




jakubklimek Iniciální verze dokumentace NKOD

 History

 1 contributor

 204 lines (160 sloc) | 14.4 KB

...

Aplikační příručka - Národní katalog otevřených dat

1. ROZSAH PLATNOSTI A ÚČEL

Tato aplikační příručka slouží k popisu Národního katalogu otevřených dat (NKOD), části projektu OD2.0.

2. DEFINÍCIA POJMOV A SKRATIEK

CORS

Cross-Origin Resource Sharing

DCAT-AP-SK 2.0

Specifikace pro metadatový záznam datasetu otevřených dat

k8s

Kubernetes

LKOD

Lokální katalog otevřených dat

NKOD

Národní katalog otevřených dat

POD

Portál otevřených dat

RDF

Resource Description Framework - datový model využívaný NKOD

SPARQL

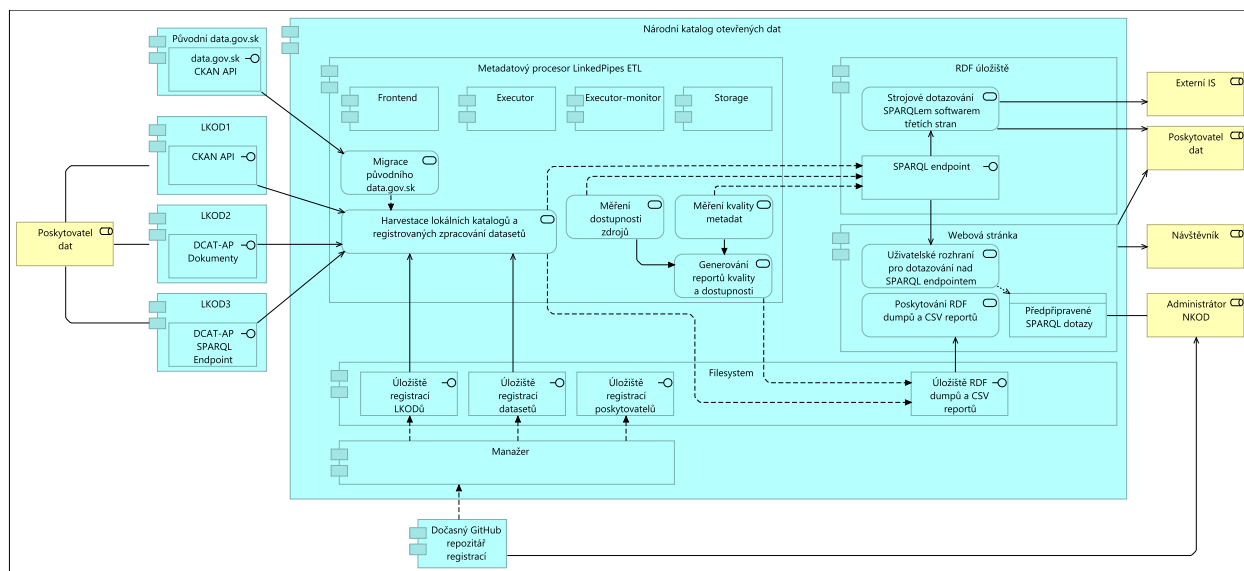
Dotazovací jazyk nad daty v RDF

3. POPIS FUNKCIONALITY IS

3.1 STRUČNÝ POPIS IS

Národní katalog otevřených dat (NKOD) obsahuje zejména databázi metadatových záznamů datasetů otevřených dat poskytovaných různými institucemi veřejné správy, která poskytuje [SPARQL](#) endpoint pro dotazování. V databázi se zrcadlí metadataové záznamy datasetů registrovaných jednotlivě přímo v NKOD a záznamy pocházející z Lokálních katalogů otevřených dat (LKOD) provozovaných přímo poskytovateli dat. Metadatové záznamy odpovídají specifikaci [DCAT-AP-SK 2.0](#). Databáze NKOD je tvořena pravidelně denně. Po každém vytvoření databáze NKOD je zhodnocena i kvalita metadatových záznamů vzhledem k [DCAT-AP-SK 2.0](#) a dostupnost registrovaných zdrojů. Na základě naměřených hodnot jsou vygenerovány reporty obsahující zjištěné skutečnosti.

