Evropská unie
Evropský sociální fond
Operační program Zaměstnanost

MINISTERSTVO VNITRA
ČESKÉ REPUBLIKY

C2V3
Návrh a realizace prototypu lokálního katalogu otevřených dat

Vytvořeno v rámci projektu

Rozvoj datových politik v oblasti zlepšování kvality a interoperability dat veřejné správy
CZ.03.4.74/0.0/0.0/15_025/0013983

Klíčová aktivita: 03 Návrhy a realizace opatření pro zlepšování kvality otevřených dat

Indikátor: 8 05 00 Počet napsaných a zveřejněných analytických a strategických dokumentů (vč. evaluačních)

Verze výstupu: 01

Požadavky na Lokální katalogy otevřených dat

Při návrhu a realizaci prototypů Lokálních katalogů otevřených dat byl nejdůležitější požadavek soulad jejich rozhraní s Otevřenou formální normou Rozhraní katalogů otevřených dat DCAT-AP-CZ[1]. Ta definuje dvě základní rozhraní - DCAT-AP Dokumenty, založené na sadě souborů vystavených na webu, a DCAT-AP SPARQL Endpoint, která staví na tom, že potřebné katalogizační záznamy jsou nahrané v tzv. SPARQL Endpointu, což je webová služba umožňující dotazování nad RDF databází.

CKAN ani DKAN nestačí

Aktuální rozšířené implementace datových katalogů CKAN[2] a DKAN[3] nestačí. Standardem pro metadata v Evropské unii, a základem pro DCAT-AP-CZ je DCAT-AP 2.0.1[4], který vyžaduje například využívání EU slovníků a číselníků (EU Vocabularies[5]), umožňuje mít vícejazyčná metadata a je založen na principu Propojených dat. CKAN ani DKAN i přes existenci různých DCAT rozšíření[6] nedosahují potřebné úrovně kompatibility.

Co je potřeba pro kompatibilitu s NKOD

Nejprve je třeba si ujasnit, co je nezbytné pro dosažení kompatibility s NKOD. Je to pouze poskytnutí strojového rozhraní odpovídajícího OFN Rozhraní katalogů otevřených dat. Jedná se tedy buď o rozhraní DCAT-AP Dokumenty[7], nebo DCAT-AP SPARQL Endpoint[8]. Uživatelské rozhraní či vizualizační aplikace nejsou potřeba, datové sady budou uživateli nalezitelné přímo v Národním katalogu otevřených dat[9]. OFN se v čase vyvíjí podle toho, jak se vyvíjejí standardy DCAT a DCAT-AP a jak je podle nich rozšiřován NKOD. Teď se to děje zhruba jednou za rok.

