OPSI – navodila za zunanji zajem podatkov

14. junij 2018

Uvod

Portal OPSI poleg ročnega uredniškega vnosa zbirk podpira tudi samodejni zajem podatkov iz zunanjih virov. Ta se izvaja redno vsako noč in osvežuje zbirke na portalu, ki pripadajo določenemu viru. Ta vrsta objavljanja je najprimernejša v primeru, ko imate množico zbirk, ki so že dostopne na spletu, a ročno vnašanje na portal OPSI zaradi njihovega števila ni praktično.

Samodejni zajem lahko vzpostavijo le glavni uredniki portala OPSI, zato za uskladitev zajema pišite na naslov odprti-podatki.mju@gov.si, kjer vam bomo pomagali pri vzpostavitvi procesa. Vaš spletni vir mora za samodejni zajem podatke izpostavljati na enotni dostopni točki v enem od podprtih formatov, zato vas prosimo, da nam v sporočilu sporočite, če že podpirate katerega od njih. Tudi v primeru, če nudite strojni vmesnik v katerem nepodprtem, a standardnem formatu, lahko morda portal OPSI razširimo in dodamo podporo zanj.

Če trenutno svojih podatkov nimate izpostavljenih s primernim vmesnikom za strojni zajem, bo verjetno najenostavneje, če vzpostavite popis svojih zbirk v formatu CSV. Oblika, ki se uporablja zanj na portalu OPSI, je spodaj podrobneje opisana in najenostavnejša za vzpostavitev zajema z najmanj dela.

Pred uporabo na javnem portalu OPSI bodo vaši viri najprej preizkušeni v razvojnem in testnem okolju, pri čemer vam bomo lahko sporočili morebitne napake in neskladja v vaših objavah ter pomagali pri njihovi odpravi za zagotovitev nemotenega delovanja v produkciji.

Podprti formati

CKAN

Portal OPSI je osnovan na uveljavljenem ogrodju CKAN[1], ki je trenutno eno od najbolj razširjenih za objavo odprtih podatkov. Uporabljajo ga ZDA, mnoge evropske države in skupni evropski portal https://www.europeandataportal.eu/, ki samodejno zajema tudi podatke s portala OPSI.

CKAN ponuja obširen API vmesnik, zato je prenos podatkov med posameznimi portali tega ogrodja enostaven. Če za vaše podatke uporabljate ogrodje CKAN, dodatnih prilagoditev ne potrebujete.

CSV (splošno)

Format CSV (angl. comma separated value) je eden najenostavnejših strojnih formatov zapisa. Podatki so zapisani kot preprosta besedilna datoteka, v kateri so vnosi urejeni kot tabela. Stolpci so med seboj ločeni z vejicami ali podpičji, prva vrstica pa običajno predstavlja glavo z imeni stolpcev. Na portalu OPSI ga uporabljamo za zajem podatkov Banke Slovenije in Državnega zbora.

Na našem portalu uporabljamo izključno ločevanje s podpičji ter obvezno glavo v prvi vrstici. Če je stolpec prazen, naj bi se podpičje še vedno uporabilo, torej mora biti število podpičij v vseh vrsticah (vključno z glavo) enako. Če želite, da določen vnos (vrednost stolpca) vsebuje podpičje, obdajte to vrednost z dvojnimi narekovaji (""). Primer datoteke lahko najdete na:

[1] https://ckan.org/

https://podatki.gov.si/sites/default/files/2018-05-19_NEW.CSV

Vsaka vrstica datoteke predstavlja eno zbirko na portalu OPSI. Stolpci, katerih preslikava je prikazana na Slika 1, pa so sledeči (kot naj bi bili poimenovani v glavi datoteke):