3.2 ZOZNAM A ZÁKLADNÝ POPIS SUBSYSTÉMOV A FUNKCÍ

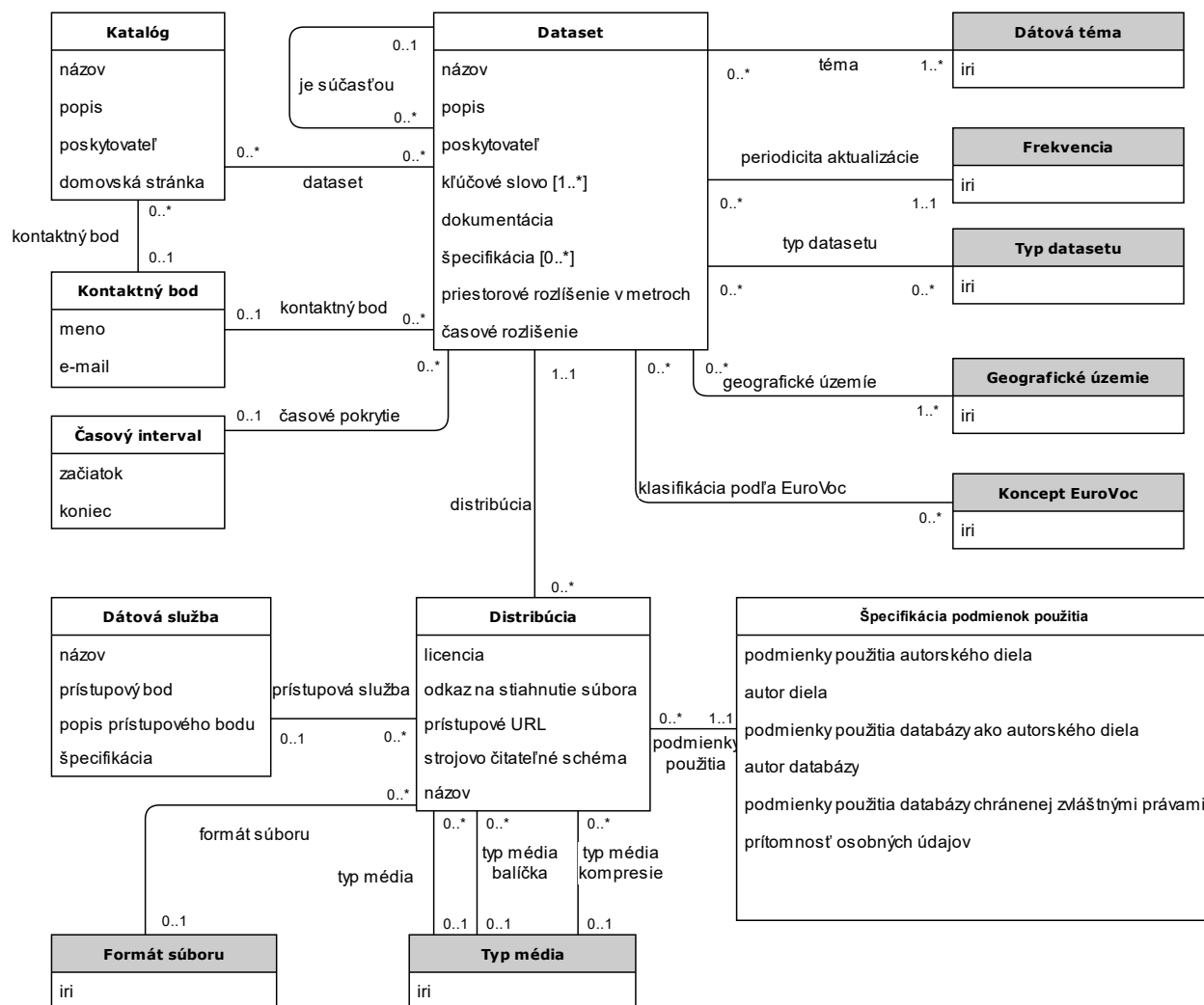


NKOD se skládá z následujících komponent:

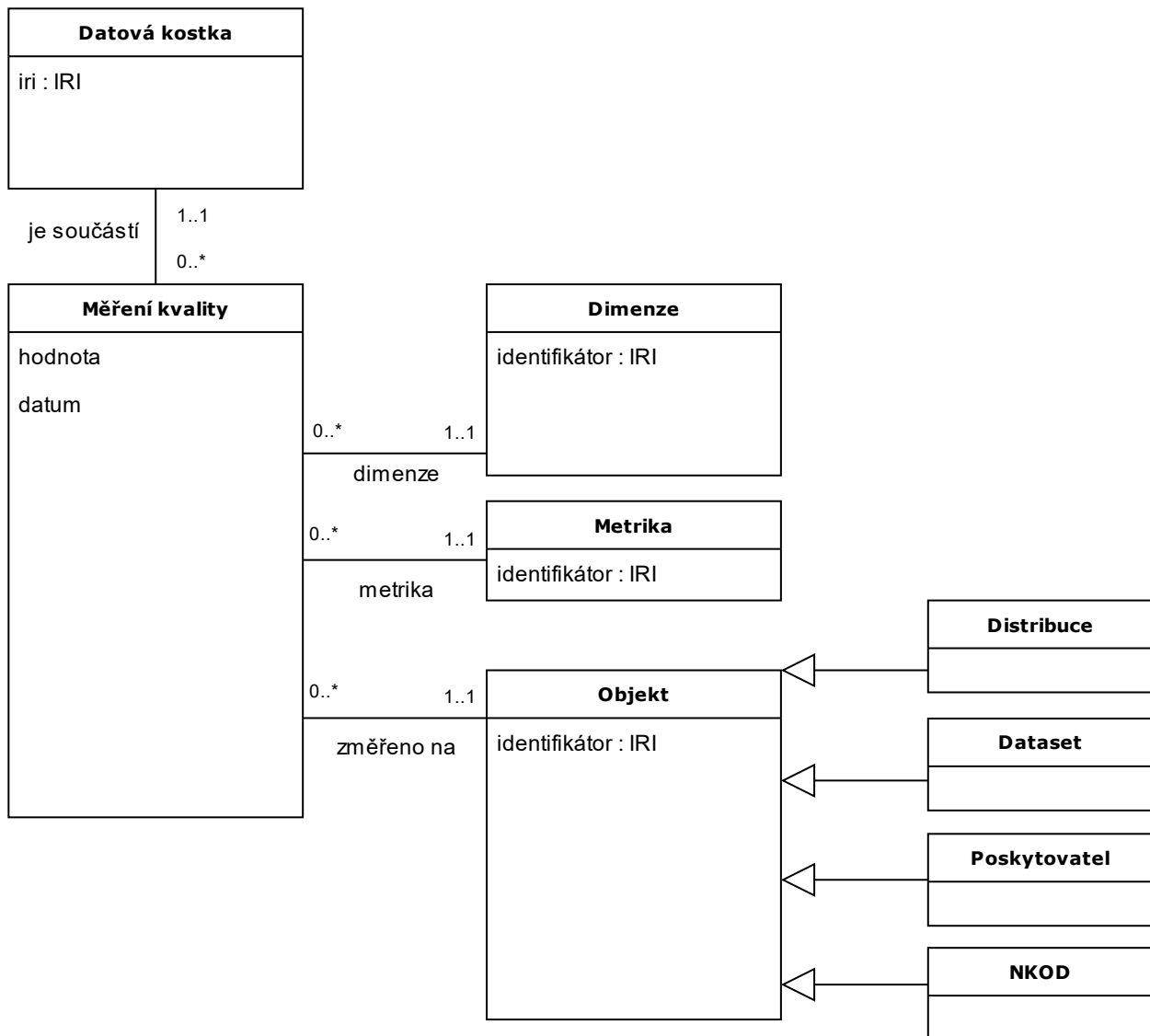
1. RDF úložiště implementované pomocí [Ontotext GraphDB Free](#) poskytující SPARQL endpoint pro dotazování
2. Metadatový procesor a harvester implementovaný pomocí [LinkedPipes ETL](#). Samotné datové procesy, harvestování lokálních katalogů, měření kvality a generování reportů jsou implementovány jako datové pipeline, které jsou uloženy a zdokumentovány v [GitHub repozitáři](#), včetně pipeline zajišťujících migraci ze stávajícího data.gov.sk.
3. Webová stránka zpřístupňující stahování obsahu NKOD a SPARQL dotazování včetně předpřipravených SPARQL dotazů
4. Manažer zajišťuje, že datové pipeline implementující harvestaci a měření kvality se před jejich spuštěním aktualizují z [GitHub repozitáře](#). Dále na pokyn plánovače spouští samotnou harvestaci. Nakonec tato komponenta zajišťuje, že registrační záznamy datových sad, lokálních katalogů otevřených dat a jednotlivých poskytovatelů, které dočasně administrátor NKOD ukládá v [GitHub repozitáři](#), se před harvestací dostanou do prostředí NKOD. Tato poslední funkcionality bude později upravena/nahrazena správou registračních záznamů v jiné části projektu OD2.0 - v Portálu otevřených dat (POD).

