

Národní repozitářová platforma

David Antoš

CESNET

2. října 2025



Spolufinancováno
Evropskou unií



MINISTERSTVO ŠKOLSTVÍ,
MLÁDEŽE A TĚLOVÝCHOVY



Registrační číslo projektu NRP

CZ.02.01.01/00/23_014/0008787

O čem to bude

- ukládání dat v kontextu e-INFRA CZ
- Národní repozitářová platforma a její služby
- NRP pro uživatele
 - vyhledávání dat
 - výběr repozitáře pro ukládání dat
 - data z repozitářů a výpočty
- stav a plány

Datová úložiště a NRP

- různá úložiště v infrastruktuře
 - úložiště výpočetních systémů
 - home/scratch
 - většinou souborové systémy
 - obecná úložiště
 - „objektová“, cca 136 PB fyzické kapacity
 - S3, RBD, CephFS
 - přímý přístup k úložišti
 - repozitáře v Národní repozitářové platformě
 - naše hlavní téma

Proč repozitáře pro vědecká data

- tradičně jsme mluvili o modelech užití
 - zálohy, archivace, sdílení dat
 - nové potřeby
 - uchování dat s ohledem na reprodukovatelnost výsledků
 - data FAIRness
- ~> posouvá se role obecných úložišť
- ~> potřebujeme specializované služby

Použití obecných úložišť

- krátkodobé a střednědobé uložení vědeckých dat, např.
 - velká výpočetně zpracovávaná data
 - velká dočasná data, která nedávají smysl v repozitáři
 - “to be FAIRified”
- pro pokrytí doby, než bude NRP plně k dispozici
- nejsou ale přijatelné stovky TB ve formě
zaloha2023/zaloha2020/zaloha-stars/nevim/zbytky
- funkci archivu přebere NRP
 - „nějaké soubory ve složkách, nevím, co tam je“ nejsou archiv

Co je repozitář? I

- v první řadě úložiště dat s (obsahovými) metadaty
- podporuje FAIR principy
 - naležitelnost/dostupnost/interoperabilita/znovupoužitelnost
- obvykle má webové rozhraní a API pro strojový přístup
- má odpovědnost za uložená data
- potenciálně certifikovatelný na Core Trust Seal
 - viz např. <https://www.clarin.eu/content/checklist-clarin-b-centres>

Co je repozitář? II

- datová sada: soubory s daty a popisná metadata
- repozitář musí obsahovat „citovatelné datové sady“
 - zajišťuje jejich neměnnost(*) a dlouhodobé(**) uchování
 - kontrolní otázka k (*): dataset má přidělené DOI, plyne z toho, že jsou to přesně stejná data, jaká jsou popsána v článku, který na toto DOI odkazuje?
- repozitář je technický, personální a procesní systém pro dlouhodobé uchování a zpřístupnění citovatelných digitálních objektů

Co je NRP I

- Národní repozitářová platforma
 - distribuovaný, multi-tenant systém pro vytváření repozitářů
 - distribuovaný: geograficky
 - multi-tenant: řada speciálních repozitářů (nikoli jeden velký)
 - vytváření repozitářů: dovoluje vytváření repozitáře z připravených komponent „jako službu“

Co je NRP II

- typy uživatelů
 - koncový uživatel repozitáře
 - hledá data, stahuje, ukládá data
 - typicky jej zajímá konkrétní repozitář
 - tomu se teď budeme věnovat
 - administrátor repozitáře
 - vytváří a spravuje repozitář pro konkrétní téma, obor nebo instituci
 - pro zájemce: <https://www.eosc.cz/sluzby/ukladani/repozitare-v-nrp/zalozeni-repozitare-v-nrp>

Co NRP není

- NRP není prostředí pro výpočty
 - na to máte Metacentrum a superpočítačové centrum
- NRP není jen pro „otevřená data“
 - FAIR data: musí být jasné podmínky přístupu (a to nemusí být „veřejně všem“)
 - „data v repozitáři“ $\neq$ „zveřejněná“
 - rozsah přístupu určuje repozitář a uživatel
- NRP nevynucuje podmínky repozitáře
 - politiku repozitáře určuje jeho správce

