Evropská unie
Evropský sociální fond
Operační program Zaměstnanost

MINISTERSTVO VNITRA
ČESKÉ REPUBLIKY

Rešerše vývoje standardů v oblasti katalogizace otevřených dat

Vytvořeno v rámci projektu

Implementace strategií v oblasti otevřených dat II
CZ.03.4.74/0.0/0.0/15_025/0004172

Klíčová aktivita: 02 Rozvoj Národního katalogu otevřených dat
Indikátor: 6 05 00 Počet napsaných a zveřejněných analytických a strategických dokumentů (vč. evaluačních)

Verze výstupu: 02

Verze k 2.8.2019

Popis výstupu

Rešerše bude analyzovat existující standardy používané v EU v oblasti katalogizace otevřených dat (např. CKAN API, DCAT-AP, GeoDCAT-AP). Budou popsány souvislosti mezi standardy a doporučený dopad na Českou republiku, resp. na Národní katalog otevřených dat. Rešerše bude zpracována formou dokumentu a zveřejněna. Její první verze bude vypracována již v prvním roce řešení projektu. Ve druhém a třetím roce budou vypracovány nové verze reflektující aktuální vývoj standardů.

Kvalita výstupu bude posouzena vybraným členem pracovní skupiny. Hodnocena bude především úplnost pokrytí existujících standardů relevantních k tématu otevřených dat v rešerši.

Výstup bude do běžného užívání implementován tak, že se jeho doporučení projeví v návrhu rozšíření Národního katalogu otevřených dat, zejména ve výstupech C3V2 a C3V3. Výstup bude uvažován i v dalších návrzích rozšíření a rozvoje Národního katalogu otevřených dat. Výstup bude v praxi využit též při naplňování opatření č. 5.3 Národní katalog otevřených dat Akčního plánu pro rozvoj digitálního trhu.

Pojmy

- Aplikační profil - zkonkretizování pravidel užití standardu pro zvolený kontext
- Distribuce datové sady - popis souboru ke stažení jakožto fyzické reprezentace datové sady
- EDP - Evropský datový portál
- Harvestace - postup sběru dat z LKOD do NKOD
- LKOD - Lokální katalog otevřených dat
- LOD - propojené otevřená data - 5* stupeň otevřenosti
- NKOD - Národní katalog otevřených dat
- PVS - stávající rejstříkové řešení Portálu veřejné správy
- RDF - Resource Description Framework - datový model pro 5* propojená data
- SEMIC - Semantic Interoperability Community, iniciativa Evropské komise

Relevantní standardy a doporučení

V této sekci je přehled standardů a doporučení relevantních pro rozvoj NKOD, a pro úplnost popis CKAN API, datové struktury pro předávání metadat datových sad aktuálně používané v NKOD. Ze standardů je vybráno jednak doporučení DCAT konzorcia W3C, celosvětově uznané autority pro webové standardy, pro reprezentaci metadat datových sad na webu a pak jeho aplikační profily pro použití v datových portálech v Evropské unii, které připravuje komunita SEMIC (Semantic Interoperability Community), která je iniciativou Evropské komise pro zvýšení sémantické interoperability.

CKAN API

CKAN je software pro katalogizaci dat od Open Knowledge. Pro strojové čtení uložených záznamů o datových sadách poskytuje tzv. CKAN API. Jedná se o API v podobě JSON souborů. Tento software se v minulosti značně rozšířil, a díky tomu bylo jeho API bylo zvoleno pro reprezentaci metadat datových sad v případě lokálních katalogů otevřených dat, které jsou harvestovány do NKOD. Toto API ale není kompatibilní s DCAT a navazujícími aplikačními profily, které jsou nyní doporučovány Evropskou komisí, a na které je tedy třeba v NKOD přejít.

Výhody

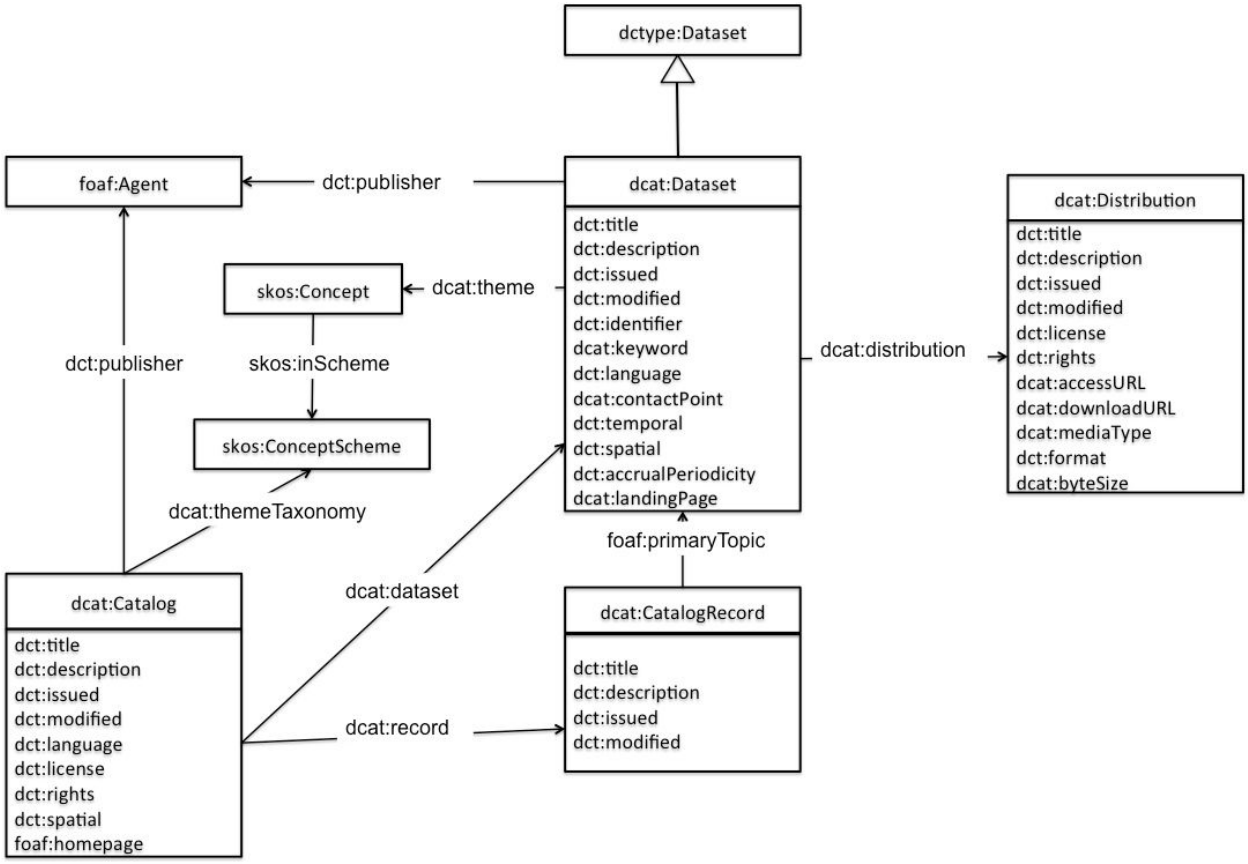
- Jedná se o jednoduchý formát a zaběhnuté technologie

Nevýhody

- Pro sémantickou interoperabilitu nedostačující
- Zvýšené riziko nekonzistencí
- Nekompatibilita s aktuálními doporučeními

Data Catalog Vocabulary (DCAT)

Data Catalog Vocabulary (DCAT) je doporučení konsorcia W3C specifikující, jak publikovat metadata o datových sadách, jejich distribucích, záznamech v katalozích a metadata o katalozích samotných. Platí celosvětově, ale jde o velmi volné doporučení. Specifikuje však základní dělení popisovaných entit na katalog, záznam v katalogu, datovou sadu a distribuci datové sady, kterým se řídí všechny navazující aplikační profily.