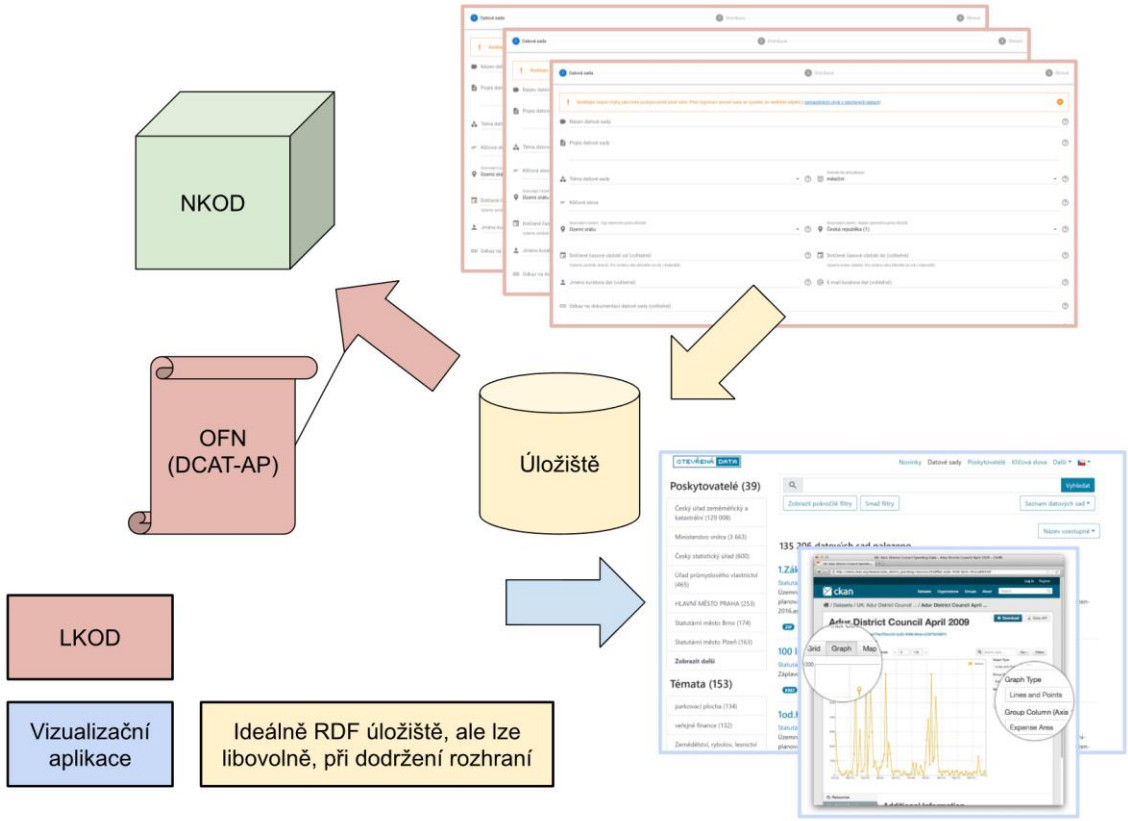
Způsob, jakým je rozhraní implementováno, OFN nespecifikuje. Lze ho tedy implementovat libovolně. Několik vybraných způsobů:

1. Ručně tvořené soubory s obsahem dle OFN, umístěné na web - toto lze použít pro malé, a ne často aktualizované LKODy. Toto řešení je však velmi levné a snadné. Jako formulář pro zadávání dat totiž poslouží formulář NKOD[10], ve kterém se pouze na konci přepne do režimu "Stáhnout nový záznam pro LKOD", vyplní se IRI datové sady a IRI poskytovatele, a stáhne se hotový katalogizační záznam, který lze vystavit

[1] https://ofn.gov.cz/rozhraní-katalogů-otevřených-dat/2021-01-11/
[2] https://ckan.org/
[3] https://getdkan.org/
[4] https://joinup.ec.europa.eu/collection/semantic-interoperability-community-semic/solution/dcat-application-profile-data-portals-europe/release/201-0
[5] https://publications.europa.eu/en/web/eu-vocabularies/about
[6] https://github.com/ckan/ckanext-dcat
[7] https://ofn.gov.cz/rozhraní-katalogů-otevřených-dat/2021-01-11/#dcat-ap-dokumenty
[8] https://ofn.gov.cz/rozhraní-katalogů-otevřených-dat/2021-01-11/#dcat-ap-sparql-endpoint
[9] https://data.gov.cz
[10] https://data.gov.cz/formulář/registrace-datové-sady

na web. Pokud je jako úložiště využit GitHub[11], je toto vystavení zcela zdarma a lze zajistit i automatizované generování souboru s katalogem. V opačném případě je třeba zajistit existenci souboru s katalogem, který odkazuje na jednotlivé záznamy datových sad. URL souboru s katalogem je pak zaregistrováno v NKOD.

2. Proprietární systém, který dané soubory (či SPARQL endpoint) zpřístupňuje.
3. Vlastní silou rozšířený katalog CKAN, DKAN či jiný, který poskytne rozhraní pro vkládání dat, a navíc se zajistí transformace těchto dat do formy dle OFN. Toho lze docílit například nástrojem LinkedPipes ETL (open-source)[12].
4. Použití jedné z variant prototypu LKOD, nebo jejich částí.



LKOD je vzhledem k NKOD tedy pouze rozhraní, které splňuje Otevřenou formální normu (OFN) Rozhraní katalogů otevřených dat: DCAT-AP-CZ. Měl by poskytovateli umožnit pohodlně zadávat katalogizační záznamy, tak jako to teď dělá formulář pro vkládání záznamu do NKOD. Další funkcionalita, jako uživatelské prohlížení záznamů či vizualizace dat jsou volitelnou, nikoliv nutnou součástí.

