodinách bez jakéhokoli roztřídění nebo jsou tam jen některé stanice?

Vyhledávač dokáže najít pouze záznamy pořadů v Audioarchivu na adrese http://hledani.rozhlas.cz/iradio, kam je vkládají webeditoři ručně. Není v něm kompletní vysílání žádné stanice, takovou kapacitu Nová média nemají. Návštěvníkům webu ale nabízíme službu Celodenní záznamy na adrese www.rozhlas.cz/zaznamy. Zde je dostupný kompletní záznam celého vysílání všech celoplošných stanic, Radia Wave a Rádia Junior. Tyto záznamy se vytvářejí

automaticky po hodinách, lze v nich snadno najít konkrétní pořad, pokud známe den a přibližně i čas vysílání, ale nelze v nich vyhledávat fulltextově.

Z celodenních záznamů jsou vyřazeny pořady, jejichž publikování na webu je omezeno publikačními právy (četby, rozhlasové hry apod.). Tyto pořady publikujeme zpravidla ručně v Audioarchivu.

Publicistika na internetových stránkách ČRo

V okamžiku, kdy se stanice s webeditorem dohodne na konkrétním pořadu, **je do audioarchivu vkládán každý díl pořadu**. To platí i pro Dvacet minut Radiožurnálu, Pro a proti, Jak to vidí atd. Vynechání dílu je vyloučeno, i když se z technických důvodů může publikace výjimečně o několik dní opozdit. Zvolené pořady se vkládají kontinuálně a kompletně, tj. neselektují se jednotlivé díly.

Ukázky výsledků hledání publicistických pořadů

- **Dvacet minut Radiožurnálu je publikováno od 29. 8. 2006 a má 2709 dohledatelných audio záznamů**
- **Více zde:** http://hledani.rozhlas.cz/?offset=0&porad=Dvacet+minut+Radio%C5%BEurn%C3%A1lu&stanice=%C4%8CRo+Radio%C5%BEurn%C3%A1l%3B&zdroj=audia
- **Pro a proti je publikováno od 1. 3. 2013 a má 1129 dohledatelných audio záznamů**
- **Více zde:** http://hledani.rozhlas.cz/?offset=0&stanice=%C4%8CRo+Plus%3B&porad=Pro+a+proti&zdroj=audia

Zpracoval:

Jiří Malina, pověřený řízením Nových médií, 3. 8. 2017

Doplnění materiálu na základě zaslaných dotazů PhDr. P. Šafaříka dne 23. 8. 2017

OTÁZKA č. 1
Je to nastaveno tak, že vyhledávání prostřednictvím klíčového slova v gramatickém jednotném čísle najde i texty, kde se daný výraz vyskytuje v gramatickém množném čísle?

ODPOVĚĎ
Ano, nástroj FAST, který využíváme pro hledání na webu Českého rozhlasu, umožňuje hledání s akceptací množného a jednotného čísla pomocí interního slovníku.

OTÁZKA č. 2
Je možné vyhledávat s pomocí uvozovek přesné spojení, jak je to známo např. při vyhledávání v Googlu?

ODPOVĚĎ
Ano. Přesné zadání textu do hledání pomocí uvozovek nástroj standardně používá.

OTÁZKA č. 3
Jak je to s případnou možností vyhledávat s pomocí logických operátorů (AND, OR, NOT atp. /známo z knihovních katalogů/)?

ODPOVĚĎ
Nástroj pro hledání má defaultně nastavený operátor AND. Syntaxe pro NOT se zapisuje znaménkem mínus před slovem. Např. "praní –peněz" znamená, že se najdou dokumenty o praní, kde se nevyskytuje slovo "peněz". Operátor OR není nástrojem podporovaný.

OTÁZKA č. 4
Zohledňuje daný systém synonyma, tj. bude-li někdo vyhledávat spojení např. „daňový únik", nabídne systém i příspěvky se spojením „daňový podvod"?

ODPOVĚĎ
Naše webové rozhraní synonyma nepodporuje, přičemž samotný nástroj FAST se synonymy umí pracovat, ale musel by se postupně učit dle konkrétních příkladů, a to by znamenalo zvýšené personální nároky na obsluhu ze strany ČRo.