NRP pro koncového uživatele

- hledání dat
 - hledat datové sady lze v Národním metadatovém adresáři
 - NMA (<https://nma.eosc.cz/>)
 - agregátor metadat
 - dovede k repozitáři s daty
 - repozitář s daty umí jemnější vyhledávání
- ukládání dat
 - přes webové rozhraní nebo API
 - vyplnění metadat je dobré automatizovat
 - detaily závisí na vlastnostech konkrétního repozitáře

Jak si zvolit vhodný repozitář pro svá data

- dle zvyklostí vědecké disciplíny
- obecná vodítka
 - první volba: známý komunitou uznávaný repozitář
 - (vágní) koncept „důvěryhodného repozitáře“
 - pokud není, tak institucionální nebo catch-all
 - pracovní skupiny EOSC budou ustavovat své oborové repozitáře
- role catch-all repozitáře
 - „repozitář poslední volby“
 - pro malé komunity, kterým se nevyplatí dedikovaný repozitář

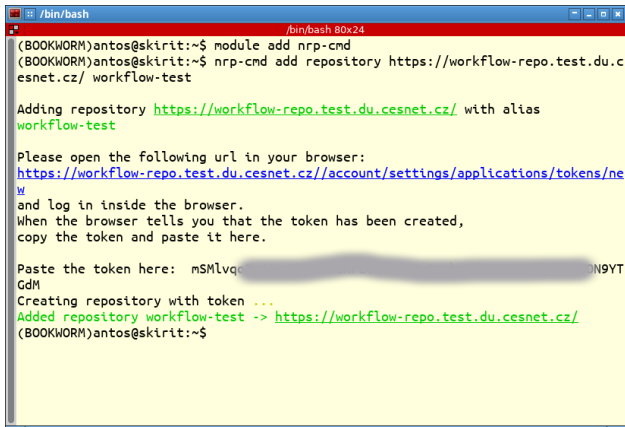
Ukládání do repozitáře

- kdy ukládat data do repozitáře
- je třeba vyvážit
 - „co nejdříve“
 - data už se nemají měnit
 - data mají hodnotu do budoucna
 - chceme umět prokázat jejich původ
 - chceme s nimi pevně svázat metadata
 - „ale ne dřív“
 - např. pokud většinu z nich zahodíme
 - nebo se budou výrazně měnit
 - ekonomický aspekt

Repozitář a výpočetní prostředí

- repozitář řídí přístup k datům
 - tj. data se nestahují „z S3“, ale „z repozitáře“
 - lze technicky optimalizovat
- repozitář je úložný systém
- nástroje pro stažení z repozitáře
 - API repozitáře
 - klientská knihovna
 - např. použitelná v Jupyter Notebooks
 - integrace do Galaxy
 - CLI nástroje

nrp-cmd I



```
/bin/bash
/bin/bash 80x24
(BOOKWORM)antos@skirit:~$ module add nrp-cmd
(BOOKWORM)antos@skirit:~$ nrp-cmd add repository https://workflow-repo.test.du.cesnet.cz/ workflow-test

Adding repository https://workflow-repo.test.du.cesnet.cz/ with alias workflow-test

Please open the following url in your browser:
https://workflow-repo.test.du.cesnet.cz//account/settings/applications/tokens/new
and log in inside the browser.
When the browser tells you that the token has been created,
copy the token and paste it here.

Paste the token here: mSMLvq...ON9YT
GdM
Creating repository with token ...
Added repository workflow-test -> https://workflow-repo.test.du.cesnet.cz/
(BOOKWORM)antos@skirit:~$
```

nrp-cmd II

← → ↺ 🏠 workflow-repo.test.du: cesnet.cz/account/settings/applica: 📄 🗖️ ☆ 🔍 Search ⬇️ 📁 🗨️ 🌈 ⚙️ 🗖️ 👤 📺 📄 ☰

INVENIO 🔍 Komunity Nástěnka 👤 david.antos@cesnet.cz ▾

Můj účet

Nastavení

- 👤 Profil
- 🔒 Zabezpečení
- 🔗 Linked accounts
- 🔗 Aplikace**

Vývojářské aplikace ➕ Nová aplikace

Prozatím jste nezaregistrovali žádné aplikace. Kliknutím na tlačítko „Nová aplikace“ vytvoříte aplikaci s přístupem k API Data Harbour.