1) **TITLE**: naslov zbirke, kot naj bi se prikazal na OPSI. Isti naslov se bo uporabil tudi za id zbirke na OPSI (+ predpona npr. "mju"), pri čemer se bodo nedovoljeni znaki samodejno pretvorili v dovoljene (npr. presledki v pomišljaje).
2) **DESCRIPTION**: opis zbirke, kot se prikaže pod naslovom ob ogledu. Zaradi omejitev CSV formata, kjer ne moramo uporabljati prelomov vrstic, za prelom uporabite oznako "\n".
3) **LAST-UPDATED**: datum zadnje osvežitve podatkov, ki jih omenja zbirka ("datum metapodatkov"). Navaja se v formatu YYYYMMDD HH:mm. Ure in minute lahko izpustite. Če datum ne bo podan, bo uporabljen aktualni datum datoteke CSV.
4) **CONTACT_MAIL** in **CONTACT_NAME**: skrbnik, kot se bo navajal pri objavljeni zbirki in bo javno viden. OPSI ima tudi vmesnik za pošiljanje elektronske pošte v zvezi z zbirkami (oranžni gumb "Predlagaj popravek"). Za ciljni naslov teh komentarjev se lahko uporabi skrbnika (torej CONTACT_MAIL), objavitelja (urednik, ki je zbirko objavil oz. sprožil zajem) ali pa fiksni naslov za vse zbirke vašega zajema. Sporočite prosim, kaj vam ustreza.
5) **THEME**: eno od področij, ki jih uporablja portal OPSI. Njihov seznam je viden na glavni strani portala. Navajanje več področij ni podprto. Lahko nam tudi sporočite eno področje, ki bo veljalo za vse zajete zbirke. V tem primeru vam stolpca ni potrebno navajati, ampak bomo dodali vrednost v konfiguracijo zajema.
6) **KEYWORDS**: ključne besede, ločene z vejico. V primeru, da ključna beseda vsebuje vejice, jo obdajte z narekovaji "". Lahko nam sporočite tudi nekaj ključnih besed, ki se bodo programsko dodale vsem vašim zbirkam.
7) **SOP**: številka pravne podlage (če obstaja). Lahko jih je več, ločenih z vejico.
8) **SOP_TITLE**: naziv pravne podlage. Lahko jih je več, ločenih z vejico, pri čemer se jih pričakuje enako število kot pri polju SOP. Nazive z vsebovanimi vejicami obdajte z narekovaji "". Na podlagi posameznega para naziva in številke se bo zgenerirala povezava na www.pisrs.si. Če nazivov ne navedete, bodo povezave poimenovane le s SOP številkami. Če ustrezne SOP številke za naziv ni, bo ta prikazan brez povezave.
9) **REFRESH_INTERVAL**: pogostost osveževanja podatkov, ena od vrednosti: stalno, dnevno, tedensko, dvotedensko, mesečno, četrtletno, polletno, letno, po potrebi, neredno, ni načrtovano. Vrednost ne vpliva na dejanski samodejni zajem portala OPSI.
10) **TEMPORAL_START**: datum začetka veljavnosti oz. relevantnosti podatkov, v enakem formatu kot LAST-UPDATED
11) **TEMPORAL_END**: datum konca veljavnosti oz. relevantnosti podatkov

