Evropská unie
Evropský sociální fond
Operační program Zaměstnanost

C3V7
Návrh zajištění podpory principů 5* otevřených dat v NKOD

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: 04: Návrhy a realizace opatření pro zvyšování povědomí o otevřených datech

Verze výstupu: 01

Návrh podpory pro 5* datové sady v NKOD

Datové sady, které jsou publikované jako 5* otevřená data mají tu vlastnost, že se s nimi dá strojově pracovat více automatizovaně nežli s otevřenými daty publikovanými na nižší úrovni otevřenosti. Tento efekt se projeví zejména v případě, že takových datových sad je registrováno více, jelikož lze hledat datové sady přímo na základě podobnosti jejich obsahu, nejen metadat. To vytváří potenciál s takto publikovanými datovými sadami pracovat již na úrovni NKOD a umožnit vyhledávání 5* datových sad na základě podobnosti jejich obsahů přímo tam. Způsob publikace 5* dat, který umožňuje se nad jejich obsahem snadno strojově dotazovat, a tedy implementovat zmíněnou funkcionalitu, je publikace pomocí SPARQL endpointu, tj. webové datové služby umožňující dotazování se nad obsahem RDF databáze obsahující 5* otevřená data.

Předpokladem pro uplatnění navrhované funkcionality tak je, že datová sada bude publikována skrze SPARQL endpoint, a tento způsob publikace bude správně registrován v NKOD jako distribuce přístupná přes datovou službu typu SPARQL endpoint. Příkladem takových metadat distribuce dle Otevřené formální normy Rozhraní katalogů otevřených dat: DCAT-AP-CZ v RDF Turtle je:

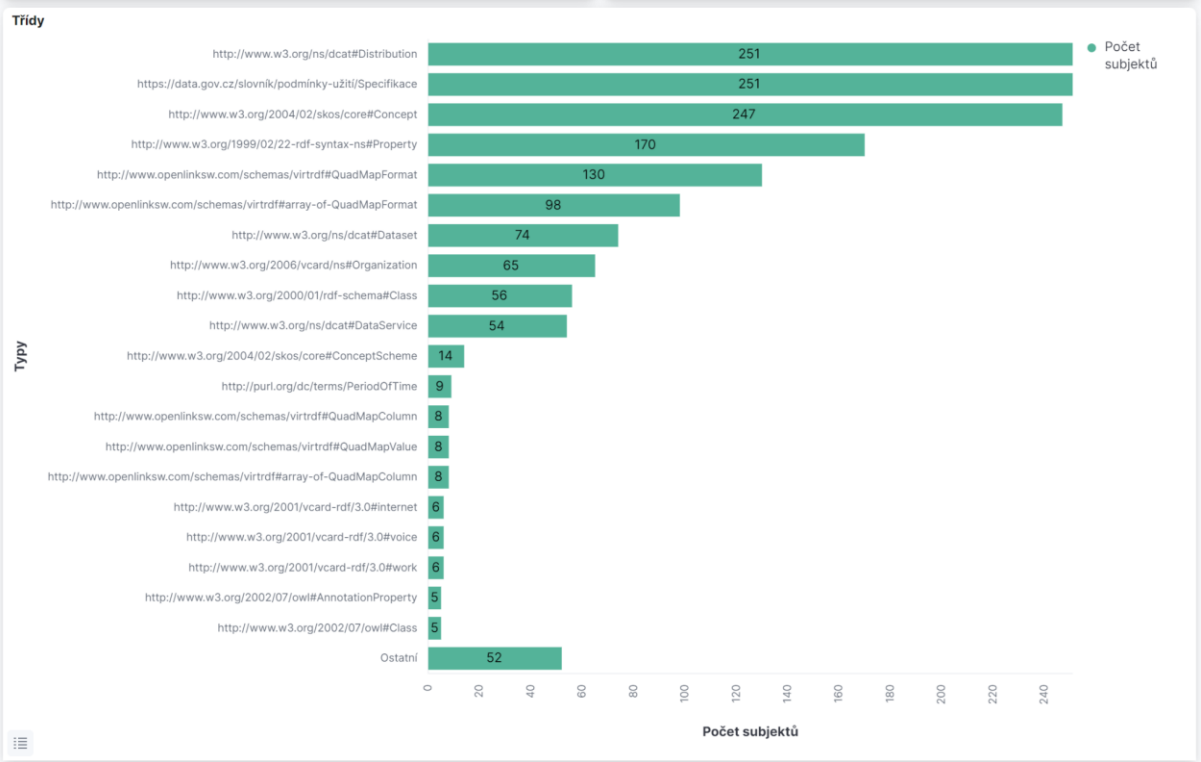
```
<https://data.gov.cz/lkod/mdcr/datové-sady/vld/distribuce/sparql> a dcat:Distribution ;
     pu:specifikace [ a pu:Specifikace ;
                pu:autorské-dílo <https://data.gov.cz/podmínky-užití/neobsahuje-autorská-díla/> ;
                pu:databáze-chráněná-zvláštními-právy <https://data.gov.cz/podmínky-užití/není-chráněna-zvláštním-právem-pořizovatele-databáze/> ;
                pu:databáze-jako-autorské-dílo <https://data.gov.cz/podmínky-užití/není-autorskoprávně-chráněnou-databází/> ;
                pu:osobní-údaje <https://data.gov.cz/podmínky-užití/neobsahuje-osobní-údaje/> ] ;
    dcat:accessURL <https://portal.cisjr.cz/sparql> ;
    dct:title "SPARQL endpoint pro jízdní řády"@cs, "SPARQL endpoint for timetables"@en ;
    dcat:accessService <https://data.gov.cz/lkod/mdcr/datové-sady/vld/služba/sparql> .

<https://data.gov.cz/lkod/mdcr/datové-sady/vld/služba/sparql> a dcat:DataService ;
    dct:title "SPARQL endpoint pro jízdní řády"@cs, "SPARQL endpoint for timetables"@en ;
    dcat:endpointURL <https://portal.cisjr.cz/sparql> ;
    dcat:endpointDescription <https://portal.cisjr.cz/sparql> ;
    dct:conformsTo <https://www.w3.org/TR/sparql11-protocol/> .
```