Obrázek 1: UML diagram tříd DCAT. Zdroj: Data Catalog Vocabulary (DCAT), https://www.w3.org/TR/vocab-dcat/

Ze standardu DCAT plyne zejména pravidlo říkající, že jednotlivé distribuce jedné datové sady se liší jen formátem souborů ke stažení, tj. CSV, JSON, RDF, nikoli obsahem. Zejména tedy není možné mít datovou sadu, jejíž distribuce jsou například všechny v CSV a jsou rozdělené časovým nebo prostorovým pokrytím. Takové dělení musí být uskutečněno již na úrovni datových sad.

Výhody

- W3C doporučení, stejným způsobem mají být metadata datových sad publikovány celosvětově

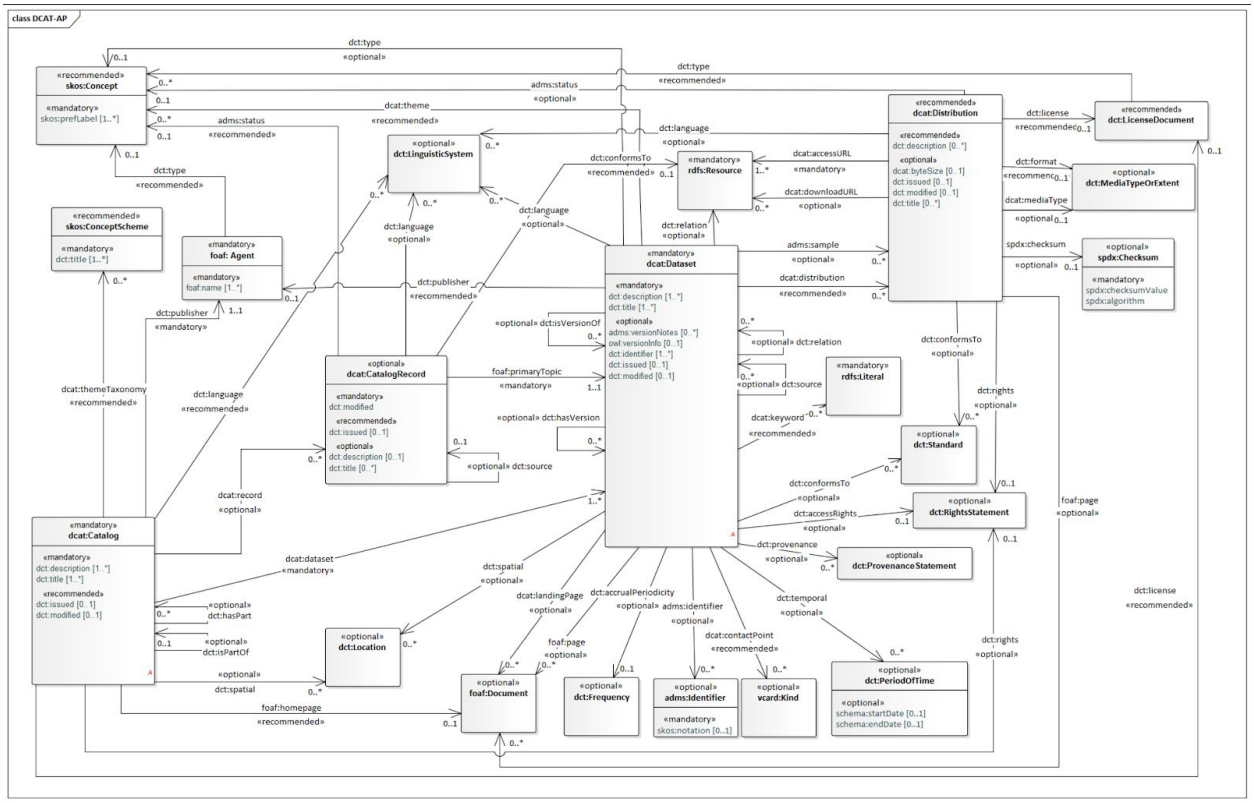
Nevýhody

- Formát RDF se teprve začíná široce rozšiřovat

DCAT-AP v1.2.1

Aplikační profil DCAT pro evropské datové portály ve verzi 1.1 byl vyvinut v roce 2015 společností PwC pro konsorcium DG CONNECT, EU Publications Office a ISA Programme, a je

nadále pravidelně aktualizován. Poslední verzí je DCAT-AP v1.2.1. Specifikuje, jak použít DCAT v evropském kontextu. Poskytuje detailnější pravidla pro používání standardu DCAT, čímž zajišťuje zvýšení interoperability mezi jednotlivými portály, které DCAT-AP implementují. Tam, kde standard DCAT uvádí, kterou RDF vlastností se má co popsat, DCAT-AP například předepisuje i konkrétní číselník, ze kterého mají být hodnoty takovýchto vlastností. Krom toho specifikuje i další vlastnosti a třídy, které by se v kontextu DCAT měly pro popis metadat datových sad použít. Pomocí tohoto aplikačního profilu jsou sbírána metadata z jednotlivých evropských národních katalogů otevřených dat do Evropského datového portálu (EDP), kde jsou dále zpracovávána. Příkladem dalšího zpracování může být strojový překlad metadat do ostatních evropských jazyků.



Obrázek 2 - UML diagram tříd aplikačního profilu DCAT. Zdroj: https://joinup.ec.europa.eu/sites/default/files/distribution/access_url/2019-05/ef3c9392-e38a-40d0-8bf0-b5a847ffae49/DCAT-AP_1.2.1.PNG

Aby mohl být NKOD harvestován do EDP, je třeba vytvořit příslušné rozhraní pro poskytování metadat právě dle DCAT-AP v1.2.1 ve formátu RDF a zajistit používání následujících řízených slovníků:

- IANA Media Types, http://www.iana.org/assignments/media-types/media-types.xhtml pro formáty souborů ke stažení z webu
- Dataset Theme Vocabulary http://publications.europa.eu/resource/authority/data-theme pro kategorie datových sad

- MDR Frequency Named Authority List http://publications.europa.eu/mdr/authority/frequency pro periodicitu aktualizace datové sady
- MDR File Type Named Authority List http://publications.europa.eu/mdr/authority/file-type/ pro formáty datových souborů
- MDR Languages Named Authority List http://publications.europa.eu/mdr/authority/language/ pro jazyky používané v datových sadách
- MDR Corporate bodies Named Authority List http://publications.europa.eu/mdr/authority/corporate-body/ pro evropské instituce a mezinárodní organizace
- MDR Continents Named Authority List http://publications.europa.eu/mdr/authority/continent/ pro kontinenty
- MDR Countries Named Authority List http://publications.europa.eu/mdr/authority/country/ pro země
- MDR Places Named Authority List http://publications.europa.eu/mdr/authority/place/ pro místa
- Geonames http://sws.geonames.org/ pro místa neobsažená ve výše zmíněných slovnících
- ADMS change type vocabulary http://purl.org/adms/changetype/ pro typ změny metadatového záznamu v katalogu
- ADMS status vocabulary http://purl.org/adms/status/ pro stav datové sady
- ADMS publisher type vocabulary http://purl.org/adms/publishertype/ pro typ poskytovatele datové sady
- ADMS licence type vocabulary http://purl.org/adms/licencetype/ pro typ podmínek užití datové sady