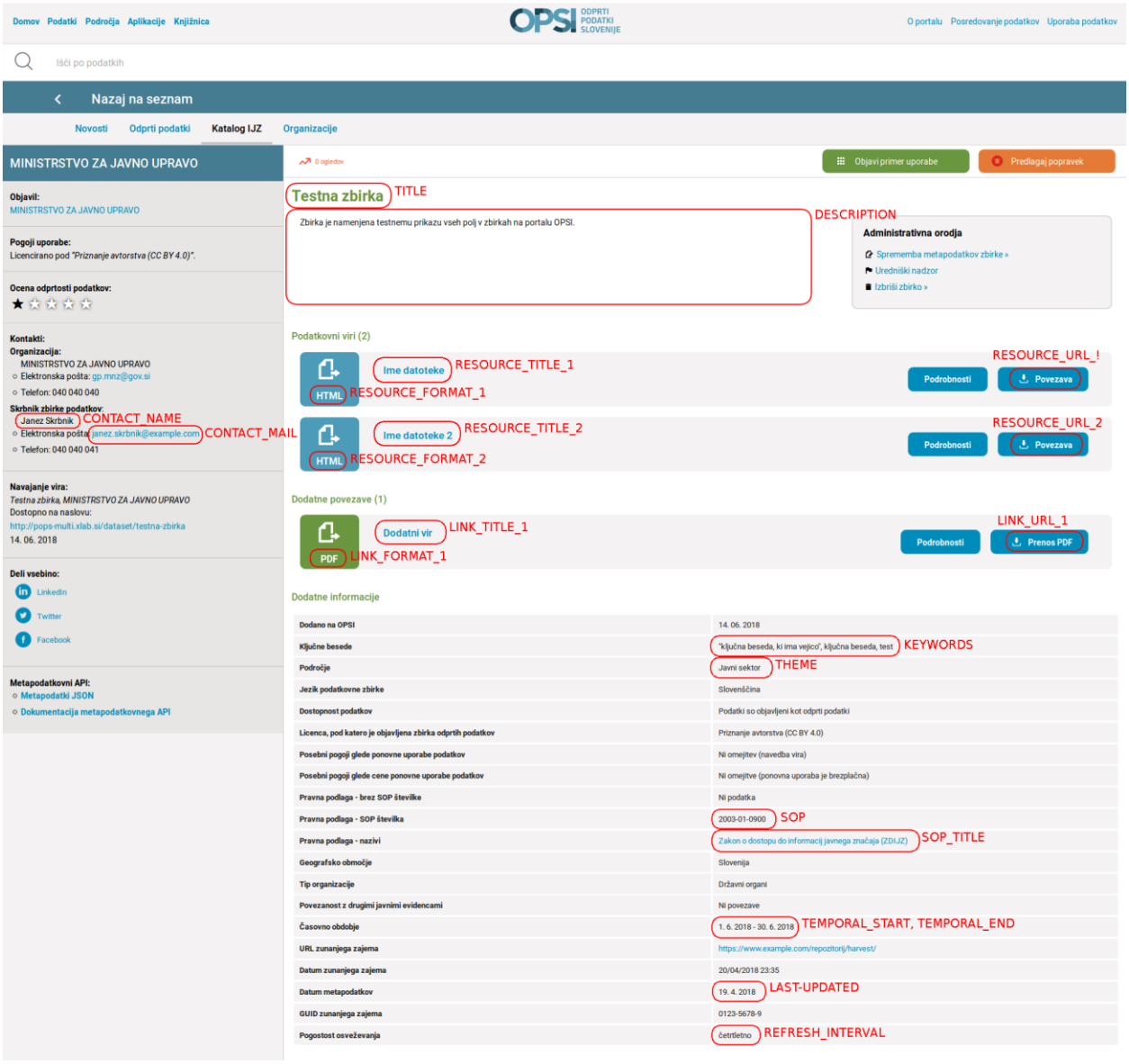
Zbirka naj bi vsebovala tudi enega ali več podatkovnih virov. Za vsakega od njih navedete sledeče stolpce (# nadomestimo z 1,2,...):

12) **RESOURCE_TITLE_#**: ime povezave.
13) **RESOURCE_URL_#**: povezava na datoteko, spletno stran itd.

14) **RESOURCE_FORMAT_#**: format datoteke (HTML če gre za spletno stran)
15) **RESOURCE_DATE_#**: datum datoteke, v enakem formatu kot LAST-UPDATED (opcijsko, a če ima ena datoteka, naj imajo vse, ker je v tem primeru prikaz urejen po datumih).

Poleg podatkovnih virov so lahko v zbirki tudi dodatne povezave (kjer niso podatki, a so relevantne za zbirko, npr. pojasnila):

16) **LINK_TITLE_#**: ime povezave.
17) **LINK_URL_#**: povezava na datoteko, spletno stran itd.
18) **LINK_FORMAT_#**: format datoteke (HTML če gre za spletno stran)



Slika 1 preslikave stolpcev CSV na portal OPSI

Pri zbirkah imamo tudi nekaj polj, ki so fiksno določena:

a) Jezik metapodatkov: slovenščina
b) Jezik zbirke: slovenščina
c) Dostopnost zbirke: 'Podatki so objavljeni kot odprti podatki'
d) Licenca: 'Priznanje avtorstva (CC BY 4.0)'
e) Pogoji licence: 'Ni omejitev uporabe'

V primeru, da ta ne ustrezajo za vaše zbirke, nam sporočite in se lahko dogovorimo o morebitni uvedbi dodatnih stolpcev.