3.3 DÁTOVÝ MODEL APLIKÁCIE

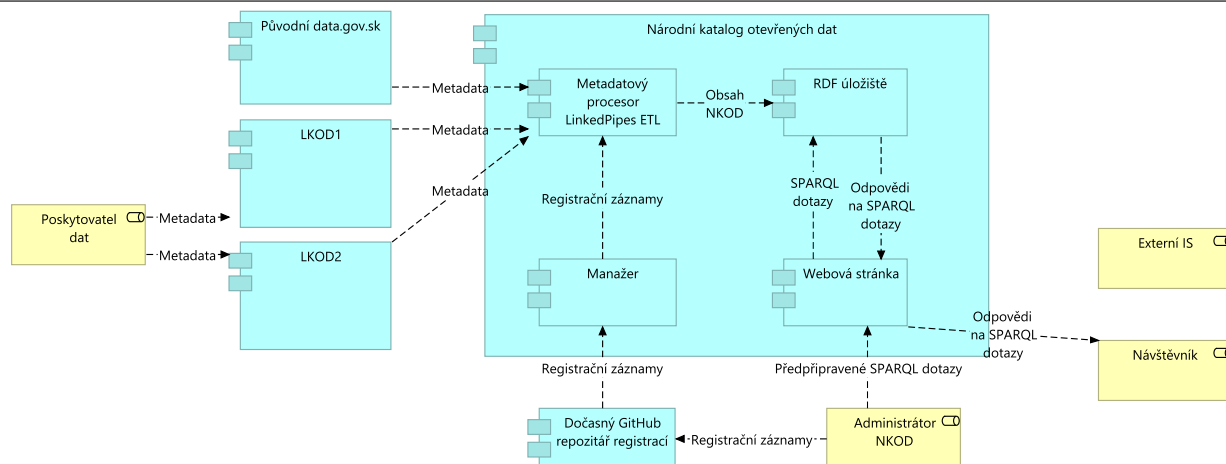
Data NKOD jsou uchovávaná v datovém modelu [Resource Description Framework \(RDF\)](#) a skládají se z metadat datových sad registrovaných do NKOD či harvestovaných z LKOD. Jejich struktura se řídí specifikací [DCAT-AP-SK 2.0](#).



Časť týkajúci sa méréni kvality záznamů pak vychází ze slovníku [Data Quality Vocabulary](#).



3.4 ZÁKLADNÝ FLOWCHART IS



V diagramu vidíme uživatelské role

Administrátor NKOD

Edituje seznam předpřipravených SPARQL dotazů a dočasně spravuje registrační záznamy, než správu převzmou poskytovatelé dat pomocí budoucího Portálu otevřených dat.

Poskytovatel dat

Provozuje registrované LKODY. V budoucnosti bude skrz Portál otevřených dat spravovat své registrační záznamy. Může vystupovat i v roli Návštěvníka.

Externí IS

Ptá se SPARQLelem na SPARQL endpoint RDF úložiště, který zpřístupňuje Webová stránka.

Návštěvník

Ptá se SPARQLelem na SPARQL endpoint RDF úložiště. SPARQL dotaz píše ve formuláři na Webové stránce, nebo si vybere jeden z Administrátorem NKOD připravených SPARQL dotazů.