Výhody

- Doporučení podporované Evropskou komisí
- Kompatibilní s DCAT
- Jednotné evropské číselníky

Nevýhody

- Formát RDF se teprve začíná široce rozšiřovat

GeoDCAT-AP 1.0.1

GeoDCAT-AP 1.0.1 je profil DCAT-AP v1.2.1 pro geodata. Je primárně určený pro reprezentaci metadat dle Evropské směrnice INSPIRE a standardu ISO 19115:2003, která se používají v

evropských geoportálech, v RDF dle DCAT-AP v1.2.1. Je to proto, aby tato metadata mohla být použita tam, kde je očekávána reprezentace dle DCAT-AP v1.2.1.

GeoDCAT-AP má 2 profily, jeden přímo kompatibilní s DCAT-AP v1.2.1 (Core) a jeden DCAT-AP rozšiřující (Extended). Staví na INSPIRE+DCAT-AP, což byl pokus o INSPIRE v DCAT-AP 1.0. Navíc zahrnuje mapování tezaurů a klasifikací EuroVoc, Gemet, INSPIRE themes a AGROVOC.

GeoDCAT-AP Core je mapování podmnožiny INSPIRE -> DCAT-AP 1.2.1
Rozšiřuje DCAT-AP 1.2.1 zejména o popis geografického pokrytí datové sady o polygony.

GeoDCAT-AP Extended přidává další vlastnosti pro mapování všech INSPIRE metadat do DCAT-AP a dalších slovníků, kde DCAT-AP nestačí. Také používá některé DCAT-AP třídy neobvyklým způsobem, například používá třídu dcat:Catalog pro INSPIRE discovery service.

Pro převod INSPIRE metadat do GeoDCAT-AP existuje XSLT šablona, která má i implementaci pro nasazení na INSPIRE geoportály.

Geodata jsou jednou z hlavních oblastí otevřených dat v ČR, a tak by i tento standard měl být uvažován při rozvoji NKOD, minimálně tedy jeho Core profil.

Výhody

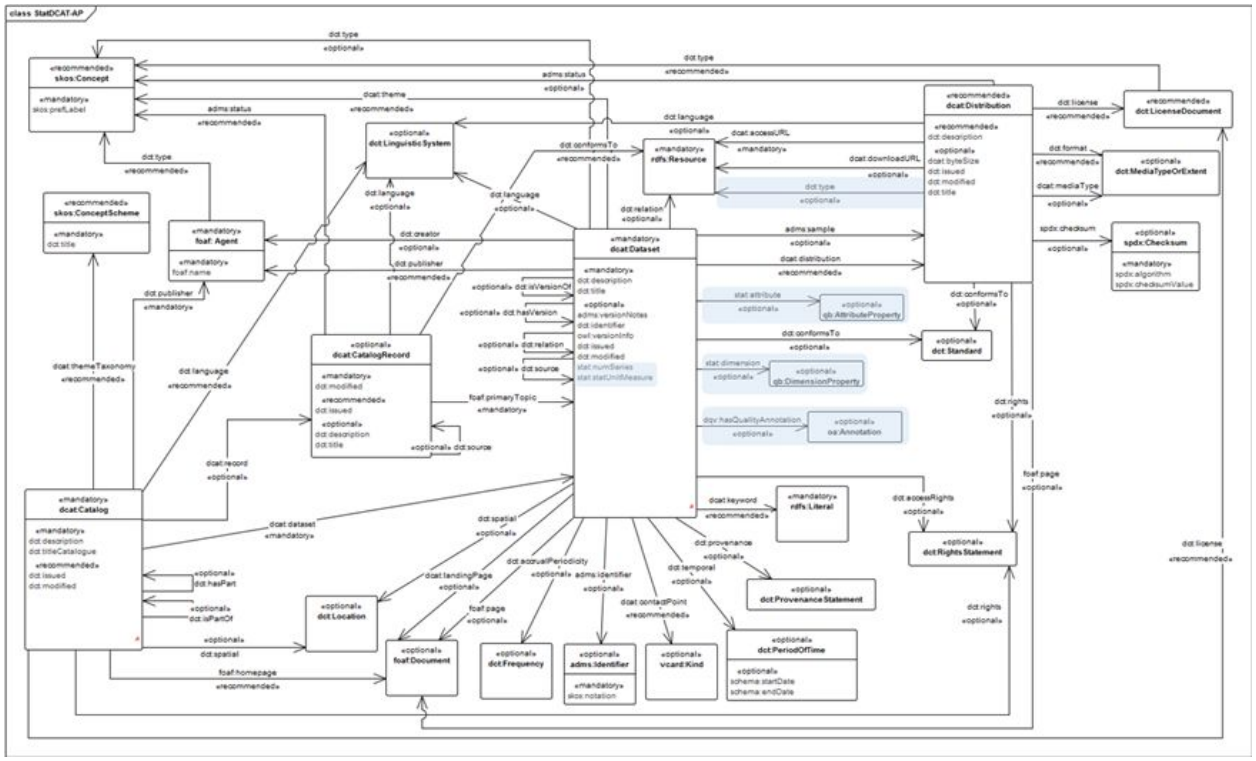
- Doporučení podporované Evropskou komisí
- Kompatibilní s DCAT
- Kompatibilní s DCAT-AP
- Jednotné evropské číselníky
- Umožňuje zaznamenávat geodata ve světě otevřených dat

Nevýhody

- Nutnost sledovat a implementovat změny v další specifikaci

StatDCAT-AP 1.0.0

StatDCAT-AP 1.0.0 je profil DCAT-AP pro statistická data, primárně tedy pro RDF datové kostky dle Data Cube Vocabulary, ale lze takto popisovat i SDMX a podobné datové sady.



Obrázek 2 - UML diagram tříd aplikačního profilu StatDCAT-AP v1.0.0

Rozšiřuje DCAT-AP 1.2.1 o následující atributy:

- Distribuce
  - Typ distribuce - soubor ke stažení, webová služba, vizualizace
- Dataset
  - Počet datových sérií
  - Jednotka měření
  - Atributy (odkaz na qb:AttributeProperty)
  - Dimenze (odkaz na qb:DimensionProperty)
  - Odkaz na anotaci kvality (integrace s DQV - Data quality vocabulary)
- Obsahuje také mapování Eurostat themes na MDR themes

Při zavádění DCAT-AP do NKOD lze zahrnout i tento profil, vzhledem k tomu, že statistická data tvoří nezanedbatelnou část českých otevřených dat.

