inFormo s.r.o.
Drobného 27, 841 01 Bratislava
IČO: 35762756, IČ DPH: SK2020273761

# Webový portál pre Národný katalóg otvorených dát

## Aplikačná príručka

Verzia: 29. septembra 2023

# Úvod

Tento dokument obsahuje stručný opis Webového portálu pre Národný katalóg otvorených dát, časti projektu Otvorené dáta 2.0.

# Popis portálu

Webový portál vznikol ako nadstavba Národného katalógu otvorených dát na účel prezentácie obsahu katalógu a správy katalógu poskytovateľmi dát, ktorí sú a budú na portáli registrovaní. Na tento účel je možné portál rozdeliť na verejnú a neverejnú časť. Verejnú časť tvorí úvodná stránka, vyhľadávanie datasetov, poskytovateľov dát a lokálnych katalógov, zobrazovanie detailov datasetov, distribúcií a lokálnych katalógov, frontend prostredie pre SPARQL endpoint a zobrazovanie kvality metadát.

Neverejnú časť, určenú pre poskytovateľov dát a superadministrátora, tvorí: správa datasetov, distribúcií, lokálnych katalógov, používateľov poskytovateľov dát, profilu, číselníkov a poskytovateľov dát. Na autentifikáciu poskytovateľov dát slúži ÚPVS (jeho IAM modul). V prípade nedostupnosti alebo odstávky ÚPVS nebude k dispozícii možnosť správy obsahu portálu, ale verejná časť portálu bude funkčná aj naďalej.

Informácie o uloženom obsahu portálu sú z princípu verejné a je ich možné stiahnuť zo SPARQL rozhrania jedným dopytom. V špecifických prípadoch je možné obsah (dočasne) označiť ako nepublikovaný, napr. pre potrebu publikovania viacerých dát naraz. Nepredpokladáme však, že to bude primárny prípad použitia a údaje takto vložené na portál by mali byť publikovateľné (bez informácií, ktoré nie je možné publikovať).

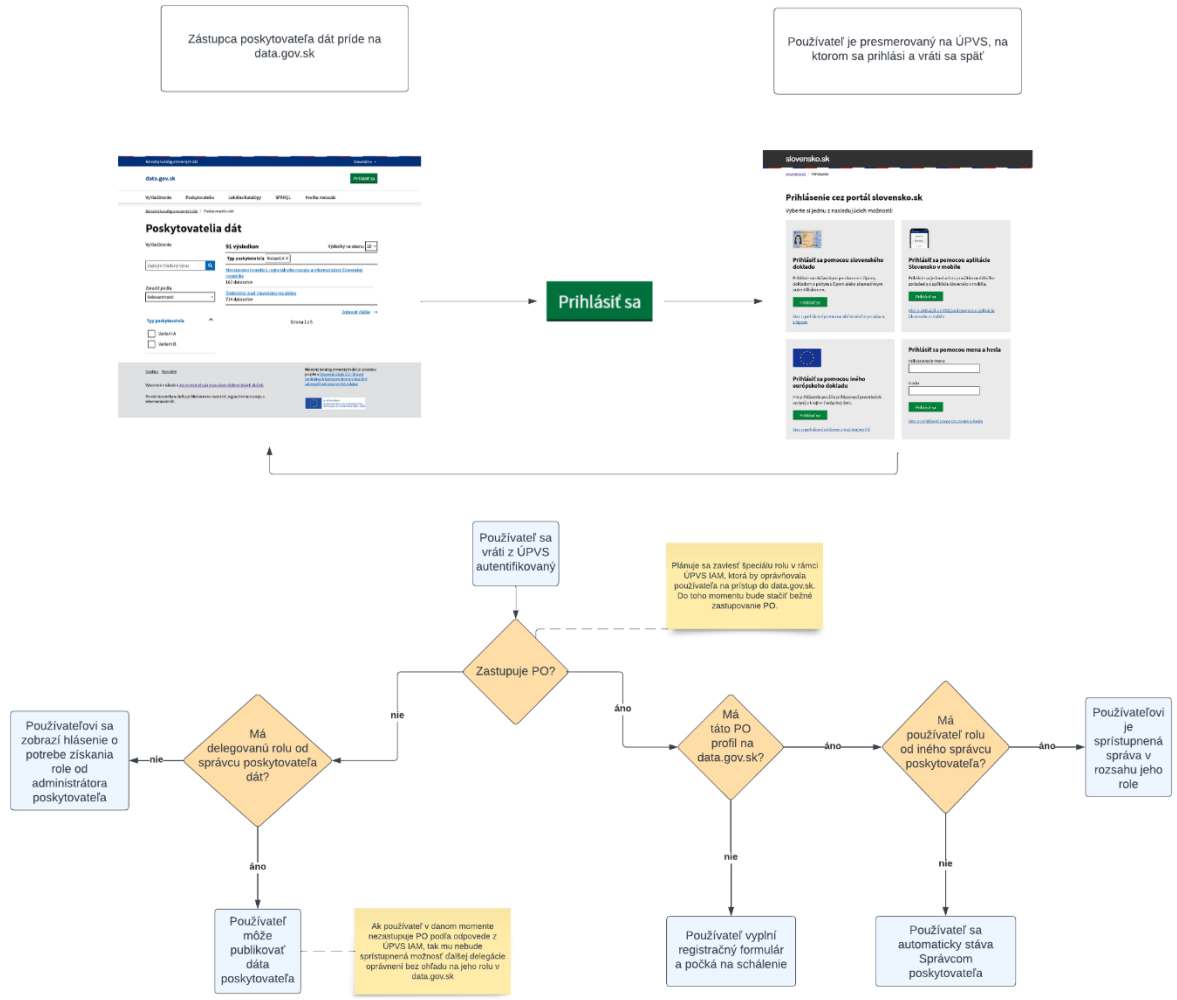
Ako je uvedené vyššie, portál pre autentifikáciu používateľov – poskytovateľov dát používa ÚPVS IAM modul prostredníctvom štandardu SAML 2.0. Z informácií poskytnutých portálom ÚPVS je vytváraný token OAuth 2.0 v podobe JWT tokenov. Obsah tokenu je verejný a je digitálne podpísaný pomocou algoritmu RSA-SHA512. Pri vydaní prístupového tokenu je vydávaný aj refresh token, ktorý slúži na obnovu prístupového tokenu aj po vypršaní jeho platnosti. Platnosť JWT a refresh tokenu je možné konfiguračne ovplyvniť.