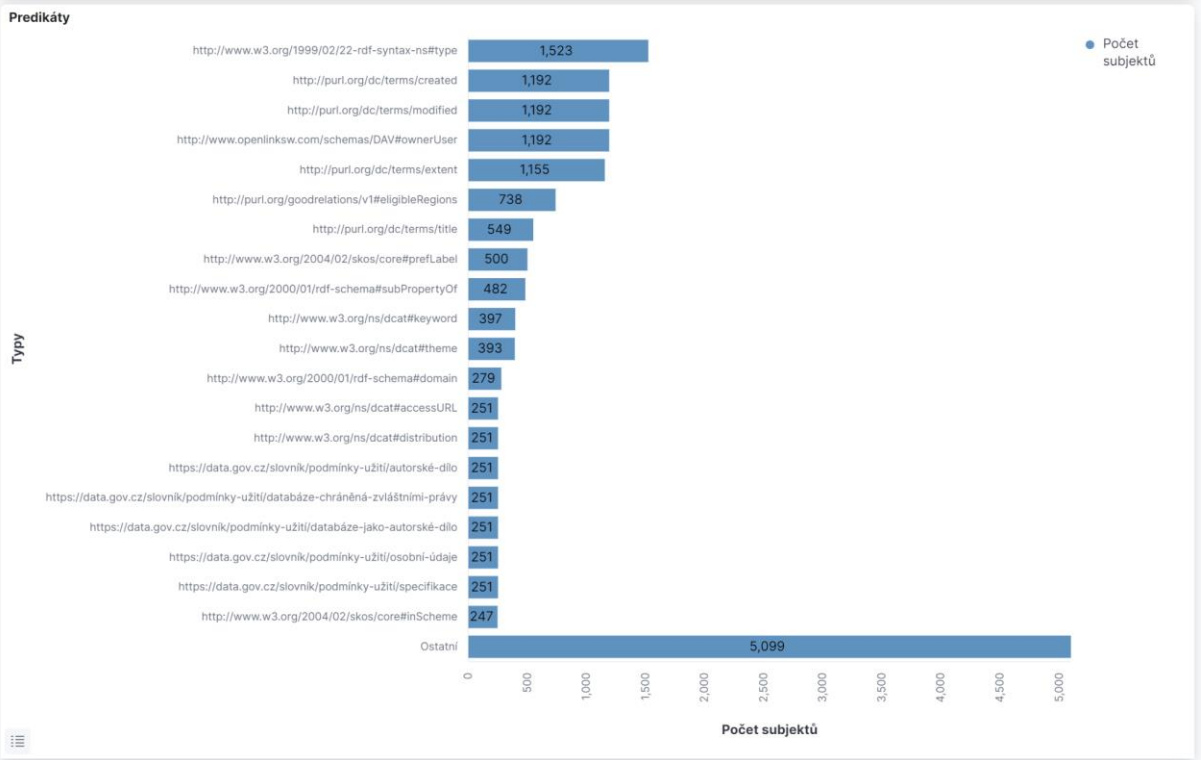
Základní vizualizace datové sady včetně zobrazení její struktury