Výhody

- Doporučení podporované Evropskou komisí
- Kompatibilní s DCAT
- Kompatibilní s DCAT-AP
- Jednotné evropské číselníky
- Umožňuje lépe katalogizovat statistická data

Nevýhody

- Stále není ustálený IRI namespace pro nové vlastnosti
- Nutnost sledovat a implementovat změny v další specifikaci

Slovník VoID

Pro popis a katalogizaci propojených otevřených dat (stupeň otevřenosti 5*) se používá slovník VoID (Vocabulary of Interlinked Datasets) Není to W3C doporučení, jen poznámka zájmové skupiny, nicméně v prostředí LOD se pro popis metadat používá. Příkladem může být portál pro sběr metadat z LOD LODStats, nebo pravidla pro popis datových sad, které se objeví v tzv. LOD cloudu - mapě LOD ve světě. Umožňuje zejména popsat, kde se nachází SPARQL endpoint, jaké slovníky datová sada používá, jaké jsou vzorové entity reprezentující datovou sadu apod. Tento slovník se dá použít dohromady se standardem DCAT-AP tak, že RDF distribuce dle DCAT-AP bude popsána jako VoID datová sada.

Speciální význam má popis slovníkem VoID pro tzv. linksety - datové sady propojující 2 jiné datové sady pomocí vazeb stejného typu.

Výhody

- Umožňuje lépe popsat data na úrovni otevřenosti 4* a 5

Nevýhody

- Nutnost sledovat a implementovat změny v další specifikaci

Budoucí vývoj standardů

V této sekci jsou identifikována místa, která je třeba sledovat kvůli vývoji standardů a dopadům změn v nich na rozvoj NKOD. Uvedené standardy prochází přirozeným vývojem a změny v nich je třeba průběžně promítat do způsobu katalogizace datových sad v NKOD a do způsobu, jakým NKOD poskytuje metadatové záznamy uživatelům.

DCAT-AP v1.2.1 Implementation guidelines

DCAT-AP v1.2.1 obsahuje některé nejasnosti a nepřesnosti, které se upřesňují pomocí tzv. implementation guidelines, na kterých pracuje komunita SEMIC. Při přechodu NKOD na DCAT-AP je třeba k těmto upřesněním přihlédnout.

W3C Dataset Exchange Working Group

V květnu 2017 byla ustanovena pracovní skupina konsorcia W3C pro výměnu datových sad, která si klade za úkol dále pracovat na standardech pro popis metadat datových sad, které zahrnují DCAT, DCAT-AP, ale i další. Práci této pracovní skupiny je nutné při rozvoji NKOD sledovat a aktivně přispívat do vývoje DCAT. Hlavním vylepšením z hlediska NKOD je podpora pro katalogizaci datových služeb, reprezentaci vztahů mezi datovými sadami a podpora pro specifikaci balíčkovacího a kompresního formátu distribuce datové sady. Změny v DCAT budou následně promítnuty i do další verze DCAT-AP. V srpnu 2019 je skupina ve stavu finalizace verze Candidate Recommendation.

Návrh přechodu frontendu NKOD na DCAT-AP v1.2.1

Národní katalog otevřených dat (NKOD) je informační systém veřejné správy provozovaný Ministerstvem vnitra ČR. Nyní je provozován na Portálu veřejné správy (PVS), ale je potřeba řešit jeho rozvoj, především v kontextu aktuálně nejisté budoucnosti PVS a práci na zadání projektu Portál občana.

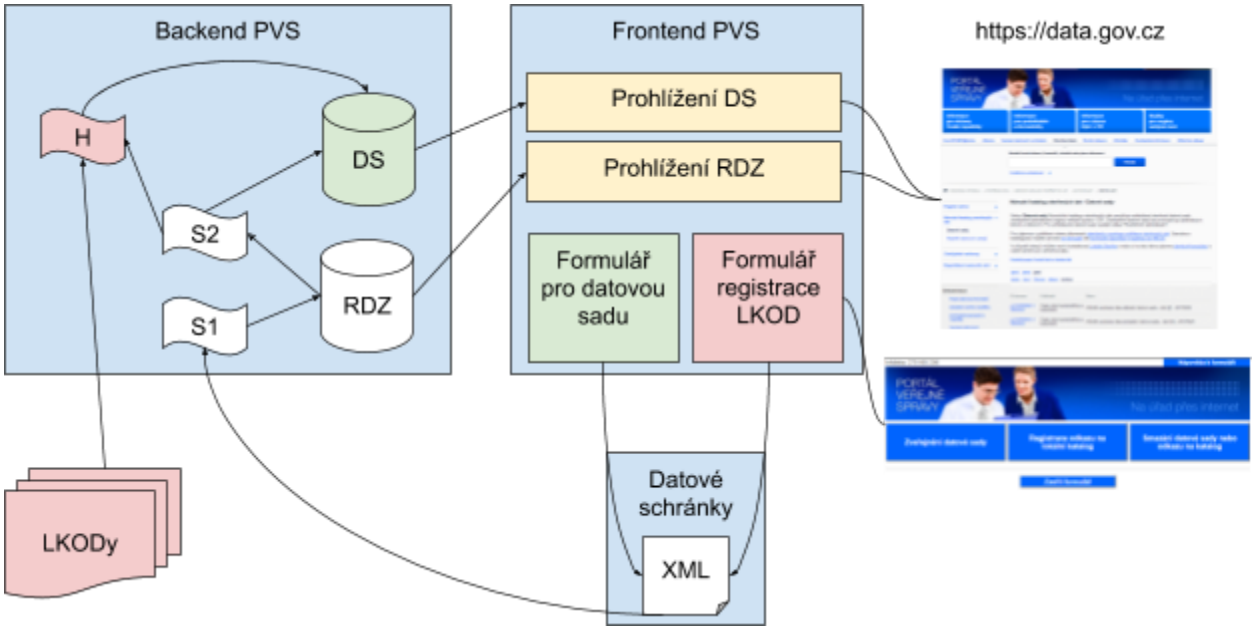
Stávající řešení

V této sekci je popis částí stávajícího řešení na PVS. Pro registraci datových sad do NKOD se aktuálně používá 2 způsobů.

1. Individuální datové sady lze registrovat a spravovat přes formuláře na PVS
2. Přes formuláře na PVS lze zaregistrovat LKOD splňující rozšířené CKAN API, datové sady se pak denně harvestují do NKOD

Rozšířené CKAN API je zjednodušeně řečeno JSON soubor s položkami pokrývajícími již zastaralý standard DCAT-AP 1.0 pro evropské datové portály. Tento je v současnosti již nahrazen standardem DCAT-AP z roku 2015, na který by měl NKOD přejít.

Pro podporu stávajícího řešení jsou k dispozici instalovatelné balíčky do nejrozšířenějších datových katalogů, CKAN a DKAN, rozšiřující jejich API o položky vyžadované NKOD.



Obrázek 3: Stávající řešení NKOD na PVS

Backend na PVS

- RDZ: Rejstřík datových zdrojů poskytující XML API
- S1: Skript vyzvedávající zprávy z datové schránky MVČR a ukládající je do Rejstříku datových zdrojů (každou hodinu)
- DS: rejstřík Datové sady poskytující XML API
- S2: Skript přenášející datové sady z Rejstříku datových zdrojů do rejstříku Datové sady
- H: Skript harvestující LKODy zaregistrované v Rejstříku datových zdrojů a v nich obsažené datové sady. Skript registruje či maže v rejstříku Datové sady.