Samodejni zajem poteka tako, da na dogovorjen spletni naslov odlagate datoteke po zgornjem formatu, pri čemer za vsak dan, ko želite spremembe, odložite tri datoteke, označene z datumom, npr. za 19. maj 2018:

- 2018-05-19_NEW.CSV: zbirke, ki jih še ni na portalu OPSI in jih želite objaviti;
- 2018-05-19_UPDATE.CSV: zbirke, ki so že na portalu OPSI, a so bile popravljene;
- 2018-05-19_DELETE.CSV: zbirke, ki jih želite umakniti s portala OPSI.

Portal OPSI vsakodnevno preverja datoteke na dogovorjenem naslovu za "današnji" in "včerajšnji" datum, zato ne spreminjajte starih datotek, ampak pripravite ustrezne nove.

Za konfiguracijo zajema potrebujemo:

1) Naziv vaše organizacije, ki bo navedena kot objavitelj vsake zbirke.

2) Naslov, na katerem bodo dostopne datoteke CSV (NEW, UPDATE, DELETE).

3) Elektronski naslov za pošiljanje sporočil "Predlagaj popravek" (fiksni ali skrbnik zbirke, kot bo naveden pri posamezni zbirki), če želite ta sporočila prejemati vi neposredno namesto urednikov OPSI.

4) Skupno področje in/ali ključne besede, v primeru da bodo iste vrednosti uporabljene pri vseh zbirkah in jih ne želite navajati v stolpcih CSV.

5) Kratko predpono za naslove vaših zbirk v naslovih URL (s tem se izognemo morebitnim podvajanjem z ročno vnešenimi zbirkami). Če tega ne boste navedli, jo bomo določili sami.

Za prvi preizkus lahko objavite tudi le datoteko NEW z nekaj primeri, da preverimo, če ste zajem ustrezno uredili. Po potrditvi delovanja boste nato pripravili datoteko s celotno množico zbirk.

CSV (PCAXIS)

Za zajem Statističnega urada Republike Slovenije smo definirali posebno različico formata CSV, ki je namenjena virom, ki objavljajo datoteke v formatu PCAXIS. Samodejni zajem poteka enako kot pri splošnem formatu CSV (z dnevnimi datotekami NEW, UPDATE, DELETE), razlikuje pa se nabor uporabljenih stolpcev:

1. **PXNAME**: ime PX datoteke, kot je dostopna na strežniku vira (le ime datoteke brez poti), npr. "F2_A2S.PX".Za generiranje povezav do nje bomo potrebovali predpono URL, na kateri so vse datoteke dostopne. Ta naslov nam pošljete in ga bomo dodali v konfiguracijo, npr. "https://apl.bsi.si/pxweb/dialog/Database/slo/serije/"
2. **PXPATH**: morebitna dodatna pot do datoteke, če datoteke na strežniku niso vse v isti mapi, npr. "04_financni_racuni/02_esa2010/". Iz PXNAME, PXPATH in predpone v konfiguraciji se ob zajemu sestavi polna pot do datoteke, npr. "https://apl.bsi.si/pxweb/dialog/Database/slo/serije/04_financni_racuni/02_esa2010/F2_A2S.PX". Vrednost PXPATH naj bi se začela brez poševnice in končala s poševnico. Stolpec PXPATH lahko izpustite, če podmap ne uporabljate.
3. **DESCRIPTION**: vrednost tega polja iz datoteke PX. Za razliko od splošnega CSV se ta uporabi za naslov zbirke na portalu OPSI (in ne njen opis), zato je dobro, da ne vsebuje kakšnih internih oznak (npr. številčnih predpon), ampak le dejansko razumljive informacije. Naslov naj bi bil unikaten.
4. **LAST-UPDATED**: zadnji datum posodobitve, kot je v PX datoteki. Pričakovan format je YYYY-MM-DD.
5. **CONTACT_MAIL**: elektronski naslov skrbnika zbirke, kot bo naveden v zbirki na portalu OPSI. Na portalu bo poleg tega naveden še kontakt organizacije (če so ti podatki v bazi OPSI - prevzeto iz AJPES). Ti naslovi bodo vidni javnosti. OPSI ima vmesnik za pošiljanje elektronske pošte v zvezi z zbirkami (oranžni gumb "Predlagaj popravek"). Za ciljni naslov teh komentarjev se lahko uporabi skrbnika (torej CONTACT_MAIL), objavitelja (urednik, ki je zbirko objavil oz. sprožil zajem) ali pa fiksni naslov za vse zbirke vašega zajema. Sporočite prosim, kaj vam ustreza.
6. **CONTACT_NAME**: ime skrbnika zbirke.
7. **INFO**: metodološka pojasnila k zbirki. Lahko je naveden polni naslov URL ali pa nam posredujete predpono, ki naj bo dodana tem vrednostim za generiranje polne povezave. Pri zbirkah na portalu OPSI bodo te navedene v razdelku "Dodatne povezave".
8. **REFRESH_INTERVAL**: pogostost osveževanja podatkov, ena od vrednosti: stalno, dnevno, tedensko, dvotedensko, mesečno, četrtletno, polletno, letno, po potrebi, neredno, ni načrtovano. Vrednost ne vpliva na dejanski samodejni zajem portala OPSI.