Z odpovede z ÚPVS sú vyberané tieto informácie:

- Actor.UPVSIdentityID – jedinečný identifikátor osoby,
- Actor.FirstName – meno osoby,
- Actor.LastName – priezvisko osoby,
- Actor.Email – e-mailová adresa osoby,
- Subject.ICO – IČO zastupovanej organizácie,
- Subject.FormattedName – názov organizácie.

Z identifikačného čísla organizácie je vytvorený URI poskytovateľa dát v podobe https://data.gov.sk/id/legal-subject/ICO, ktorý sa používa na všetky väzby obsahu k poskytovateľovi.

Registráciu poskytovateľa dát musí vykonať osoba, ktorá zastupuje poskytovateľa – organizáciu a automaticky sa stáva administrátorom poskytovateľa. Ten môže delegovať zverejňovanie dát inej fyzickej osobe bez práva na zastupovanie organizácie. Diagram prihlásenia a vyhodnotenia prístupu je možné nájsť na nasledujúcom obrázku.




Obrázok č. 1 Prihlasovanie cez ÚVPS a delegácia oprávnení

Interný modul IAM uchováva priradenie identít k poskytovateľom a ich role v rámci portálu. Rola môže byť jedna z týchto: „Zverejňovateľ dát“ a „Administrátor poskytovateľa dát“.