Referenční implementace LKOD - DCAT-AP Dokumenty

Tato varianta používá pouze formulář NKOD a GitHub.

[11] https://github.com/
[12] https://etl.linkedpipes.com/

Požadavy referenční implementace DCAT-AP Dokumenty

- Repozitář na GitHubu. Nejjednodušší je udělat fork minimalistického LKOD[13] (viz Přilohy). Sem bude poskytovatel ukládat katalogizační záznamy. GitHub zde řeší i přihlašování uživatelů a řízení přístupu uživatelů k repozitáři.
- Uživatel plnící LKOD musí umět pracovat s GitHubem.

Workflow referenční implementace DCAT-AP Dokumenty

S touto implementací se pak pracuje následovně:

1. Pro tvorbu katalogizačního záznamu se využije formulář pro vkládání záznamu do NKOD.
2. V posledním kroce se tlačítkem se symbolem ozubeného kolečka přepne do režimu "Stáhnout nový záznam pro LKOD", vyplní se IRI datové sady a IRI poskytovatele.
3. Vyplněné IRI datové sady musí odpovídat URL souboru výsledného katalogizačního záznamu pro přístup k jeho nezměněné podobě po jeho uložení do GitHub repozitáře. Např. `https://raw.githubusercontent.com/opendata-mvcr/lkod-mvcr/master/rpp/dokumenty-převodu-agend.jsonld` pro přístup k `https://github.com/opendata-mvcr/lkod-mvcr/blob/master/rpp/dokumenty-převodu-agend.jsonld`. Toto URL pak slouží i pro opětovné načtení katalogizačního záznamu do formuláře NKOD, například pro jeho editaci.
4. Výsledný JSON-LD soubor se nahraje na GitHub do repozitáře poskytovatele (Příklad MV ČR[14]).
5. GitHub po této akci automaticky spustí GitHub Actions skript[15] generující soubor katalogu, který odkazuje na jednotlivé záznamy datových sad. Spuštění tohoto skriptu musí být nakonfigurováno v daném repozitáři.
6. URL souboru katalogu, opět ve formě pro přístup k jeho nezměněné podobě po jeho uložení do GitHub repozitáře, je pak zaregistrováno do NKOD[16].

Referenční implementace LKOD - SPARQL Endpoint

Referenční implementace je v tuto chvíli tvořena pomocí komponent v aktuální implementaci Národního katalogu otevřených dat. Tato varianta používá výhradně open-source software a GitHub, je pro ni potřeba mít k dispozici (virtuální) server. Aktuálně na ní běží například

[13] https://github.com/opendata-mvcr/lkod-min
[14] https://github.com/opendata-mvcr/lkod-mvcr
[15] https://github.com/opendata-mvcr/lkod-github-actions/tree/master/create-catalog-file
[16] https://opendata.gov.cz/cinnost:registrace-vlastniho-katalogu-v-nkod

LKOD MV ČR[17]. Jednotlivé kroky či komponenty lze libovolně vyměňovat za jiné, za předpokladu dodržení rozhraní pro NKOD.

Požadavky referenční implementace SPARQL endpoint

- Server, kde LKOD poběží, nejlépe s OS Linux (Ubuntu, Debian, ...)
- LinkedPipes ETL - Open-source software pro transformace propojených dat
- Nějaká implementace SPARQL Endpointu. Například Openlink Virtuoso Open-Source 7[18], Apache Jena Fuseki[19], Eclipse RDF4J[20], Blazegraph[21] apod.
- Repozitář na GitHubu. Sem bude poskytovatel ukládat katalogizační záznamy.
- Web server pro implementaci zabezpečení přístupu k jednotlivým komponentám a příjem Webhooks, např. nginx[22]
- Uživatel plnící LKOD musí umět pracovat s GitHubem.