První část navrhované funkcionality se zabývá základním zobrazením dat v nalezeném SPARQL endpointu a zobrazením jejich struktury. Struktura RDF dat ve SPARQL endpointu je dána použitými třídami a predikáty. Základní představu o datech pak poskytne přehled, kdy kromě této struktury je zobrazeno i množství, ve kterém se dané třídy a predikáty v datech vyskytují. Dohromady tak dostáváme základní charakteristiku dat v daném SPARQL endpointu.

Zobrazení nalezených tříd a jejich četností pak může vypadat následovně. Pro každou nalezenou třídu je zobrazen počet jejích instancí v endpointu:



Zobrazení nalezených predikátů a jejich četností pak může vypadat následovně. Pro každý nalezený predikát je uveden počet jeho výskytů v endpointu.



Nakonec je zobrazen seznam datových sad z Národního katalogu otevřených dat, které jsou poskytovány skrze zobrazený SPARQL endpoint. Pro každou datovou sadu je zobrazen

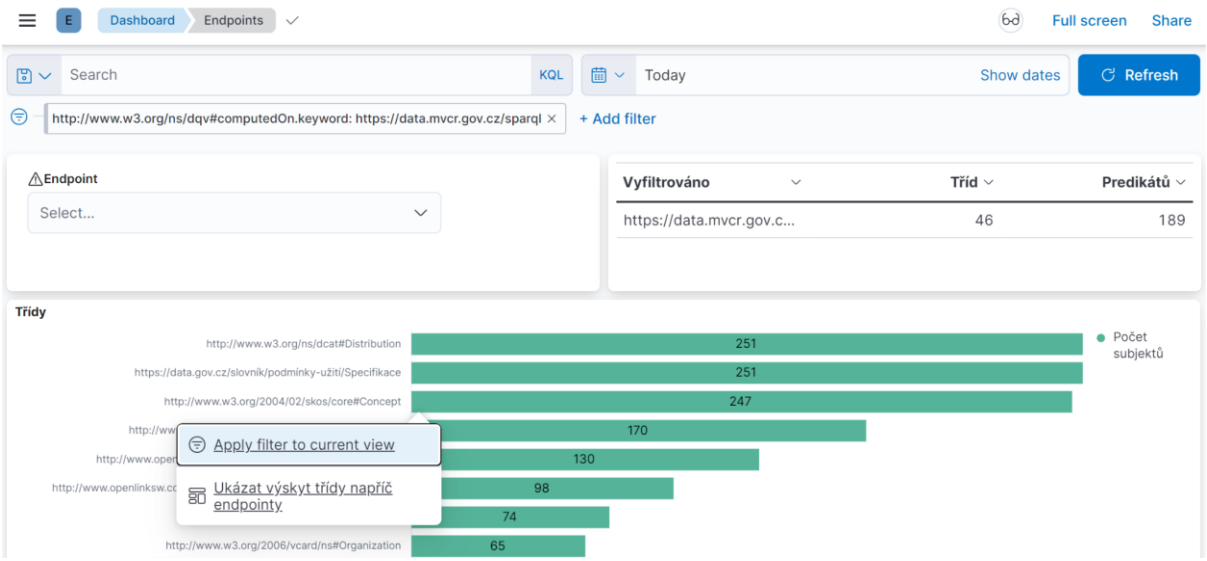
poskytovatel, název a IRI datové sady prokliknutelné do uživatelského prostředí NKOD.

