National Czech Programme | eosc

National Repository Platform: Status and Plans

David Antoš
CESNET
26 November 2024

Spolufinancováno Evropskou unií
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
cesnet
Registrační číslo projektu NRP
CZ.02.01.01/00/23_014/0008787

Overview

- data storage in context of e-INFRA CZ
- National Repository Platform and its services
- NRP for users
  - searching for data
  - choice of repository for data deposition
- NRP for repository administrators
  - establishing a repository
- data and document repositories
- state of NRP development and plans

Data Storage and NRP

- data services in the infrastructure
  - storage systems coupled with computational resources
    - home/scratch
    - understood and described as a part of computation
    - almost exclusively classic file systems
    - out of scope of this presentation
  - general purpose data storage
    - A.K.A. “object storage”
    - and special data services
  - repositories in the National repository platform
    - main topic of this talk

Current General Storage Facilities

- operating object storage clusters CL1–CL5
  - S3, RBD, CephFS
  - direct access to storage
- for generic data
  - approx. 121 PB physical capacity
  - planned gradual modernisation of the physical infrastructure
  - but no significant capacity increase
- higher-level services:
  - FileSender—`https://filesender.cesnet.cz`
    - temporary storage for file transfer
  - ownCloud—`https://owncloud.cesnet.cz`
    - sync'n'share

Supporting Scientific Data

- we used to identify basic use cases for storage
  - backup, archiving, data sharing
- we need to support new requirements for
  - data retention
  - data FAIRness

⇝ the role of generic data storage is shifting
⇝ we add specialised services

Generic Storage Use Cases

- role of generic/unstructured data storage in the infrastructure:
  - short- and mid-term storage of scientific data, e.g.
    - used for computation tasks and exceeding standard disk arrays in size
    - “to be FAIRified”
    - too big to be ever directly stored in repositories
    - shared among users
  - buffer until the National repository platform is ready
    - but not a final resting place of unstructured mess
- archiving function will be gradually taken over by repositories
  - “files in folders” are not an archive, though

What is the NRP I

- National Repository Platform
  - distributed, multi-tenant system for repository instantiation
    - distributed: geographically
    - multi-tenant: not a single big repository, but many tailored repositories
    - repository instantiation: able to build a repository out of pre-fabricated components “as a service”

What is the NRP II

- types of users:
  - repository end-user
    - searches for data, downloads, deposits data
    - is typically interested in a particular repository
  - repository administrator/curator
    - establishes and operates a repository for a particular topic: scientific community or for an institution
    - similar to a Virtual Organisation admin
    - negotiates properties of the repository with the infrastructure
    - manages user groups and deposited data

What the NRP Is Not

- not a processing/computation environment
  - computation resources are elsewhere in e-INFRA CZ
- not an environment just for “open data”
  - just for FAIR data
  - i.e. data in the repository ≠ published
  - publication is always controlled by users
- not an environment to run generic applications
  - unless they are characterised as repositories
  - and/or directly related supporting systems

What is a Repository

- system for storing data with extensive descriptive metadata
- supporting FAIR principles
- web interface and API for machine access
- bearing responsibility for stored data (esp. integrity)
- potentially CTS certifiable
  - cf. `https://www.clarin.eu/content/checklist-clarin-b-centres`
- must contain “citable data sets”
  - ensuring their immutability and long-term retention
- a repository is a technical, personal, and process solution for long-term storage and publication of citable digital objects

Implementations of NRP Repositories

- CESNET Invenio (CESNET)
- CLARIN DSpace (Charles University)
- ASEP/ARL (Czech Academy of Sciences)
- alternative implementations possible
  - again, must be a repository
  - some piloted in the project
  - to run those, the infrastructure offers S3 storage and Kubernetes containerisation as a service

Layers of the NRP

- end users of a repository
  - are using
- a repository
  - operated by repository admins and curators
  - which is based on a
- repository implementation
  - operated by system admins and has a development team
  - storing data+running in
- S3+Kubernetes
  - operated by system admins

NRP for End Users

- to search for data sets:
  - start in the National Metadata Directory
  - NMA (`https://nma.eosc.cz/`)
    - metadata aggregator
    - primary search interface for all datasets
  - more detailed search supported by particular repositories
- data deposition
  - highly dependent on metadata models and procedures of a repository