Workflow referenční implementace SPARQL Endpoint

S touto implementací se pak pracuje následovně:

1. Pro tvorbu katalogizačního záznamu se využije formulář pro vkládání záznamu do NKOD.
2. V posledním kroce se tlačítkem se symbolem ozubeného kolečka přepne do režimu "Stáhnout nový záznam pro LKOD", vyplní se IRI datové sady a IRI poskytovatele.
3. Výsledný JSON-LD soubor se nahraje na GitHub do repozitáře poskytovatele (Příklad MV ČR[23]). GitHub zde řeší oprávnění uživatelů k editaci jednotlivých záznamů a automatizaci následného aktualizačního procesu.
4. GitHub po této akci automaticky zavolá tzv. Webhook, který aktualizuje klon repozitáře na serveru a spustí transformační proces v nástroji LinkedPipes ETL. Podívejte se na vzorový skript[24] pro obsluhu Webhooku v PHP. Stačí na tento skript nasměrovat GitHub Webhook, vyplnit secret, informace o adresáři s klonem GitHub repozitáře katalogu a IRI obslužné pipeline[25] v LinkedPipes ETL - vzorová pipeline pracuje s Openlink Virtuoso a je třeba v ní upravit metadata tvořeného lokálního katalogu.
5. Proces v LinkedPipes ETL stáhne záznamy z GitHub a nahraje je do SPARQL endpointu (rozhraní pro NKOD).
6. SPARQL endpoint je pak zaregistrován do NKOD.

[17] https://data.mvcr.gov.cz/
[18] https://github.com/openlink/virtuoso-opensource/
[19] https://jena.apache.org/documentation/fuseki2/
[20] https://rdf4j.org/
[21] https://blazegraph.com/
[22] http://nginx.org/
[23] https://github.com/opendata-mvcr/lkod-mvcr
[24] https://github.com/opendata-mvcr/lkod/blob/master/lkod-sparl/webhook.php a viz Příloha
[25] https://github.com/opendata-mvcr/lkod/blob/master/lkod-sparql/LKOD%20z%20GitHub%20repozitáře%20do%20SPARQL%20endpointu.jsonld a viz Příloha

Přehled aktuálního nasazení prototypů LKOD

Oba prototypy byly nasazeny v produkčním prostředí, což demonstruje jejich použitelnost. Další řada poskytovatelů se prototypem inspirovala při tvorbě svých vlastních řešení lokálních katalogů.

1. Katalog otevřených dat Ministerstva vnitra ČR
    a. **Poskytovatel:** Ministerstvo vnitra
    b. **Varianta prototypu:** DCAT-AP SPARQL endpoint
    c. **URL:** https://data.mvcr.gov.cz/sparql
2. Katalog otevřených dat Ministerstva zemědělství
    a. **Poskytovatel:** Ministerstvo zemědělství
    b. **Varianta prototypu:** DCAT-AP Dokumenty
    c. **URL:** https://raw.githubusercontent.com/ondrejsilhacek/lkod-min/main/katalog.jsonld
3. Ministerstvo pro místní rozvoj
    a. **Poskytovatel:** Ministerstvo pro místní rozvoj
    b. **Varianta prototypu:** DCAT-AP Dokumenty
    c. **URL:** https://raw.githubusercontent.com/opendata-mmr/lkod-min/main/katalog.jsonld

Přílohy

V přílohách jsou zdrojové kódy použitého open-source software vyjma RDF úložiště a 2 prototypy LKOD podle 2 typů rozhraní DCAT-AP v Otevřené formální normě Rozhraní katalogů otevřených dat: DCAT-AP-CZ.

Použitý open-source software

1. LinkedPipes ETL (fork v https://github.com/opendata-mvcr/etl)
    - soubor `etl-develop.zip`
2. LinkedPipes DCAT-AP Forms (fork v https://github.com/opendata-mvcr/dcat-ap-forms)
    - soubor `dcat-ap-forms-develop.zip`

Popis prototypů lokálního katalogu otevřených dat včetně varianty DCAT-AP SPARQL Endpoint

1. https://github.com/opendata-mvcr/lkod
    - soubor `lkod-master.zip`

Repozitář použitelný pro variantu LKOD DCAT-AP Dokumenty

1. https://github.com/opendata-mvcr/lkod-min - samotný repozitář katalogu
    - soubor `lkod-min-main.zip`

2. https://github.com/opendata-mvcr/lkod-github-actions - GitHub Actions skript generující soubor katalogu
    - soubor `lkod-github-actions-master.zip`