Datasety

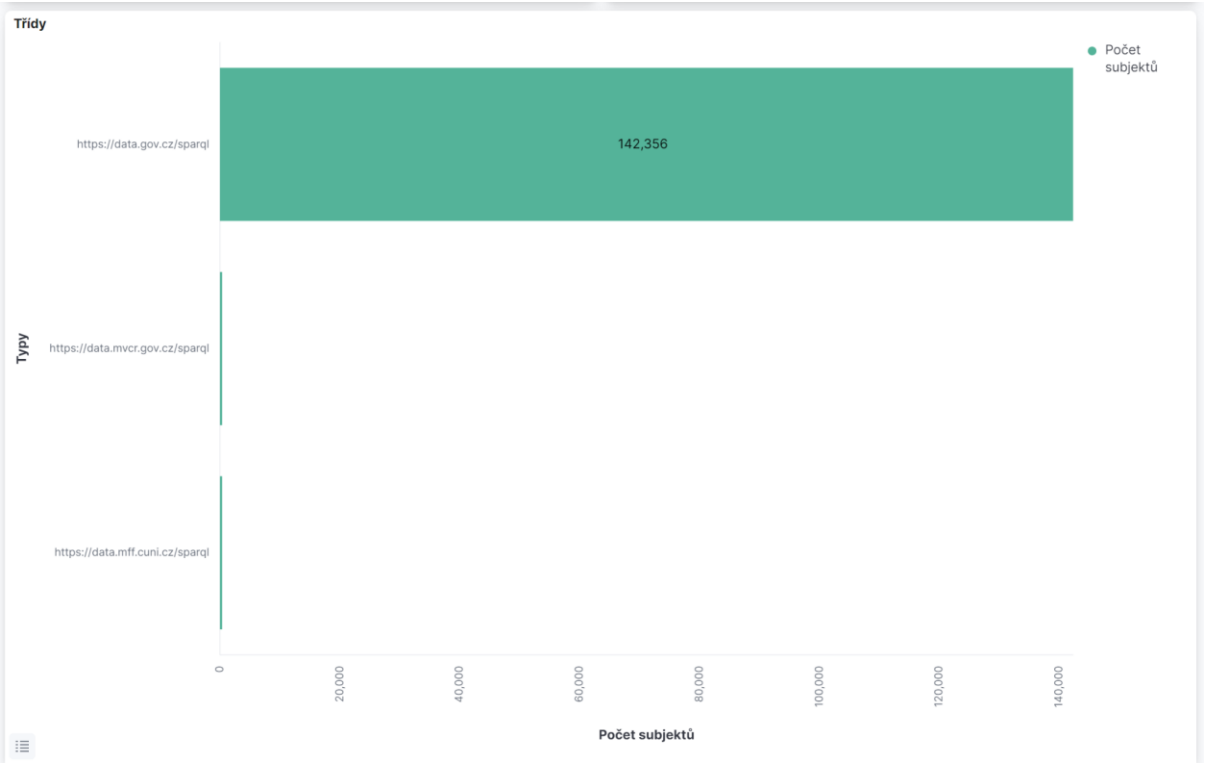
Název	Odkaz
Ministerstvo vnitra: Dostupnost CORS u registrovaných zdrojů - Národní katalog otevře...	https://data.gov.cz/datová-sada?iri=https://data.gov.cz/zdroj/datové-sady/00007064/...
Ministerstvo vnitra: Dostupnost CORS u registrovaných zdrojů - agregované indikátory ...	https://data.gov.cz/datová-sada?iri=https://data.gov.cz/zdroj/datové-sady/00007064/...
Ministerstvo vnitra: Dostupnost registrovaných zdrojů - Národní katalog otevřených dat	https://data.gov.cz/datová-sada?iri=https://data.gov.cz/zdroj/datové-sady/00007064/...
Ministerstvo vnitra: Dostupnost registrovaných zdrojů - agregované indikátory - Národ...	https://data.gov.cz/datová-sada?iri=https://data.gov.cz/zdroj/datové-sady/00007064/...
Ministerstvo vnitra: Indikátory kvality metadat - Národní katalog otevřených dat	https://data.gov.cz/datová-sada?iri=https://data.gov.cz/zdroj/datové-sady/00007064/...
Ministerstvo vnitra: Indikátory kvality metadat - agregované indikátory - Národní katalo...	https://data.gov.cz/datová-sada?iri=https://data.gov.cz/zdroj/datové-sady/00007064/...
Ministerstvo vnitra: Kompletní metadata - Národní katalog otevřených dat	https://data.gov.cz/datová-sada?iri=https://data.gov.cz/zdroj/datové-sady/00007064/...