Primer datoteke lahko najdete na:

https://podatki.gov.si/sites/default/files/2018-05-20_NEW.CSV

Za testni zajem potrebujemo:

1. Naslov, na katerem bodo dostopne datoteke CSV (NEW, UPDATE, DELETE).
2. Predpono URL, na kateri bodo dostopne datoteke PX.
3. Predpono za metodološka pojasnila, če ne boste navajali polnih naslovov URL
4. Predpono in pripono za generiranje naslova do izbora podatkov (vaš vmesnik za pregledovanje vsebine PX). V primeru Banke Slovenije sta to npr. "https://apl.bsi.si/pxweb/Dialog/varval.asp?ma=" in

"&ti=&lang=12&path=Database/slo/serije/". Vmes med njiju se pri zajemu doda še ime datoteke PX (brez končnice), na konec pa PXPATH, kar da polni naslov do pregledovalnika, npr.:

1. https://apl.bsi.si/pxweb/Dialog/varval.asp?ma=F2_A2S&ti=&lang=12&path=Database/slo/serije/04_financni_racuni/02_esa2010/

5. Elektronski naslov za pošiljanje sporočil "Predlagaj popravek", če želite ta sporočila prejemati vi neposredno namesto urednikov OPSI ali skrbnikov zbirk.
6. Področje, na katerega naj uvrstimo vaše zbirke, iz šifranta:
   - Energetika
   - Finance in davki
   - Gospodarstvo
   - Izobraževanje, kultura in šport
   - Kmetijstvo, ribištvo, gozdarstvo in prehrana
   - Mednarodne zadeve
   - Okolje in prostor
   - Pravosodje, pravni sistem in javna varnost
   - Prebivalstvo in družba
   - Promet in infrastruktura
   - Sociala in zaposlovanje
   - Javni sektor
   - Zdravje
   - Znanost in tehnologija

   Če vaše zbirke spadajo pod različna področja, lahko dodamo podporo stolpcu THEME, v katerem boste navajali področje po posamezni zbirki. Še vedno mora pa biti iz šifranta in le eno področje na zbirko.

7. Morebitne ključne besede, ki jih dodamo vsem vašim zbirkam. Lahko dodamo tudi podporo stolpcu KEYWORDS za navajanje ključnih besed v datotekah CSV.

SURS je za pripravo datotek CSV vzpostavil samodejni mehanizem, zato vam lahko v primeru, da uporabljate format PCAXIS, pomagamo pri njihovi pripravi.

CSW – INSPIRE

Format CSW[2], ki ga je razvil konzorcij OGC , je odprti standard za objavo prostorskih podatkov v formatu XML. Katalog v tem formatu opisuje zbirke s podatki ali spletnimi storitvami. INSPIRE je evropska direktiva, ki spodbuja objavo prostorskih podatkov v enotni obliki in določa podrobnejšo različico formata CSW. Ta format za hrambo in strojno objavo podatkov uporablja tudi slovenski Geoportal[3], katerega podatke zajemamo na portalu OPSI.