Frontend na PVS

- Prohlížeč rejstříku datové sady http://portal.gov.cz/portal/obcan/rejstriky/data/97898/
- Prohlížeč Rejstříku datových zdrojů http://portal.gov.cz/portal/ovm/rejstriky/data/97899/
- Formuláře pro vyplnění XML se záznamem datové sady, registraci LKOD a smazání záznamu datové sady nebo LKOD https://portal.gov.cz/webfiller/FormService/Filler.Open?name=nkod.fo

Návrh vylepšení prezentačního rozhraní dle DCAT-AP v1.2.1

Je třeba vyvinut nový alternativní prohlížeč datových sad (v obrázku modře vyznačený), na který bude nově mířit https://data.gov.cz místo prohlížeče na PVS. Bude ale závislý na stávajícím PVS backendu.

Obrázek 4: Alternativní prohlížeč založený na DCAT-AP

Backend na PVS

- Zůstává beze změny

Frontend na PVS

- Formuláře pro sběr dat zachovány ve stávající podobě, zůstávají beze změny.
- Prohlížeče rejstříků zachovány ve stávající podobě, ale https://data.gov.cz už na ně nebude směřovat

Backend prohlížeče

- Využívá XML API pro přístup do rejstříku Datové sady na PVS
- LP-ETL: Nástroj LinkedPipes ETL pro převod metadat z XML API PVS do RDF, dle standardu DCAT-AP v1.2.1
- RDF: RDF databáze (nyní OpenLink Virtuoso) pro uložení RDF dat a zpřístupnění pomocí SPARQL endpointu
- SOLR: Index Apache Solr pro datové sady
- DOC: Dokumentová databáze pro poskytování metadat uživatelům

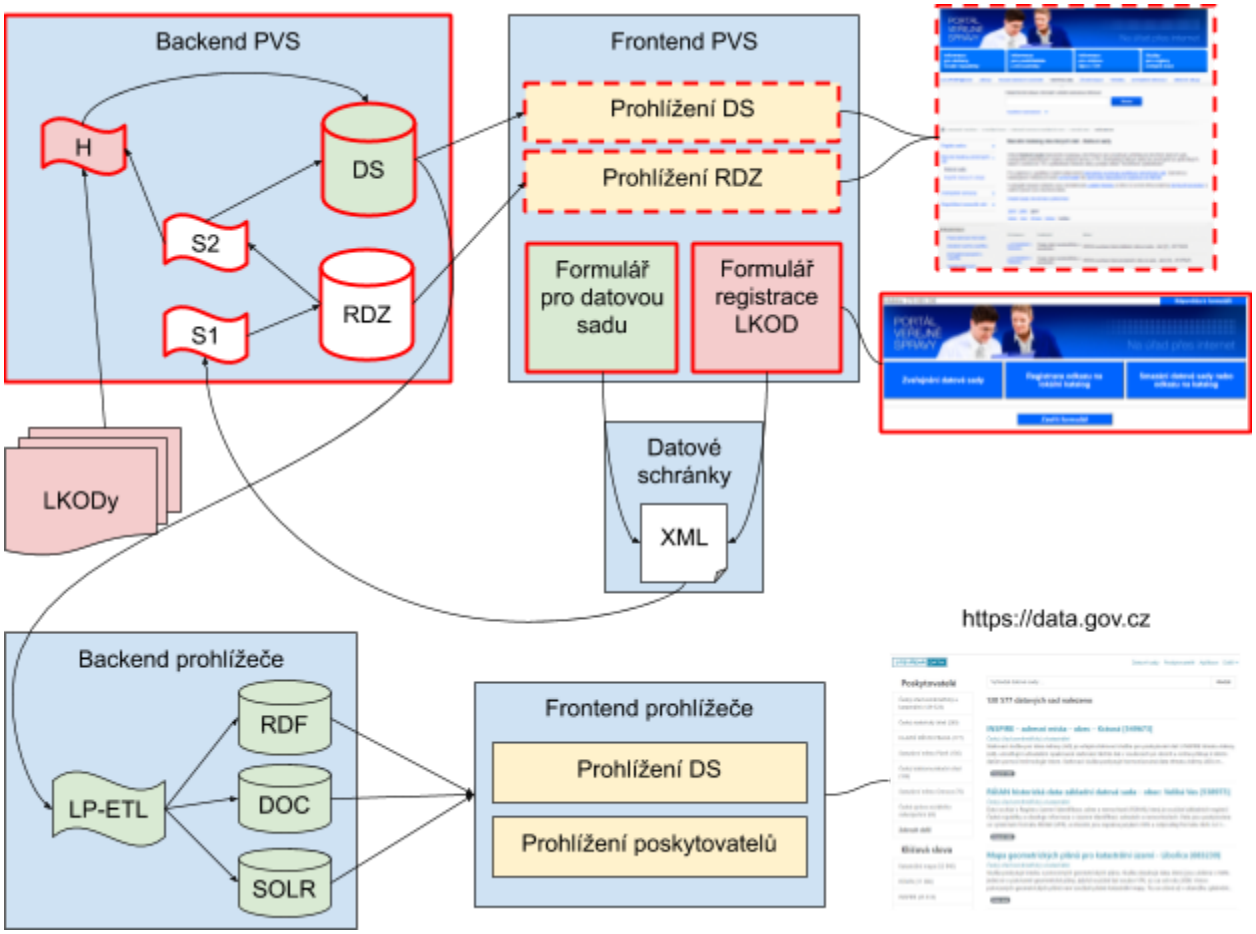
- Node.js pro poskytování vyhledávacích služeb pro frontend

Frontend prohlížeče

- Homepage, na kterou bude směřovat https://data.gov.cz v nové šabloně Portál občana
- HTML + JavaScript aplikace umožňující zobrazení a vyhledání metadat datových sad a poskytovatelů v nové šabloně Portál občana

Popis řešení 2019+

Řešení 2019+ zatím není jasné. Pokud zanikne stávající rejstříkové řešení PVS, následující části (v obrázku červeně vyznačené) bude třeba vyřešit jiným způsobem.



Obrázek 5: Součásti NKOD, které je třeba přenést či vylepšit

Backend na PVS

- *Rejstřík datových zdrojů poskytující XML API*
- *Skript vyzvedávající zprávy z datové schránky MVČR a ukládající je do Rejstříku datových zdrojů (každou hodinu)*
- *rejstřík Datové sady poskytující XML API*

- *Skript přenášející datové sady z Rejstříku datových zdrojů do rejstříku Datové sady*
- *Skript harvestující v Rejstříku datových zdrojů zaregistrované LKODy a v nich obsažené datové sady registrující/mazající v rejstříku Datové sady*

Frontend na PVS

- Prohlížeče rejstříků nejsou potřeba (v obrázku čárkovaně)
- *Formuláře pro vyplňování XML zpráv*

Backend prohlížeče

- Není jasné odkud bude brát data, zbytek může být beze změny

Frontend prohlížeče

- Beze změny

Návrh přechodu implementace NKOD v PVS na DCAT-AP v1.2.1

Současná implementace NKOD na PVS (formuláře, rejstříky a skripty) pracují s již zastaralým standardem DCAT-AP v1.0. Pro přechod na aktuální DCAT-AP v1.2.1 je potřeba je přepracovat a stávající data upgradovat. Zadání této změny je možné vypracovat v rámci projektu OD II. Pracovní název cílového stavu, tj. NKOD pracující s DCAT-AP v1.2.1 je NKOD v1.1.