Vyhledávání datových sad na základě souvislostí obsažených přímo v datech

Pro vyhledávání datových sad se souvisejícím obsahem bude umožněno každou třídu a predikát prokliknout akcí “Ukázat výskyt třídy napříč endpointy”:



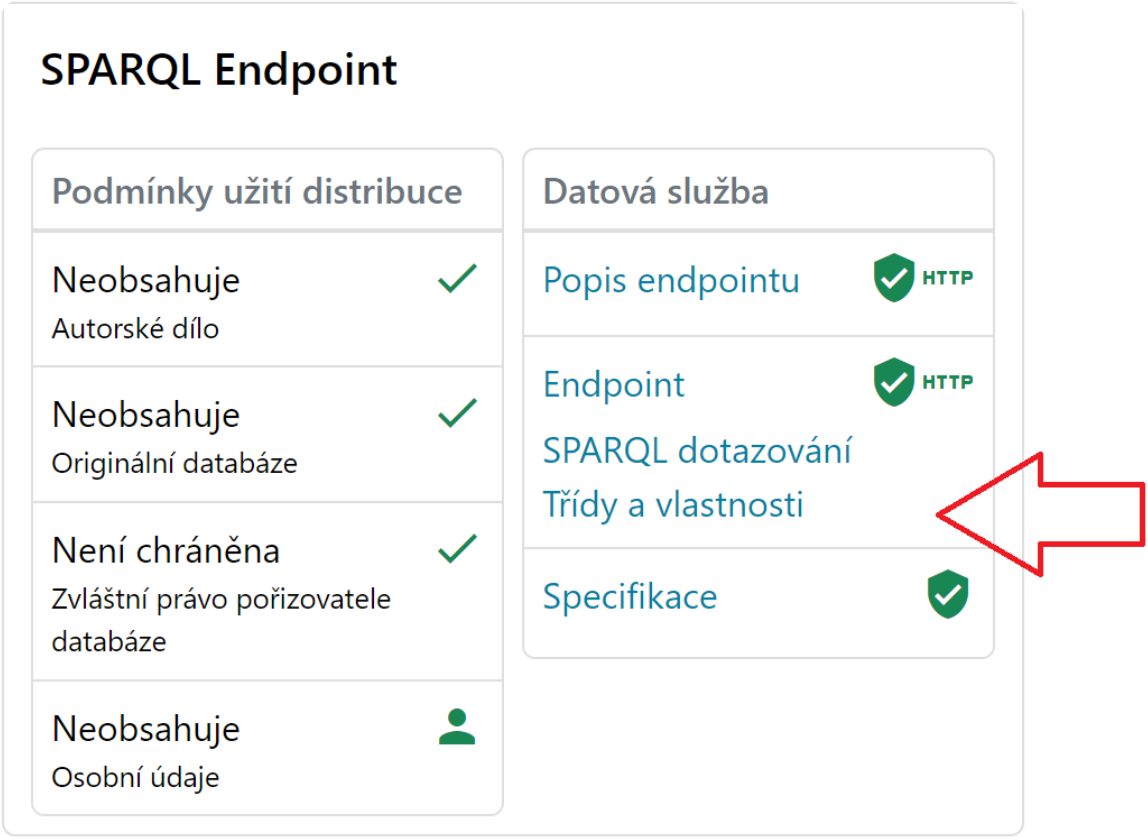
Uživatel se takto dostane na zobrazení výskytu dané třídy nebo daného predikátu ve SPARQL endpointech registrovaných v NKOD, čímž využije této souvislosti k jejich nalezení:



Po kliknutí na daný endpoint se uživatel může vrátit zpět na zobrazení tříd či predikátů v daném endpointu, kde uvidí datové sady, které jsou pomocí tohoto endpointu poskytovány.

Integrace do Národního katalogu otevřených dat

Navržená funkcionalita může být integrována do Národního katalogu otevřených dat tak, že přibude příslušný proklik ze zobrazení distribuce typu SPARQL Endpoint “Třídy a vlastnosti”, který povede na základní vizualizaci obsahu endpointu. Tato funkcionalita je implementována v přiloženém prototypu uživatelského rozhraní NKOD, v nástroji LinkedPipes DCAT-AP Viewer.



Prototyp implementace

Jako základ technického řešení byl zvolen ELK stack, tj. index Elasticsearch a vizualizační systém Kibana. Dále byl implementován software pravidelně naplňující Elasticsearch index na základě procházení registrovaných SPARQL endpointů v NKOD. Výsledkem je software Dashboard Indexer publikovaný na GitHubu jako open source. GitHub repozitář obsahuje instrukce pro instalaci pomocí systému Docker.

Systémové požadavky

Zvolené řešení vyžaduje (virtuální) stroj se 4 jádry, alespoň 8 GB paměti RAM, 20 GB místa na disku, instalovaným OS Linux (preferované je Ubuntu), systémem Docker a docker-compose. Řešení je testováno na verzi Ubuntu 22.04.