3.5 POPIS EXISTUJÍCÍCH ROZHRANÍ V IS

3.5.1 Webová stránka

Toto je testovacia verzia SPARQL Endpointu, ktorý bude súčasťou nového portálu otvorených dát. Zdrojom dát je stále súčasný data.gov.sk ktorý sa do NKODu harvestuje raz za deň.

opendata.mirri.tech

Slovenčina



SPARQL Endpoint pre Národný katalóg otvorených dát

Príklady dotazov

[100 datasetov a ich poskytovateľov](#)

100 datasetov a ich poskytovateľov

[Zoznam lokálnych katalógov údajov](#)

Zoznam lokálnych dátových katalógov a počty dátových sád v nich

[Počet datasetov na poskytovateľa](#)

Počet datasetov na poskytovateľa

[Dáta v CSV](#)

Datsety s distribúciami v CSV

[Nedostupnosť stiahnuteľných súborov](#)

Nedostupnosť stiahnuteľných súborov na základe meraní kvality

Zadaj SPARQL dotaz

Query

```
1 PREFIX dcterms: <http://purl.org/dc/terms/>
2 PREFIX dcat: <http://www.w3.org/ns/dcat#>
3 PREFIX foaf: <http://xmlns.com/foaf/0.1/>
4
5 SELECT ?mime_type (count (distinct ?distribuce) as ?počet_distribucí)
6 WHERE
7 {
8   ?datová_sada a dcat:Dataset ;
9   dcat:distribution ?distribuce .
10
11   ?distribuce dcat:mediaType ?mime_type.
12 }
13 GROUP BY ?mime_type
14 ORDER BY DESC(?počet_distribucí)
```

Table

Response

15 results in 0.05 seconds

Simple view

Ellipse

Filter query results

Page size: 50



mime_type

počet_distribucí

1	<http://www.iana.org/assignments/media-types/application/xml>	"4051"^^<http://www.w3.org/2001/XMLSchema#integer>
2	<http://www.iana.org/assignments/media-types/text/csv>	"3016"^^<http://www.w3.org/2001/XMLSchema#integer>
3	<http://www.iana.org/assignments/media-types/application/vnd.openxmlformats-officetype/ms-excel+xml>	"1002"^^<http://www.w3.org/2001/XMLSchema#integer>
4	<http://www.iana.org/assignments/media-types/text/html>	"707"^^<http://www.w3.org/2001/XMLSchema#integer>
5	<http://www.iana.org/assignments/media-types/application/pdf;type=pdf1x>	"668"^^<http://www.w3.org/2001/XMLSchema#integer>
6	<http://www.iana.org/assignments/media-types/application/vnd.ms-excel>	"532"^^<http://www.w3.org/2001/XMLSchema#integer>
7	<http://www.iana.org/assignments/media-types/application/zip>	"341"^^<http://www.w3.org/2001/XMLSchema#integer>
8	<http://www.iana.org/assignments/media-types/application/geo+json>	"20"^^<http://www.w3.org/2001/XMLSchema#integer>
9	<http://www.iana.org/assignments/media-types/application/vnd.openxmlformats-officetype/ms-excel+xml>	"18"^^<http://www.w3.org/2001/XMLSchema#integer>
10	<http://www.iana.org/assignments/media-types/application/xhtml+xml>	"16"^^<http://www.w3.org/2001/XMLSchema#integer>

```
10 <http://www.iana.org/assignments/media-types/application/xml>
11 <http://www.iana.org/assignments/media-types/application/json>
12 <http://www.iana.org/assignments/media-types/image/tiff>
13 <http://www.iana.org/assignments/media-types/text/rtf>
14 <http://www.iana.org/assignments/media-types/text/plain>
15 <http://www.iana.org/assignments/media-types/application/msword>
```

Showing 1 to 15 of 15 entries

< 1 >

[Našli ste na stránke chybu?](#)

Webová stránka NKOD je hlavné rozhraní pro návštevníky NKOD - umožňuje vybrať predpripravený SPARQL dotaz alebo napsať vlastní, a nechať ho vykonať na rozhraní SPARQL endpoint nad RDF úložiskom NKOD a zobraziť výsledky. Pro samotnou prácu sa SPARQL dotazom využíva knihovňa [Yasgui](#). Webová stránka je k dispozícii v niekoľkých jazykoch, teraz v 5 jazykoch. Prostredníctvom webovej stránky je prístupný aj obsah NKOD v podobu súborov na stiahnutie.