Dáta entít portálu sú ukladané na samostatnom oddiele v rámci samostatného komponentu portálu. K dátam nie je priamy prístup zvonka, pristupuje sa k nim na základe požiadavky prostredníctvom API rozhrania frontendového API, ktoré následne požiadavku skontroluje a preloží ju na interné volanie API poskytovaného komponentom DocumentStorage. Dáta sú ukladané ako RDF Turtle súbory samostatne pre jednotlivé entity (datasety, distribúcie, katalógy, poskytovatelia). Spolu s dátami sú ukladané aj prípadné súbory obsahujúce zverejňované dáta (napr. CSV súbory). Aby bolo možné jednotlivé súbory rozlíšiť a efektívne spravovať, sú ku každému uloženému súboru k dispozícii metadáta – dáta uložené vo formáte JSON, ktoré obsahujú jeho ID, väzbu na nadradený súbor, priradeného poskytovateľa, číselníkové údaje pre filtráciu a iné. Komponent DocumentStorage pri štarte tento obsah číta a udržiava si index súborov v pamäti, podľa ktorej poskytuje dáta na základe požiadaviek. Komponent DocumentStorage neposkytuje z princípu žiadne chránené dáta bez správneho a zvalidovaného tokenu používateľa. Rovnako neumožňuje žiadnu manipuláciu údajov bez platného tokenu, a to len v rozsahu jeho platnosti (napr. vo väzbe na poskytovateľa dát).

Fulltextové vyhľadávanie s využitím knižnice Lucene poskytuje priamo DocumentStorage, a to s využitím in-memory indexu, ktorý sa vytvára počas štartu aplikácie komponentu. V prípade nárastu

obsahu je možné tento index uchovávať na disku, ale mal by byť vždy vytvorený znova. Tým je zabezpečené, že dáta sú ukladané len na jednom mieste a nie je potrebné riešiť konflikty.

# Číselníky

Pre správne fungovanie portálu je potrebné zadefinovať a udržovať číselníky dát, ktoré sa viažu na vopred daný zoznam vlastností entít. Ich zoznam je možné zistiť z platného štandardu DCAT-AP-SK 2.0 (https://htmlpreview.github.io/?https://github.com/datova-kancelaria/dcat-ap-sk-2.0/blob/main/index.html).

Pre správne fungovanie portálu je potrebné viesť tieto číselníky:

| Id číselníka | Názov / opis |
|---|---|
| http://publications.europa.eu/resource/authority/data-theme | Téma datasetu |
| https://data.gov.sk/set/codelist/dataset-type | Typ datasetu |
| http://publications.europa.eu/resource/authority/frequency | Periodicita aktualizácie datasetu |
| http://publications.europa.eu/resource/authority/place | Miesta súvisiaceho geografického územia |
| http://eurovoc.europa.eu/100141 | Témy EuroVoc datasetu |
| https://data.gov.sk/set/codelist/authors-work-type | Typ autorského diela |
| https://data.gov.sk/set/codelist/original-database-type | Typ originálnej databázy |
| https://data.gov.sk/set/codelist/database-creator-special-rights-type | Typ osobitého práva nadobúdateľa databázy |
| https://data.gov.sk/set/codelist/personal-data-occurence-type | Typ výskytu osobných údajov |
| http://publications.europa.eu/resource/authority/file-type | Formát súboru |
| http://www.iana.org/assignments/media-types | Typ média súboru, kompresného a balíkovacieho formátu |

Pre aktualizáciu číselníka je potrebné vytvoriť RDF Turtle súbor so všetkými položkami tak, aby obsahoval jeden objekt triedy *http://www.w3.org/2004/02/skos/core#ConceptScheme* – minimálne vlastnosť *http://publications.europa.eu/ontology/authority/prefLabel* a samostatné objekty položiek číselníka *http://www.w3.org/2004/02/skos/core#Concept* – minimálne vlastnosti *http://www.w3.org/2004/02/skos/core#prefLabel*, ideálne vo všetkých potrebných jazykoch. Prílohou tohto dokumentu bude stav číselníkov, ktoré vzniknú po (finálnej) migrácii údajov. Súbory číselníkov sú vždy verejne dostupné.

V prípade potreby je možné vytvoriť aplikáciu na prípravu alebo konverziu súboru číselníka s využitím knižníc, ktoré používajú ostatné časti projektu. Požiadavky na ňu budú zistené počas post-implementačnej podpory.

Jednotlivé entity sú vytvárané ako RDF Turtle súbory s využitím týchto pravidiel:

1. Súbor poskytovateľa dát je vytváraný ako neverejný, schválením Superadmina sa stáva verejný.

2. Dataset je zakladaný vždy ako neverejný, ak nie je označený ako séria. Pridaním prvej distribúcie sa môže stať verejným, ak ho zverejňovateľ tak označil.
3. Distribúcia je vždy verejná.
4. Registrácia lokálneho katalógu je verejná podľa toho, či bola tak označená poskytovateľom dát.