**MR06/2017 – Vyhledávání pro zpravodajský web**

**Příloha č. 4 – Technická specifikace**

Indexace obsahu bude probíhat skrze modul “Views OAI-PMH”, který je views pluginem umožňující publikovat data skrze OAI-PMH protokol.

Změna architektury webu v roce 2016, která většinu klíčového obsahu převedla z ostatních entit na entitu “node” umožňuje zjednodušení v tom smyslu, že indexace bude nově realizována výhradně přes přístupový bod na URL: oai/node.

Přístupové body na URL oai/files a oai/taxonomy nebudou dále využívány.

# oai/node

V rámci tohoto přístupového bodu bude indexován pouze obsahu typu článek, který je v novém CMS klíčovým typem obsahu. Skrze referenční pole článku se budou indexovat další entity - konkrétně obrázky, audia, profily osob a pořady.

Vzhledem k tomu, že obsah typu článek se bude lišit u zpravodajského webu a webu rozhlas.cz, určí dodavatel, zda budou články poskytovány v rámci jednoho přístupového bodu (pak ale některé položky v jednom typu článku nebudou) nebo rozdílné body - tedy rozdílné soubory:

Přehled indexovaných polí článku a polí referenčních entit (varianta rozhlas.cz):

* **Datum aktualizace**.
* **Titulek**.
* **Krátký titulek**.
* **Marketingový titulek**.
* **Perex**.
* **Krátký perex**.
* **Body**.
* **Hlavní obrázek**.
  + **URL obrázku**.
  + **Časové omezení**. Datum začátku a datum konce.
  + **Práva**. K downloadu, kde streamu, apod.
* **Audio** (záznam pořadu+audio - krátké, embedované)
  + **URL audia**.
  + **Časové omezení**. Datum začátku a datum konce.
  + **Práva**. K downloadu, kde streamu, apod.
* **Autor** (autor příspěvku z vysílání, autor příspěvku na webu, autor uměleckého díla)
  + **URL profilu**.
  + **Titulek**.
  + **Body**.
  + **Typ profilu**. Lidé z rádia vs. osobnosti. Možné poznat také skrze source doménu.
  + **Hlavní obrázek**.
    - **URL obrázku**.
  + **Body**.
* **Pořad**.
  + **Jméno pořadu**.
  + **ID**.
  + **URL pořadu**.
  + **Body**.
  + Field collection.
    - **Jméno stanice**.
    - **URL stanice**. Doména.

Přehled indexovaných polí článku a polí referenčních entit (varianta zprávy.rozhlas.cz):

* **Datum aktualizace**.
* **Titulek**.
* **Domicil**
* **Štítky**
* **Perex**
* **Body**
* **Hlavní obrázek**.
  + **URL obrázku**.
  + **Časové omezení**. Datum začátku a datum konce.
  + **Práva**. K downloadu, kde streamu, apod.
* **Autor** (autor příspěvku z vysílání, autor příspěvku na webu, autor uměleckého díla)
  + **URL profilu**.
  + **Titulek**.
  + **Body**.
  + **Typ profilu**. Lidé z rádia vs. osobnosti. Možné poznat také skrze source doménu.
  + **Hlavní obrázek**.
    - **URL obrázku**.
  + **Body**.

# Kategorie vyhledávání

V Drupalu jsou nově klíčové souborové entity vždy “zabaleny” do entity článek. Tento stav u historického obsahu zajistila migrace a u nového obsahu bude zajištěno tím, že souborové entity typu audio a video půjde do systému přidat výhradně skrze článek (netýká se obrázků).

Ve výsledcích hledání zůstanou zachovány současné kategorie - vše, články, audia, zprávy a lidé z rádia.

V čase ovšem bude v kategorii “audia” ubývat obsahu, neboť ten se bude v Drupalu nacházet v kategorii “články”. V určité chvíli se předpokládá, že kategorie “audia” zanikne.

Kategorie “lidé z audi” by měla z Drupalu vracet všechny profily typu “Lidé ČRo”, případně profily mající source doménu lide.rozhlas.cz.

# Filtry

### Stanice

Příslušnost článku ke stanici je v Drupalu řešena skrze tzv. “zdrojovou doménu”, např. článek ze stanice Vltavy má kanonickou URL vltava.rozhlas.cz.

### Pořady

Filtr pořady odstraníme.

### Časové omezení

Filtr pořady odstraníme.