Požadavky

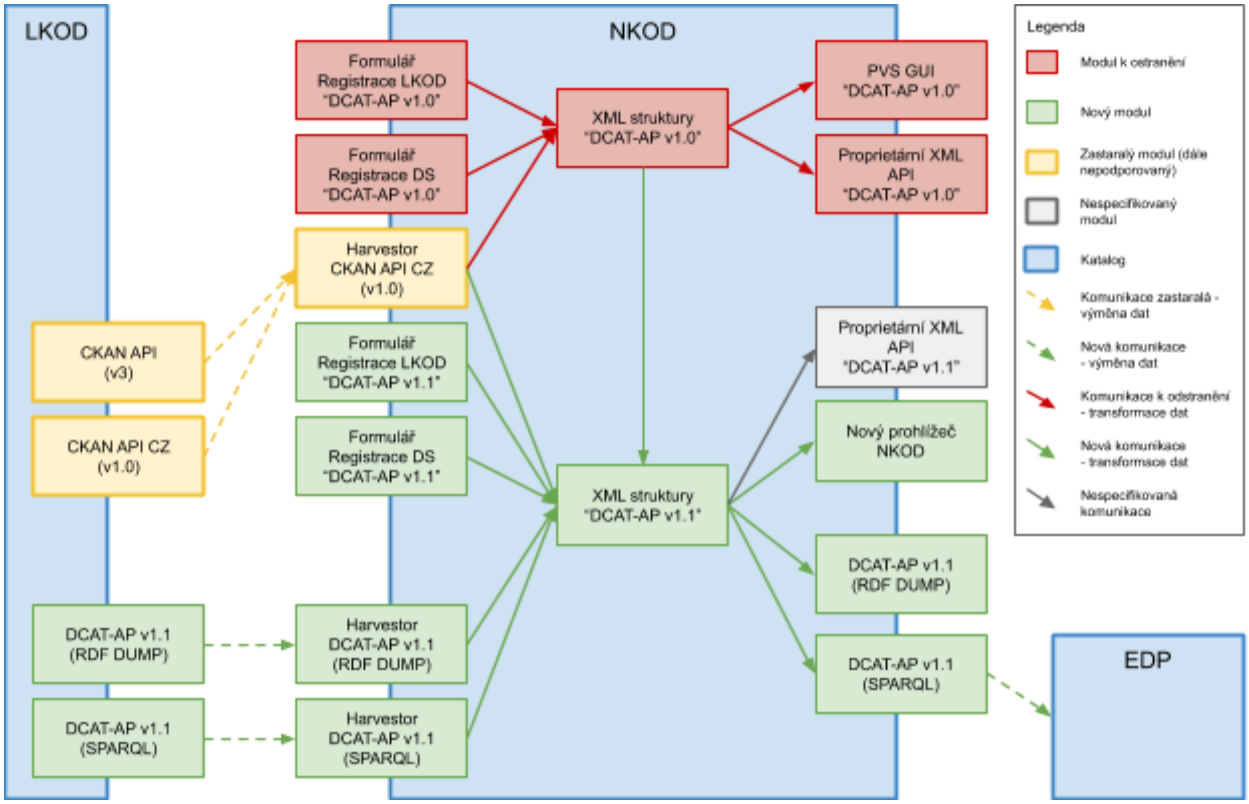
- V současné verzi, NKOD v1.0, je podporován DCAT-AP v1.0. DCAT-AP v1.2.1 však zavádí několik změn, které je nutno reflektovat. DCAT-AP v1.2.1 je standard přijatý na úrovni EU. Je podporován Evropským datovým portálem (EDP).
    - Je nutno zajistit podporu harvestování obsahu LKOD poskytovaného dle standardu DCAT-AP v1.2.1.
    - Je také nutno zajistit poskytnutí obsahu NKOD dle tohoto standardu jednak pro potřeby harvestování obsahu NKOD ze strany EDP a jednak pro ostatní uživatele otevřených dat
    - Je nutno rozšířit formuláře pro registraci LKOD a registraci datové sady tak, aby jejich položky odpovídaly položkám standardu DCAT-AP v1.2.1.
    - Je nutno rozšířit prezentační vrstvu (GUI) NKOD na PVS tak, aby zobrazovala všechny položky v souladu se standardem DCAT-AP v1.2.1
- Další požadavky vyplývají ze současné praxe jednotlivých institucí VS ČR, které poskytují otevřená data, a praxe přicházející z EU.

- Je nutno zachovat podporu harvestování obsahu LKOD ze základního CKAN API pro případ, kdy instituce VS ČR používá pro svůj LKOD nástroj CKAN v základní variantě bez rozšíření.
- Je také nutno zachovat podporu rozšířeného CKAN API definovaného pro potřeby NKOD v1.0 (tzv. CKAN API CZ). Podpora však již nebude dále udržována a bude označena jako "*deprecated*". Toto rozšířené API bylo definováno MV ČR z důvodu zajištění kompatibility CKAN API s DCAT-AP v1.0. V době realizace NKOD v1.0 žádné použitelné rozšíření CKAN API tímto směrem neexistovalo a MV ČR tedy definovalo vlastní.

Některé instituce VS ČR se rozhodly použít pro zajištění svého LKOD nástroj DKAN. Ten sice deklaruje podporu pro poskytování obsahu katalogu v podobě základního CKAN API, avšak v praxi se v řadě položek odlišuje. Pro DKAN je ovšem k dispozici instalovatelný balíček rozšiřující API do podoby zpracovatelné NKOD.

Navrhované změny v architektuře řešení NKOD na PVS

Obrázek 6 zobrazuje architekturu řešení NKOD v1.2.1 ve vazbě na stávající architekturu řešení NKOD v1.0. Ukazuje změny, které je nutno provést, aby byly splněny požadavky uvedené výše.



Obrázek 6: Architektura řešení NKOD v1.1

1. Interní XML struktury

   a. Stávající interní XML struktury odpovídající DCAT-AP v1.0 budou rozšířeny tak, aby odpovídaly standardu DCAT-AP v1.2.1.
      i. Stávající metadatové záznamy budou transformovány ze staré XML struktury do nové.
2. Formulářové rozhraní
   a. Formulářové rozhraní pro registraci LKOD a datové sady (DS) bude upraveno tak, aby odpovídalo standardu DCAT-AP v1.2.1.
      i. Bude implementována transformace formulářových dat do nových interních XML struktur.
3. Harvestory obsahu LKOD
   a. Bude zachován harvestor obsahu LKOD poskytovaného prostřednictvím CKAN API v3.
      i. Harvestor bude označen jako "*deprecated*" a nebude dále udržován.
   b. Bude zachován harvestor obsahu LKOD poskytovaného prostřednictvím CKAN API CZ.
      i. Harvestor bude označen jako "*deprecated*" a nebude dále udržován.
   c. Bude implementován harvestor obsahu LKOD poskytovaného prostřednictvím DCAT-AP v1.2.1 dumpu.
      i. Harvestor bude podporovat všechny RDF serializace: JSON-LD, RDF/XML, Turtle, ...
      ii. Harvestor bude převádět získanou RDF serializaci do nových interních XML struktur.
   d. Bude implementován harvestor obsahu LKOD poskytovaného prostřednictvím DCAT-AP v1.2.1 SPARQL endpointu.
      i. Harvestor bude podporovat načítání pouze změněných datových sad
4. Poskytování obsahu NKOD
   a. Bude přidán export kompletního obsahu NKOD do DCAT-AP v1.2.1 dumpu.
      i. Export bude prováděn do následujících RDF serializací: JSON-LD, RDF/XML, Turtle.
   b. Bude přidán SPARQL endpoint obsahující kompletní obsah NKOD

Formulářové rozhraní - Úpravy položek dle DCAT-AP v1.2.1

U textových položek (název, jméno, popis) je třeba umožnit zadávat více hodnot - jednu pro jeden jazyk, tedy např. jeden název datové sady česky a jeden název datové sady anglicky. Pro snadnou transformaci do RDF pro tyto jazykové mutace doporučujeme jazyk ukládat jako ISO 639-1 2-místný kód.