Repository in Scientific Workflows: Deposition

- when should data be stored into a repository
- TL;DR: it depends
- aspects to balance
  - as soon as possible
    - when the data doesn’t change (any more)
    - when you expect the data to be of future value
    - early deposition makes tracking metadata easier
    - and improves provenance tracking
  - but not sooner
    - e.g. big primary data that is strongly decimated
    - e.g. majority of primary data is wrong anyway

Repository in Scientific Workflows: Accessing the Data

- using data from a repository
- data is typically identified by a persistent identifier
- there will be tools to download datasets resolving PIDs
  - staging to computations, …
- staging data to computations is similar to standard object storage
- note: repository is responsible for authorisation decisions when accessing data
- repository *is* as a storage system
  - users don’t access the underlying storage directly
  - can be technically optimised, but it doesn’t change the concept

How to Choose a Suitable Repository

- community specific
- general guidelines
  - first choice: well-established community-recognised repository
    - (still fuzzy) concept of “trustworthy repositories”
  - if not available, institutional or catch-all repository
  - EOSC CZ Working Groups should describe specifics for their research areas
- role of the catch-all repository
  - (temporarily) cover communities before their repositories are established
  - for users and groups without better options

NRP for Repository Administrators I

- if you represent a user community intending to establish a repository
- `https://www.eosc.cz/projekty/narodni-repozitarova-platforma-pro-vyzkumna-data-os-i-nrp/nrp` (in Czech)
- covering
  - necessary roles and personnel to run a repository
  - choice of tools
  - minimal integration requirements
  - roles and responsibilities of various actors
  - etc.
- contacts for consultations included

NRP for Repository Administrators II

- repository administrator needs to set up:
  - repository administrator and curator
  - metadata profiles, available licenses
  - deposition/data access workflows
  - roles of user groups in the repository
  - end user documentation
  - establish repository policies
  - first level user support
  - register the repository, metadata exports
  - additional items when running non-standard repository implementation
- infrastructure prepares standardised components
- but primary responsibility is on the repository administrator

What is Available Right Now

- `https://data.narodni-repozitar.cz/`
  - catch-all repository
  - for long-tail, for groups that don't have a repository yet
  - small storage capacity so far
- pilot repositories as NRP instances appearing
- `https://nma.eosc.cz/`
  - National Metadata Directory
  - hardware procured
  - service running

What about documents?

- document repository will be available mid 2025
  - as a part of the “national repository”
  - catch-all
  - National Library of Technology within NCIP VaVaI project
- it will replace National Repository of Grey Literature (NUŠL)
  - `https://nusl.cz/`
  - mix of documents and harvested metadata

National Data Infrastructure

- main components
  - National Repository Platform
    - expected capacity in 2028: 250 PB physical/50 PB user
  - National Metadata Directory (NMA for NM„Adresář“ in Czech)
    - metadata aggregator
    - search capabilities for end users
  - National Repository Catalogue
    - listing of available repositories
    - including metadata schemas
  - generic storage
  - supporting systems

Main Milestones of the NRP

- we are currently getting ready for groups intending to establish repositories
  - don’t wait for NRP hardware!
  - hardware will enable storing big data
  - processes and configurations can be prepared in advance
- “Repository as a Service:” 2025
- first dedicated hardware resources for the NRP: ~~Q2~~ Q3-Q4/2025
  - delay due to public tenders awaiting formal steps
  - catch-all repository, other repositories will move there
- continuous integration of project results into the infrastructure
- full capacity of the infrastructure: 2028

Where to Seek Documentation and Support

- `https://du.cesnet.cz/`
  - `support@cesnet.cz`
- `https://data.narodni-repozitar.cz/`
  - generic catch-all repository
    - generic metadata model, DOI
  - direct link to the documentation on the main page
  - `support@narodni-repozitar.cz`
- `https://www.eocs.cz/`

Summary

- shifting the role of generic storage facilities in the infrastructure
  - emphasis on working with scientific data
  - archival functionality ⇝ NRP
- National Repository Platform will become a pillar of the National Data Infrastructure
  - repositories tailored to the needs of user communities
  - providing infrastructure services
  - while retaining control in user's hands

National Czech Programme | eosc

In Case of Evacuation, Use All Available Emergency Exits

Spolufinancováno Evropskou unií

MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

cesnet

Registrační číslo projektu NRP
CZ.02.01.01/00/23_014/0008787