[2] https://en.wikipedia.org/wiki/Catalog_Service_for_the_Web
[3] http://www.geoportal.gov.si/

Zajem tega formata na portalu OPSI je osnovan na zajemu portala data.gov.uk, ki standard še podrobneje opredeljuje s formatom GEMINI. V splošnem sta formata kompatibilna in morebitne napake zaradi odstopanj odkrijemo v testni fazi.

DCAT

Format DCAT je odprt format za opisovanje splošnih podatkovnih virov kot podvrsta formata RDF, ki je v širši rabi na evropskih podatkovnih portalih. Ogrodje CKAN nudi vtičnik za ta format, s katerim lahko zajemamo vire tipa DCAT tudi na portalu OPSI. DCAT sicer nudi izredno širok nabor vrednosti, slovarjev in šifrantov, zato preslikava v zbirke portala OPSI ni povsem enoznačna. Ker na portalu OPSI zaenkrat še nismo zajemali tovrstnih virov, bo v primeru, da boste želeli zajem v tem formatu, predvidoma potrebno uvajalno obdobje, v katerem bomo odpravili morebitne težave in izpopolnili podporo. Podprta polja so podana v Tabela 1 in Tabela 2.

data.json

Format data.json je poenostavljena različica formata DCAT, zapisana v preprostejšem formatu JSON, ob čemer vsebuje podobna polja, navedena v Tabela 1 in Tabela 2. Ta format podpira npr. ogrodje za prostorske podatke ArcGIS. Na spletu so na voljo tudi mnoga orodja, ki so lahko v pomoč za delo s tem formatom:

https://project-open-data.cio.gov/

Polje v data.json	Polje v DCAT	Obvezno	Opis
title	dct:title rdfs:label	da	Naslov zbirke
description	dct:description rdfs:comment	da	Opis zbirke
identifier	URI objekta ali dct:identifier	da	Univerzalni identifikator za zbirko, enak tudi če se spremeni njen naslov
license	dct:license	da	Za zaznavo odprtosti mora biti enaka nazivu ene od podprtih licenc
keyword	dcat:keyword	ne	Ključna beseda, v data.json je lahko seznam, npr. {"okolje", "prostor"}
issued	dct:issued	ne	Datum prve objave
modified	dct:modified	ne	Datum zadnje spremembe
publisher	dct:publisher	da	Organizacija zajema. Proces zajema bo sicer vrednost nadomestil z organizacijo, definirano v konfiguraciji zajema
distribution	dcat:distribution	ne	Podatkovni vir, ki ima poseben format zapisa: glej spodaj
landingPage	dcat:landingPage	ne	Spletna stran vira te zbirke
references	foaf:Document	ne	Vir dodatne dokumentacije o zbirki
language	dct:language	ne	Jezik zbirke
frequency	dct:accrualPeriodicity	ne	Pogostost osveževanja
temporal	dct:temporal	ne	Časovno obdobje veljavnosti

spatial	dct:spatial	ne	Prostorska lokacija, ki jo opredeljujejo podatki
theme	dcat:theme	ne	Eno od področij na portalu OPSI

Tabela 1 Zbirka (dataset) v DCAT in data.json

Polje v data.json	Polje v DCAT	Opis
downloadURL	dcat:downloadURL	Neposredni spletni naslov do datoteke
accessURL	dcat:accessURL	Posredni spletni naslov, če neposredni ne obstaja
title	dct:title	Ime datoteke
description	dct:description	Se trenutno ne uporablja na portalu OPSI
format	dcat:mediaType	Format datoteke
conformsTo	dct:conformsTo	Shema, po kateri so zapisani podatki
temporal	dct:temporal	Časovno obdobje veljavnosti
spatial	dct:spatial	Lokacija, ki jo opredeljujejo podatki
identifier	URI objekta ali dct:identifier	Unikatni identifikator datoteke

Tabela 2 Datoteka (distribution) v DCAT in data.json