3.5.2 SPARQL endpoint

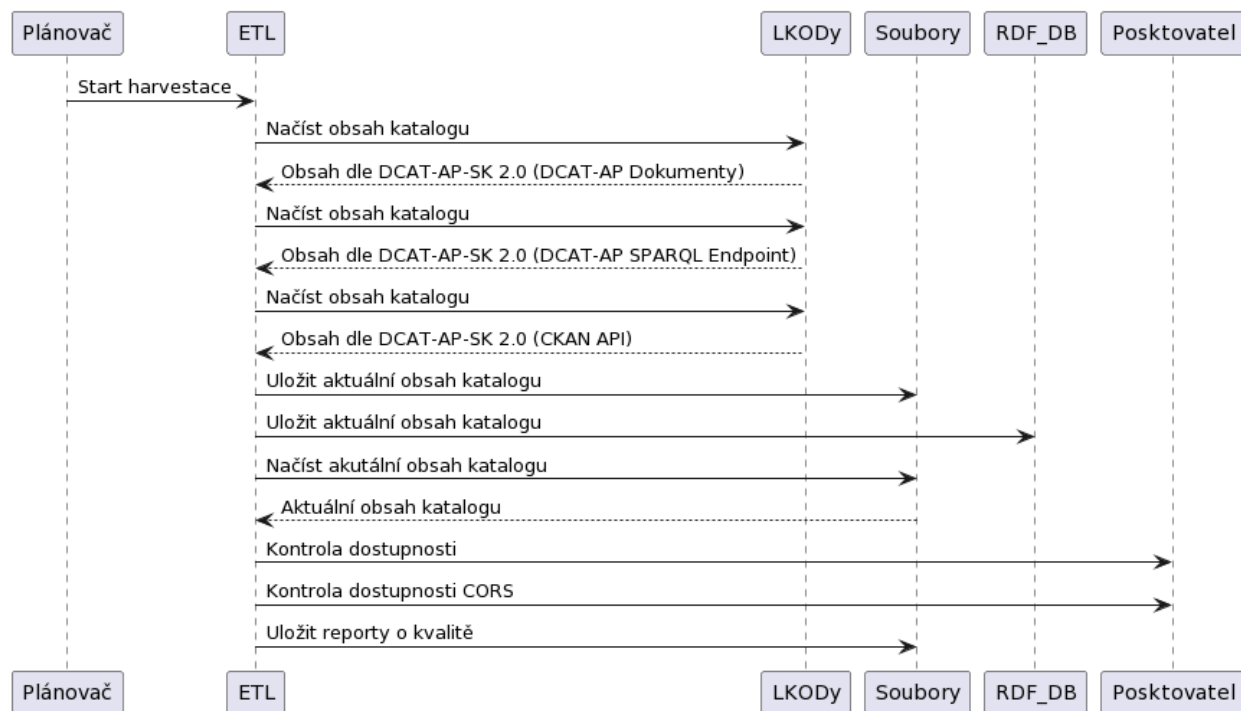
Samotný SPARQL endpoint je webová služba primárne určená pre komunikáciu s aplikáciami. Vyhodnocuje dotazy v jazyku [SPARQL](#) nad RDF úložiskom NKOD. Toto rozhranie je tiež využívané formulárom pre SPARQL dotaz na Webovej stránke.

3.5.3 LinkedPipes ETL

LinkedPipes ETL				
PIPELINES EXECUTIONS TEMPLATES PERSONALIZATION EXPORT HELP				
Label search				
Execution status				
✓	08 - Generování reportů v CSV	2023-01-25 05:45:18, 00:02:03	Size: 1135.73 mB	⌂ 🗑 ⋮
✓	07 - Kombinované indikátory kvality	2023-01-25 05:44:56, 00:00:21	Size: 280.16 mB	⌂ 🗑 ⋮
✓	06 - Kvalita metadatových záznamů v NKOD DQV	2023-01-25 05:44:38, 00:00:17	Size: 53.11 mB	⌂ 🗑 ⋮
✓	05 - Statistika dostupnosti distribucí, schémat, podmínek užití a dokumentace - CORS	2023-01-25 05:40:23, 00:04:13	Size: 646.24 mB	⌂ 🗑 ⋮
✓	04 - Statistika dostupnosti distribucí, schémat, podmínek užití a dokumentace - HEAD	2023-01-25 05:35:20, 00:05:02	Size: 604.13 mB	⌂ 🗑 ⋮
✓	03.2 - Spustit pipeline pro kvalitu	2023-01-25 05:35:18, 00:00:01	Size: 0.12 mB	⌂ 🗑 ⋮
✓	03.1 - Nahrát NKOD do GraphDB	2023-01-25 05:18:00, 00:17:18	Size: 239.64 mB	⌂ 🗑 ⋮
✓	03 - Harvestace LKOD a registrací	2023-01-25 05:17:05, 00:00:53	Size: 347.04 mB	⌂ 🗑 ⋮

Rozhraní [LinkedPipes ETL](#) není veřejně přístupné a je určeno pouze pro Administrátora NKOD pro monitoring běhu harvestace NKOD. Ten může vidět poslední tři běhy harvestace a případně vidět a diagnostikovat chyby, ke kterým může dojít.

3.6 ZÁKLADNÝ POPIS HLAVNÝCH PROCESOV V SUBSYSTEMOCH



Vyobrazen je proces tvorby obsahu NKOD realizovaný komponentou LinkedPipes ETL. Zahrnuje zpracování registračních záznamů v NKOD a následnou harvestaci registrovaných LKODů, které mohou být realizovány pomocí 3 typů rozhraní dle [DCAT-AP-SK 2.0](#). Výsledkem jsou sesbíraná metadata datových sad otevřených dat tvořící obsah NKOD. Dále jsou spuštěny kontroly kvality metadatových záznamů, dostupnosti registrovaných zdrojů a dostupnosti techniky CORS na registrovaných zdrojích. Výsledky jsou uloženy do RDF úložiště NKOD a z nich vytvořené reporty jsou zpřístupněny v podobě souborů [CSV](#) ke stažení.

3.7 POŽIADAVKY NA BEZPEČNOST A OCHRANU ÚDAJOV V IS

3.7.1 Organizačné zabezpečenie

Přístup ke cloudové infrastruktuře, přes kterou lze aplikaci konfigurovat a nasadit, má pouze Administrátor NKOD. Veřejně dostupná rozhraní jsou dostupná bez přihlášení, ale umožňují pouze čtení, nikoliv modifikaci.

3.7.2 Technické zabezpečenie

Pro zápis dat při procesu harvestace je nastaven systémový uživatel, jehož přístupové údaje jsou součástí konfigurace systému při nasazení, což se děje pomocí přístupu ke cloudové infrastruktuře, kde je aplikace nasazena. Pro přístup k webovým rozhraním aplikace je použit standardní zabezpečený protokol HTTPS.