Osobní přístupové tokeny ➕ Nový token

Toto jsou osobní tokeny pro přístup k API Data Harbour:

[workflow-test-metacentrum](#)

Povolené aplikace

Dosud jste žádné externí aplikaci neudělili přístup ke svému účtu.

nrp-cmd III

```
> nrp-cmd create record --community generic ./dataset.json hrensko.tar.gz '{"title": "Hrensko"}' --set var  
Record 4dyah-5rz49  
  
alternate_identifiers {  
  "identifier": "10.1235/alternate.doi",  
  "scheme": "doi"  
}  
  
alternate_titles {  
  "title": {  
    "lang": "en",  
    "value": "Vegetation indices in the 2022 fire-affected Hrensko area"  
  },  
  "titleType": {  
    "id": "AlternativeTitle",  
    "@v": "7b0b0f14-6a60-4fb2-9e12-797c27296aee::1",  
    "title": {  
      "cs": "Alternativní název",  
      "en": "Alternative Title"  
    }  
  },  
  {  
    "title": {  
      "lang": "cs",  
      "value": "Vegetační indexy v oblasti Hrenska, postížené lesním požárem r. 2022"  
    },  
    "titleType": {  
      "id": "AlternativeTitle",  
      "@v": "7b0b0f14-6a60-4fb2-9e12-797c27296aee::1",  
      "title": {  
        "cs": "Alternativní název",  
        "en": "Alternative Title"  
      }  
    }  
  }  
}  
  
contributors {  
  "person_or_org": {  
    "type": "personal",  
    "name": "Rozsáhlá Rada"
```

nrp-cmd IV

```
> nrp-cmd get record doi:10.23644/hrensko-jrje-ra24

Record 7nb94-p3435

alternate_identifiers {
  "identifier": "10.1235/alternate.doi",
  "scheme": "doi"
}
alternate_titles {
  "title": {
    "lang": "en",
    "value": "Vegetation indices in the 2022 fire-affected Hrensko area"
  },
  "titleType": {
    "id": "AlternativeTitle",
    "@v": "7b0b0f14-6a60-4fb2-9e12-797c27296aee::1",
    "title": {
      "cs": "Alternativní název",
      "en": "Alternative Title"
    }
  }
},
{
  "title": {
    "lang": "cs",
    "value": "Vegetační indexy v oblasti Hřenska, postižené lesním požárem r. 2022"
  },
  "titleType": {
    "id": "AlternativeTitle",
    "@v": "7b0b0f14-6a60-4fb2-9e12-797c27296aee::1",
    "title": {
      "cs": "Alternativní název",
      "en": "Alternative Title"
    }
  }
}
}
```

Co je k dispozici už teď

- <https://datarepo.eosc.cz/>
 - catch-all repozitář
- pilotní repozitáře
 - např. <https://mbdb-data.org>
- z projektu OSII vznikne řada repozitářů během nejbližšího roku
 - „repozitář jako služba“
- <https://nma.eosc.cz/>
 - Národní metadatový adresář
 - primární vyhledávací bod pro datové sady

Dokumentace a podpora

- <https://docs.du.cesnet.cz/>
 - support@cesnet.cz
- <https://docs.nrp.eosc.cz/>
 - přehled repozitářů, rozcestník na technické dokumentace
 - silně ve výstavbě
- <https://www.eosc.cz/>
 - projekty/milníky/pracovní skupiny/politika/školení/akce

Shrnutí

- předpokládáme vyšší zapojení velkých úložných systémů do výpočetních úloh
 - pro vědecká data
- NRP
 - přebírá funkci archivu
 - ale je to živé úložiště pro rozsáhlá data
 - repozitáře dle potřeb uživatelských komunit
 - poskytuje infrastrukturní služby
 - uživatel a repozitář řídí životní cyklus dat

This Is Not a Drill



Spolufinancováno
Evropskou unií



Registrační číslo projektu NRP

CZ.02.01.01/00/23_014/0008787