OTÁZKA č. 5
Povede hledání podle klíčového slova „daňový únik" k nalezení jen takových textů, kde se objevuje přesně toto spojení (popř. přesně v plurálu – „daňové úniky"/?/), nebo se vyhledají i texty pojednávající o daném či velmi těsně souvisejícím tématu, i když se v těch textech konkrétně ono spojení (ať v jednotném nebo množném čísle – „daňový únik", „daňové úniky"), ani jeho synonymum (např. „daňový podvod") nevyskytuje?

ODPOVĚĎ
Ano, ale přesnější odpověď už je techničtějšího rázu.
Pro frázi „daňový únik" se interně spouští tento dotaz, lematizace je defaultně zapnuta:
xrank(string("daňový únik" , mode="SIMPLEALL"), title:string("daňový únik" , linguistics=off , mode="SIMPLEALL"), boost=10000)
Znamená to, že slova „daňový" a „únik" jsou expandována na ostatní známé tvary, jako „danova danove danoveho danovej danovejch danovejm danovejma danovejsi danovejsich danovejsiho danovejsim danovejsima danovejsimi danovejsimu danovem danovemu danovi danovou danovych danovyho danovym danovyma danovymi danovymu nejdanovejsi nejdanovejsich nejdanovejsiho nejdanovejsim nejdanovejsima nejdanovejsimi nejdanovejsimu danovy" a „unicich unikach unikama unikem uniku unikum uniky unik", všechny tyto tvary jsou následně využity pro vyhledání fráze.
Klauzule XRANK s vypnutou lematizací navíc zajišťuje, že vyhledané dokumenty, které obsahují přesnou frázi „daňový únik", obdrží navíc 10000 bodů a posunou se ve výsledcích nahoru.

OTÁZKA č. 6
Dotaz specifikovaný na úplně konkrétním příkladu: kdyby si někdo ve všech 3 vyhledávačích, o které jde (internetový archiv ČRo http://hledani.rozhlas.cz/iRadio , vyhledavač na zpravodajském webu ČRo https://www.irozhlas.cz/ a vyhledavač z hlavní webové stránky ČRo http://www.rozhlas.cz/portal/portal), chtěl vyhledat příspěvky (psané texty i audia; tj. z odvysílaného i z materiálů, které se objevily „jen" na webu ČRo) k níže uvedenému tematickému komplexu, podle jakých vyhledávacích výrazů by měl hledat, aby bylo co nejpravděpodobnější, že jeho rešerše celkově (tj. nejen na 1 pokus/1 vyhledávaný výraz, ale s postupným použitím více vyhledávaných výrazů) povede k co nejúplnějšímu nálezu relevantních textů?

Zde je ohlášený příkladný tematický komplex: daňové ráje, organizovaná daňová kriminalita, např. organizované mezinárodní podvody na DPP (tzv. karuselové obchody/karuselové podvody) + politické a justiční snahy proti výše uvedenému (tj. např. diskuse o společném konsolidovaném základu daně z příjmů právnických osob /CCCTB/; opatření proti erozi základu daně a přesouvání zisků /BEPS/).

ODPOVĚĎ

Nástroj hledání na adrese http://hledani.rozhlas.cz/ lze aplikovat na všechny zdroje (iRadio, iRozhlas, Portál) současně, takže není nutné hledat vždy pro každý segment zvlášť. Pokud zadám do hledání řetězec: *organizovaná daňová kriminalita*, aplikuje se hledání primárně s operátorem AND, ale nepoužívají se synonyma. Pokud hledání najde přesnou frázi, zvýší se výsledek obodováním a bude ve výsledcích zobrazen na přednějších pozicích.

S ohledem na zodpovězené otázky a vysvětlení, jak vyhledávání funguje a co umožňuje, bych doporučoval zjednodušit hledané celky. Snažil bych se vybrat základní klíčová slova tématu a ty zadat s mezerou, aby se defaultně aplikoval operátor AND. Také je nutné si uvědomit, že současné rozhraní pro zadání hledání je určené především pro běžné návštěvníky a nemá nastavené náročnější prostředí pro sestavení rozsáhlejší konstrukce dotazu pro detailní analýzu obsahu.

Zároveň pro uživatele hledání na našem webu připravíme krátký info text s návodem, který může pro některé návštěvníky našich stránek zlepšit výsledek a komfort používání nástroje hledani.rozhlas.cz.

Zpracoval:

Jiří Malina, pověřený řízením Nových médií, 25. 8. 2017