3.7.3 Spôsob a pridelenie prístupových práv do IS

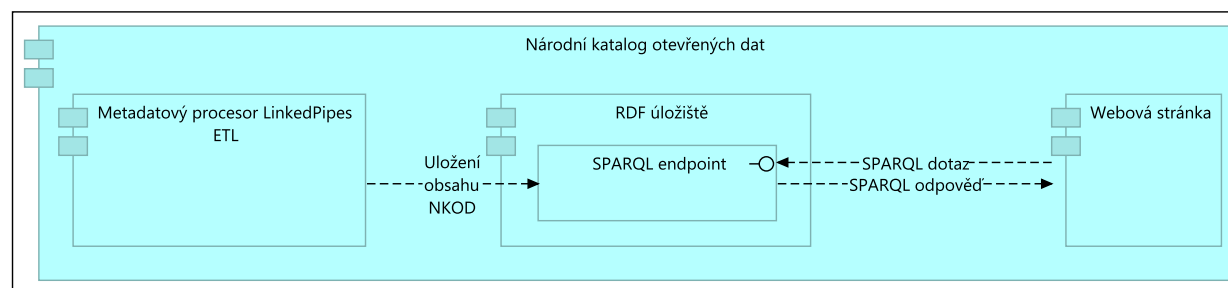
NKOD neobsahuje žádné části, ke kterým by bylo potřeba přidělovat přístup. Řízení přístup ke cloudové infrastruktuře je pak mimo rozsah aplikační příručky NKOD.

3.7.4 Ochrana údajov a úprava IS

Úložiště NKOD obsahuje data, která jsou získána ze zaslaných registračních záznamů datových sad a v procesu harvestace sesbírána z veřejně přístupných zaregistrovaných LKODů a dále výsledky měření kvality nad vzniklým obsahem NKOD. Neobsahuje tedy žádná data, která by bylo potřeba chránit z hlediska přístupu ke čtení. Je třeba tedy pouze zabezpečit, že data nikdo neautorizovaně nezmění. To je zajištěno nastavením úložiště RDF tak, že pro nepřihlášený uživatel má přístup pouze ke čtení. Systémový uživatel s právy zápisu je tvořen při nasazení aplikace dle poskytnuté konfigurace.

4. POPIS MODULOV

4.1 RDF úložiště



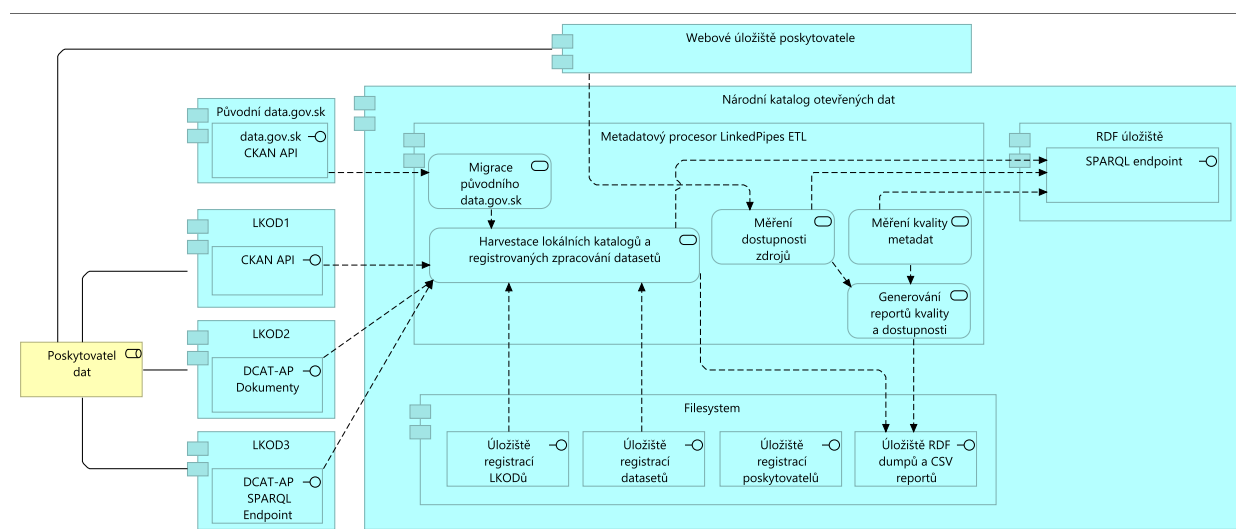
4.1.1 Popis modulu a komunikací

RDF úložiště implementované pomocí [Ontotext GraphDB Free](#) slouží k ukládání aktuálního stavu NKOD. Pro dotazování poskytuje jako rozhraní svůj SPARQL endpoint. Stejný endpoint je použit i interně pro zápis pod systémovým uživatelem.

4.1.2 Tok údajov medzi modulmi vstup – modul – výstup

Po harvestaci a měření kvality je do něj uložen aktuální stav NKOD přes rozhraní SPARQL endpoint pod systémovým uživatelem. Při příchodu SPARQL dotazu přes Webovou stránku je dotaz vyhodnocen a jsou vráceny výsledky.