Datová sada - změny stávajících položek

Název datové sady

Název datové sady (1..n) je nyní třeba umožnit zadávat v jazykových mutacích, jeden název pro každý jazyk.

Popis datové sady

Popis datové sady (1..n) je nyní třeba umožnit zadávat v jazykových mutacích, jeden název pro každý jazyk.

Kurátor

U nás je kurátor vždy člověk. DCAT-AP umožňuje dát obecnější kontakt pro datovou sadu, může to být (Individual, Organization, Location, Group).

Klíčová slova

Nyní je třeba umožnit zadávat v jazykových mutacích.

Poskytovatel dat

Poskytovatel dat se nyní do NKOD vyplňuje na základě ID datové schránky. Nově je třeba umožnit volitelně (0..1) zadat “Typ poskytovatele dat” dle číselníku ADMS pro typ poskytovatele dat: http://purl.org/adms/publishertype/1.0. Hodnotou je tedy IRI, např. http://purl.org/adms/publishertype/LocalAuthority, viditelný by měl být název položky. Jelikož se poskytovatel dat ve formuláři nezadává, navrhujeme toto zařadit do formuláře pro datovou sadu.

Klasifikace

Nově EU zavádí povinný číselník pro klasifikaci témat datových sad. Ten obsahuje 13 položek ve všech evropských jazycích, např. http://publications.europa.eu/resource/authority/data-theme/EDUC. Zde může být hodnotou jak IRI, tak jen kód, ze kterého se později dá IRI sestavit.
Může být volitelně doplněn další klasifikací, tedy aktuálně používaný EUROVOC může zůstat jako doplňkový, nicméně by už neměl být povinný.

Periodicita aktualizace

Nově EU zavádí pro tuto položku povinný číselník pro periodicitu aktualizace datových sad obsahující položky v evropských jazycích, např. http://publications.europa.eu/resource/authority/frequency/MONTHLY.

Územní pokrytí

Nepovinná položka (0..n), pro kterou je povinnost používat číselníky http://publications.europa.eu/mdr/authority/country/, http://publications.europa.eu/mdr/authority/place/, a http://publications.europa.eu/mdr/authority/continent/ a v případě, že dané území v těchto číselnících není, pak IRI z http://sws.geonames.org/. IRI z RÚIAN může zůstat jako doplňkové. Příklady: http://publications.europa.eu/resource/authority/place/CZE_PRG, http://www.geonames.org/3339576/stredocesky-kraj.html

Dotčené časové období

Beze změny

Odkaz na dokumentaci datové sady v lidsky čitelné podobě

Tato položka by nově měla umožnit zadat více dokumentů (0..n).

Datová sada - nové položky

IRI datové sady

IRI datové sady se může generovat automaticky například z ID datové schránky a kódu datové sady, což podpoří i dereferencovatelnost. Ve formuláři by mělo jít zadávat vlastní IRI, které bude připojeno owl:sameAs vazbou.

Otevřenost datové sady

Může nabývtat 3 hodnot, :public (Otevřená), :restricted (Omezená), :non-public (Neveřejná). Číselník bude k dispozici později. V NKOD se jiné než :public nebudou vyskytovat, lze nastavovat implicitně.

Použité specifikace a standardy

Tato položka může obsahovat seznam odkazů na použité specifikace a standardy.

Má verzi/Je verzí

Tato dvojice položek (každá 0..n) umožňuje odkazovat na datové sady, ktere jsou verzí aktuální datové sady / jejichž verzí tato datová sada je. Je třeba umožnit jiné datové sady odkazovat pomocí jejich IRI, i když v NKOD ještě nejsou. To vyžaduje předvídatelnou nebo nastavitelnou podobu IRI datových sad.

Hlavní identifikátor datové sady

0..n hlavních identifikátorů datové sady v katalogu. Toto může být IRI v rámci NKOD a IRI z LKOD.

Stránka datové sady

0..n URL webových stránek, přes které je přístupná datová sada a/nebo její distribuce. Jde o stránku u původního poskytovatele dat, nikoliv v NKOD.

Jazyky datové sady

Tato položka udává jazyky používané v datové sadě. Použit je EU číselník pro jazyky, položkou je IRI nebo kód, např. http://publications.europa.eu/resource/authority/language/CES.

Sekundární identifikátor datové sady

0..n Položka sekundární identifikátor datové sady, např. DOI, EZID, W3ID, DataCite apod. Je potřeba pro každý umožnit zadat i “Typ sekundárního identifikátoru datové sady” dle http://www.sparontologies.net/ontologies/datacite/source.html. Hodnotou je IRI typu (pro použití v RDF).

Dokument o původu datové sady

0..n Odkaz na dokument s informacemi o původu datové sady. Dokument může obsahovat informace o tom, jak byla data pořízena, zpracovávána a publikována.

Odkaz na související datové entity

0..n Odkazy na související datové entity (ve smyslu RDF)

Datum a čas publikace datové sady

Nepovinná položka. Jedná se o oficiální uveřejnění datové sady, což není datum a čas registrace v NKOD.

Zdrojová datová sada

0..n Pokud se jedná o datovou sadu odvozenou z jiné datové sady, tato položka obsahuje odkaz na zdrojovou datovou sadu.

Datum a čas poslední modifikace datové sady

Nepovinná položka pro datum a čas poslední změny datové sady.

Vzorek dat

0..n Jedná se o speciální případ datového zdroje obsahující vzorek dat.

Typ datové sady

Nepovinná položka stanovující typ datové sady.

Verze datové sady

Nepovinná položka obsahující číslo verze datové sady

Poznámky k verzi datové sady

0..n (pro různé jazyky, v každém 1x) Obsahuje text s popisem změn oproti minulé verzi datové sady.

Datový zdroj - změny stávajících položek

Odkaz na webovou stránku zpřístupňující data

1..n Povinná položka. Musí obsahovat buďto odkaz na stránku zpřístupňující datový zdroj, nebo na datový zdroj samotný. V případě, že se jedná o datový zdroj samotný, tento odkaz bude duplikován do odkazu ke stažení datového zdroje

Podmínky užití datového zdroje

0..1 Doporučená položka, odkaz na podmínky užití. Ve formuláři zůstane povinná.

Odkaz na soubor ke stažení

0..n Odkaz na stažení souboru v daném formátu.

Odkaz na strojově čitelné schéma

0..n - Nově seznam

Fromát datového zdroje (MIME typ)

Nepovinná položka. Skutečný mime typ z číselníku http://www.iana.org/assignments/media-types/media-types.xhtml, např. http://www.iana.org/assignments/media-types/text/turtle

Název datového zdroje

0..n (jazyky)

Datový zdroj - nové položky

IRI datového zdroje

Jedná se o IRI entity reprezentující informace o datovém zdroji z hlediska katalogu, tj. není to URL pro stažení datového zdroje.