Konfigurace

Pro konfiguraci vizualizéru Kibana je třeba použít přiložený JSON soubor `export.ndjson`. Nahrání konfigurace pak probíhá přes nabídku Stack Management -> Saved Objects -> Import.
Pro konfiguraci Dashboard Indexeru je třeba použít přiložený JSON soubor `indexer-configs.conf`, který je třeba přes uživatelské rozhraní importovat.

Pro konfiguraci LinkedPipes DCAT-AP Viewer použitého jako uživatelské rozhraní NKOD je třeba přidat do `configuration.yaml` následující (a upravit URL dle místa nasazení):

```
# Template of URL to use for "class and types", with {} as placeholder for IRI.
    class-properties-url-template: "https://<MÍSTO_NASAZENÍ>/kibana/s/endpoints/app/dashboards#/view/0a73b8d0-c185-11ec-9088-bd05feec7dc9?_a=(filters:!((query:(match_phrase:('http:%2F%2Fwww.w3.org%2Fns%2Fdqv%23computedOn.keyword':'{}')))))"
```

Indexace

Indexace SPARQL endpointů registrovaných v NKOD pak probíhá v pravidelných intervalech, ve kterých je třeba naplánovat spouštění skriptu `endpoint-script.sh`, který je přidá k indexaci. To lze realizovat například systémem `cron`. Tento skript může běžet např. denně, stejně často jako probíhá harvestace NKOD. Skript je třeba upravit pro konkrétní běhové prostředí.

Testování použitelnosti

V této sekci se věnujeme testování použitelnosti prototypu pomocí standardní metodiky System Usability Scale (SUS). Testovací uživatelé nejprve prototyp vyzkoušeli na základě scénáře, a pak vyplnili dotazník založený na SUS.

Scénář

1. Otevřeme datovou sadu Národního katalogu otevřených dat, která je přístupná přes SPARQL endpoint https://data.gov.cz/sparql
2. Na SPARQL endpoint distribuci klikneme na Třídy a vlastnosti - uvidíme přehled tříd a vlastností v endpointu
3. Zaujala nás třída https://slovník.gov.cz/legislativní/sbírka/111/2009/pojem/datová-schránka - Datová schránka, o které tedy ve SPARQL endpointu NKOD jsou nějaké údaje. Chceme vědět, v jakých jiných endpointech se instance této třídy nachází.
4. Klikneme na počet instancí této třídy a zvolíme "Ukázat výskyt třídy napříč endpointy"
5. Uvidíme, že instance této třídy se kromě endpointu NKOD vyskytují také v endpointu https://rpp-opendata.egon.gov.cz/odrpp/sparql.
6. Kliknutím na počet instancí třídy v tomto endpointu a zvolením akce "Ukázat rozložení endpointu" přejdeme na zobrazení přehledu tříd a vlastností v tomto endpointu. Jedná se o endpoint Registru práv a povinností a zjistíme, jaké třídy jsou reprezentované v něm, což můžeme využít pro další, už ruční SPARQL dotazování nad ním.
7. Ve spodní části obrazovky pak vidíme datové sady, které jsou skrz tento endpoint poskytovatné, a můžeme se prokliknout zpět na jejich zobrazení v NKOD.

Výsledky dle metodiky SUS

Počet testovacích uživatelů: **3**
Výsledné skóre a jeho interpretace: **85,83 - nadprůměrné skóre**

Jednotlivé výsledky upravené dle metodiky

Už./O.	O1	O2	O3	O4	O5	O6	O7	O8	O9	O10	Skóre
1	4	3	3	4	3	4	3	4	4	4	90
2	2	4	3	4	3	4	2	4	4	1	77.5
3	3	4	4	4	3	3	4	4	3	4	90

Přílohy

1. `export.ndjson` pro konfiguraci vizualizéru Kibana
2. `dashboard-indexer.zip` je zdrojový kód softwaru dashboard-indexer zajišťujícího indexování nalezených SPARQL endpointů
3. `indexer-configs.conf` pro konfiguraci Dashboard Indexeru
4. `endpoint-script.sh` - skript pro pravidelné přidávání endpointů k indexaci
5. `dcat-ap-viewer.zip` je zdrojový kód uživatelského rozhraní NKOD s implementovanou funkcionalitou prokliku na vizualizaci vztahů datových sad pomocí nástroje LinkedPipes DCAT-AP Viewer.