4.2 Metadatový procesor - LinkedPipes ETL



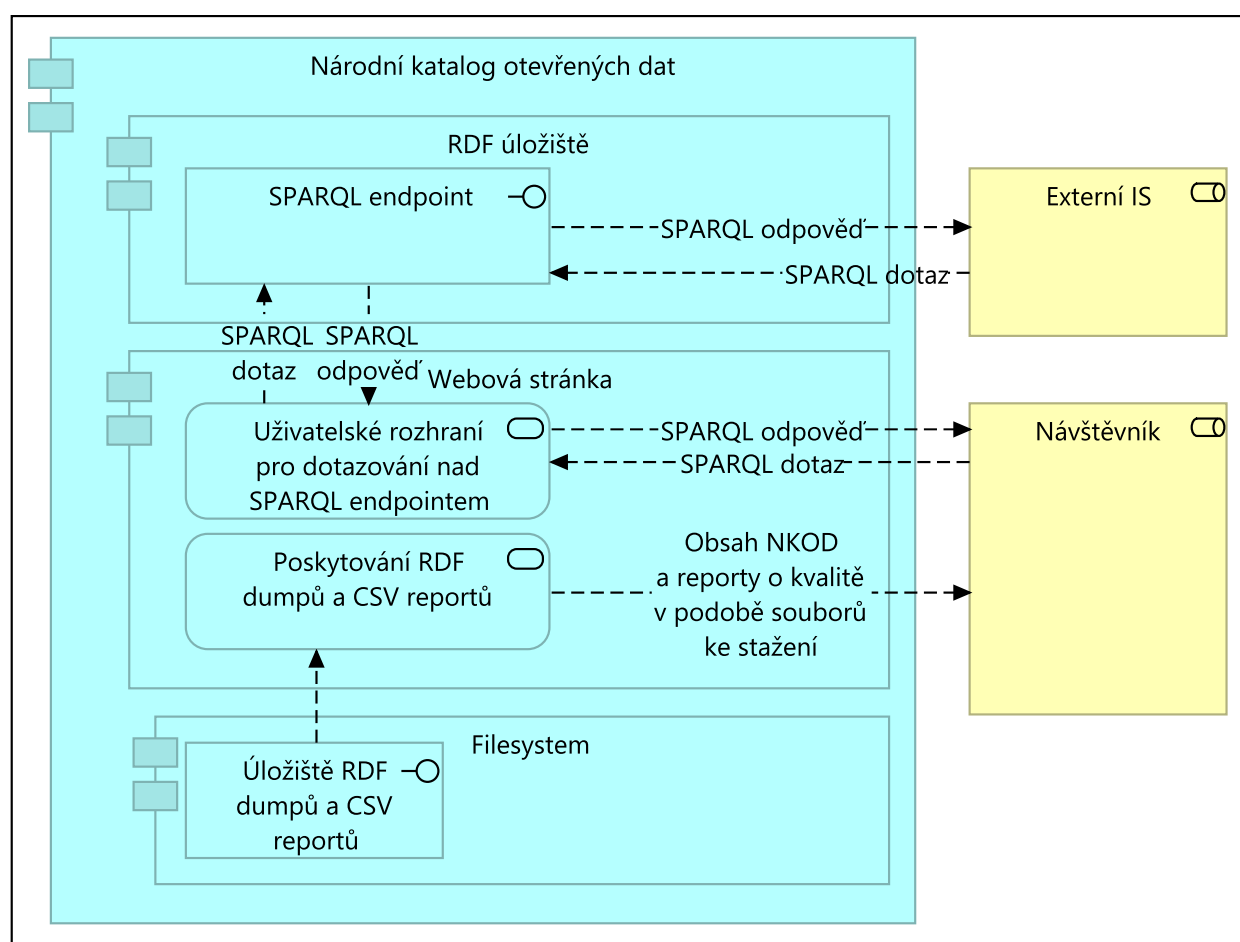
4.2.1 Popis modulu a komunikací

Metadatový procesor implementovaný pomocí [LinkedPipes ETL](#) (LP-ETL) je zodpovědný za tvorbu obsahu NKOD, měření kvality metadat a dostupnosti registrovaných datových zdrojů a techniky [CORS](#) na nich. ETL pipeline použité pro NKOD jsou publikovány a dokumentovány v [GitHub repozitáři](#). Z tohoto repozitáře se také před každým během stahují - nemá tedy smysl je modifikovat přímo v testovacím či produkčním prostředí.

4.2.2 Tok údajov medzi modulmi vstup – modul – výstup

Na vstupe pro proces tvorby obsahu NKOD LP-ETL čte registrační záznamy jednotlivých datových sad, LKODů a informace o registrovaných poskytovatelích z úložišť registrací. Na základě registrací LKODů pak čte jejich API a dostává registrační záznamy datových sad v nich obsažených. Výstupem tohoto kroku je obsah NKOD, který je uložen do RDF úložiště skrz SPARQL endpoint, a do souborového systému jako soubory ke stažení, které zpřístupňuje Webová stránka. Následuje proces měření kvality metadat, který si vystačí s právě vygenerovaným obsahem NKOD, a proces měření dostupnosti zdrojů a dostupnosti techniky [CORS](#) na nich, který navíc přistupuje ke všem registrovaným zdrojům pomocí jejich URL.