Typ souboru datového zdroje

Doporučená položka, 0..1. Specifikuje typ souboru datového zdroje dle EU číselníku pro typy souborů. Například http://publications.europa.eu/resource/authority/file-type/CSV, hodnotou může být IRI nebo kód.

Typ podmínek užití datového zdroje

Doporučená položka, popisuje typ podmínek užití dle číselníku http://purl.org/adms/licencetype/, např. http://purl.org/adms/licencetype/PublicDomain

Popis datového zdroje

0..n (jazyky). Položka pro popis datového zdroje, ve více jazycích.

Velikost souboru v bytech

Volitelná položka, udává velikost datového zdroje v bytech.

Kontrolní součet

Položka pro zadání kontrolního součtu algoritmem SHA-1 (obecně i jiným, ale zatím je podporován jen tento).

Odkaz na dokument o datovém zdroji

0..n Odkaz na dokument o datovém zdroji

Jazyk datového zdroje

Tato položka udává jazyky používané v datovém zdroji. Použit je EU číselník pro jazyky, položkou je IRI nebo kód, např. http://publications.europa.eu/resource/authority/language/CES.

Datum publikace datového zdroje

Volitelná položka, datum publikace datového zdroje (jiné než zanesení do katalogu).

Specifikace práv k datovému zdroji

Volitelná položka, odkazuje se na dokument (není jasný vztah k licenci), zřejmě se bude jednat o strojově čitelný popis dle jednoho ze slovníků Further activities in this area are undertaken by the Open Data Institute[1] with the Open Data Rights Statement Vocabulary[2] and by the Open Digital Rights Language (ODRL) Initiative[3].

Vyspělost datového zdroje

Volitelná položka, z číselníku http://purl.org/adms/status/, např. http://purl.org/adms/status/Completed

Datum poslední změny datového zdroje

Volitelná položka, datum poslední změny datového zdroje.

Typ datového zdroje

Povinná položka udávající, zda jde o soubor ke stažení, webovou službu, vizualizaci, nebo informační kanál.

Datový zdroj - položky k odebrání

Formát strojově čitelného popisu (schematu)

V DCAT-AP není, lze vypustit.

[1] Open Data Institute. http://www.theodi.org/
[2] Open Data Institute. Open Data Rights Statement Vocabulary. http://schema.theodi.org/odrs/
[3] Open Digital Rights Language (ODRL) Initiative. http://www.w3.org/community/odrl/

Časové pokrytí

V DCAT-AP řešeno verzemi datové sady. Lze vypustit.

Územní pokrytí

V DCAT-AP řešeno verzemi datové sady. Lze vypustit.

Registrace katalogu - změny stávajících položek

Poskytovatel katalogu

Poskytovatel katalogu se nyní do NKOD vyplňuje na základě ID datové schránky. Nově je třeba umožnit volitelně (0..1) zadat “Typ poskytovatele dat” dle číselníku ADMS pro typ poskytovatele dat: http://purl.org/adms/publishertype/1.0. Hodnotou je tedy IRI, např. http://purl.org/adms/publishertype/LocalAuthority, viditelný by měl být název položky. Jelikož se poskytovatel dat ve formuláři nezadává, navrhujeme toto zařadit do formuláře pro registraci katalogu nebo vyplňovat na základě datové sady Orgány veřejné moci.

Název katalogu

Nově 1..n pro jazyky.

Jméno správce

Beze změny

Email správce

Beze změny

Registrace katalogu - nové položky

IRI katalogu

Stejně jako u datové sady a datového zdroje, katalog má své IRI.

Popis katalogu

0..n pro jazyky

Hlavní stránka katalogu

Doporučená položka odkazující na originální stránku katalogu

Jazyk záznamů v katalogu

Tato položka udává jazyky používané v katalogu. Použit je EU číselník pro jazyky, položkou je URI nebo kód, např. http://publications.europa.eu/resource/authority/language/CES.

Podmínky užití obsahu katalogu

0..1 Doporučená položka, odkaz na podmínky užití. Ve formuláři bude povinná.

Typ podmínek užití obsahu katalogu

Doporučená položka, popisuje typ podmínek užití dle číselníku http://purl.org/adms/licencetype/, např. http://purl.org/adms/licencetype/PublicDomain

Datum a čas publikace katalogu

Doporučená položka

Klasifikace použitá pro datové sady

Doporučená položka 0..n, IRI skos:ConceptSchemes použitých pro klasifikaci datových sad. Pro povinnou EU klasifikaci http://publications.europa.eu/resource/authority/data-theme, EuroVoc: http://eurovoc.europa.eu/100141

Datum a čas poslední změny katalogu

Doporučená položka

Část katalogu/Je částí katalogu

Možnost mít podkatalogy, oproti datovým sadám může být katalog podkatalogem jen jednoho katalogu.

Specifikace práv k obsahu katalogu

Volitelná položka, odkazuje se na dokument (není jasný vztah k licenci), zřejmě se bude jednat o strojově čitelný popis dle jednoho ze slovníků Further activities in this area are undertaken by the Open Data Institute[4] with the Open Data Rights Statement Vocabulary[5] and by the Open Digital Rights Language (ODRL) Initiative[6].

Územní pokrytí katalogu

Nepovinná položka (0..n), pro kterou je povinnost používat číselníky http://publications.europa.eu/mdr/authority/country/, http://publications.europa.eu/mdr/authority/place/, a http://publications.europa.eu/mdr/authority/continent/ a v případě, že dané území v těchto číselnících není, pak URI z http://sws.geonames.org/. RÚIAN může zůstat jako doplňkový. Příklady: http://publications.europa.eu/resource/authority/place/CZE_PRG, http://www.geonames.org/3339576/stredocesky-kraj.html

[4] Open Data Institute. http://www.theodi.org/
[5] Open Data Institute. Open Data Rights Statement Vocabulary. http://schema.theodi.org/odrs/
[6] Open Digital Rights Language (ODRL) Initiative. http://www.w3.org/community/odrl/

Importní rozhraní

Harvestování lokálních katalogů z RDF

Z dumpu dle DCAT-AP v1.2.1 ve standardních serializacích RDF 1.1: N-Triples, N-Quads, Turtle, TriG, RDF/XML, JSON-LD, RDFa.
Pro splnění požadavků DCAT-AP je třeba PLNÁ podpora všech (i volitelných) položek.

Harvestování lokálních katalogů z CKAN

https://docs.google.com/spreadsheets/d/1QfU3Qng3JbrFSW8ikP6F5r6z1AVzpCtBwsSye8tTRhs/edit#gid=0

Úložiště katalogizačních záznamů a publikovaných dat

Rejstříkové řešení zůstává, zůstává i rozdělení na rejstřík datových zdrojů a rejstřík “Datové sady”.
Změní se ovšem ukládané informace.

Poskytnutí obsahu katalogu v RDF

Katalog je třeba poskytnout jako RDF dump dle DCAT-AP v1.2.1 ve standardních serializacích RDF 1.1: N-Triples, Turtle, RDF/XML, JSON-LD, TriG, N-Quads.
Pro harvestování EDP je třeba data vystavit i jako SPARQL endpoint, ve kterém jsou záznamy o datových sadách odděleny do separátních pojmenovaných grafů.