4.3 Webová stránka



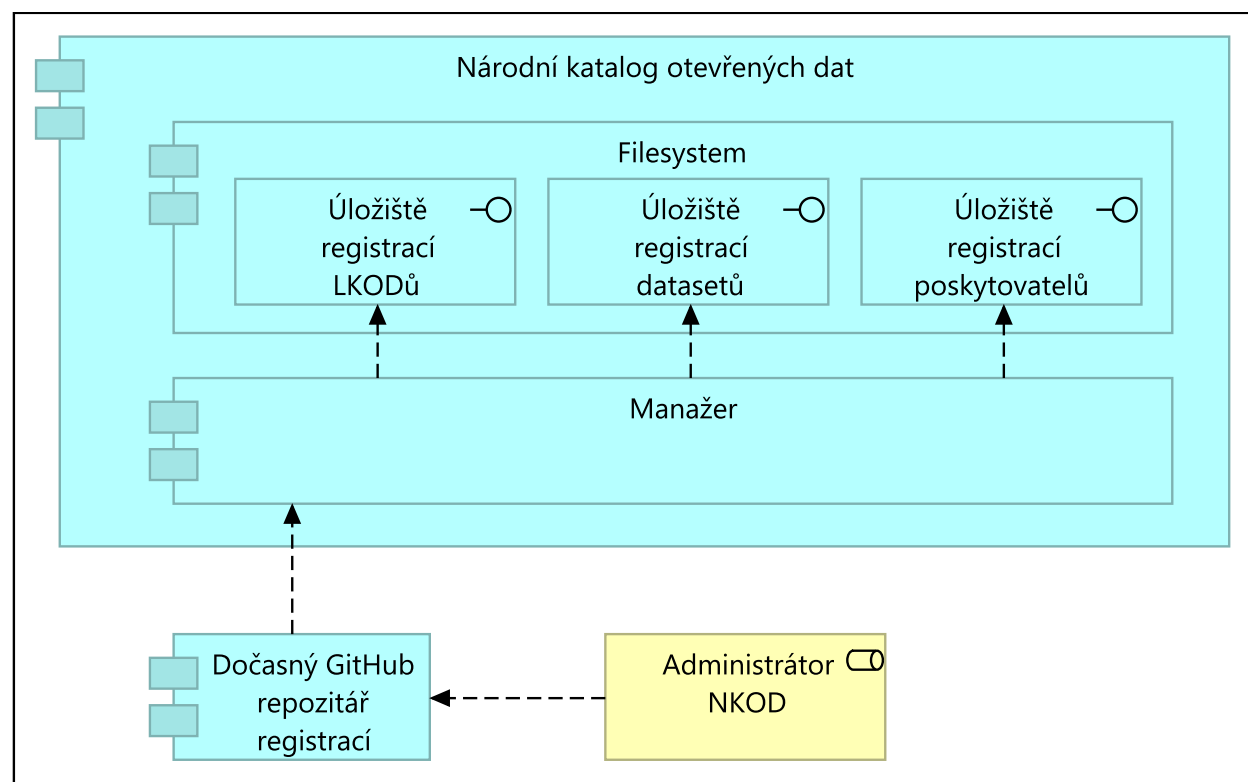
4.3.1 Popis modulu a komunikací

Webová stránka je jediné rozhraní NKOD směrem k uživateli. Obsahuje seznam předpřipravených SPARQL dotazů konfigurovaných Administrátorem NKOD a editor vlastního SPARQL dotazu [Yasgui](#). Uživatel může zvolit předpřipravený dotaz nebo vytvořit vlastní. SPARQL dotaz pak vykoná nad RDF úložištěm a vidí výsledky. Modul webové stránky pak slouží také pro zpřístupnění souborů ke stažení a technické zpřístupnění SPARQL endpointu pro externí IS formou reverse-proxy.

4.3.2 Tok údajov medzi modulmi vstup – modul – výstup

Webová stránka je vystavená na Internet. V oblasti SPARQL dotazování je vstupem SPARQL dotaz, který se předá RDF úložišti k vykonání. Vracené výsledky se zobrazí na webové stránce. Vstupem pro webovou stránku je také konfigurace předpřipravených SPARQL dotazů. V oblasti souborů ke stažení je pak vstupem soubor ve filesystému, který je zpřístupněn ke stažení přes daná URL.

4.4 Manažer



4.4.1 Popis modulu a komunikací

Manažer je modul zodpovědný za

1. Zrcadlení [GitHub repozitáře](#) s registračními záznamy datových sad, LKODů a poskytovatelů, který spravuje Administrátor NKOD, do filesystému NKOD. Tato

funkcionalita bude v budúcnosti nahradená Portálom otvorených dát (POD), ktorý bude registrační záznamy spravovať. Bude treba zabezpečiť iba prenos týchto spravovaných záznamov z POD do sústavy NKOD.

2. Aktualizáciu samotných pipeline z [GitHub repozitára](#) pred ich spustením.
3. Spustenie prvej pipeline v reťazi zabezpečujúcej celý proces harvestácie a merenia kvality.

4.4.2 Tok údajov medzi modulmi vstup – modul – výstup

Vstupom je [GitHub repozitár s registračnými záznamami dátových sad](#) a [GitHub repozitár s pipelineami NKOD pripravenými na import](#). Výstupom je klon repozitára do sústavy NKOD a aktualizované pipeline v LinkedPipes ETL.

6. NÁROKY NA POUŽÍVATEĽA

Užívatelia NKOD sú 3 druhov.

Návštevník bez znalosti SPARQL

Tento si vystačí s predpripravenými SPARQL dotazmi a teda nemusí mať žiadne špeciálne zručnosti.

Návštevník so znalosťou SPARQL

Pokiaľ chce návštevník zadať vlastný SPARQL dotaz, musí ovládať SPARQL a RDF, znáť špecifikáciu DCAT-AP-SK 2.0, a pre spracovávanie údajov z merenia kvality tiež Data Quality Vocabulary.

Externí IS

Externí IS budú prístupovať k webovej službe SPARQL endpointu. Potrebuje knihovnu na prácu so SPARQLom, znáť RDF, znáť špecifikáciu DCAT-AP-SK 2.0, a pre spracovávanie údajov z merenia kvality tiež Data Quality Vocabulary.

[Give feedback](#)