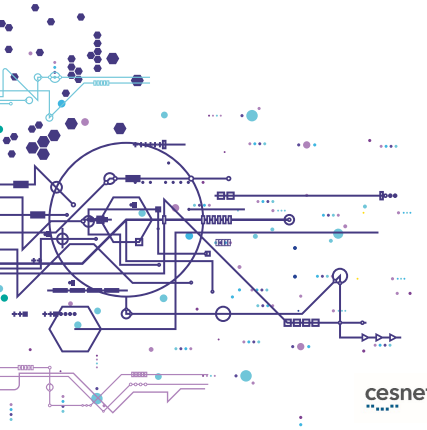
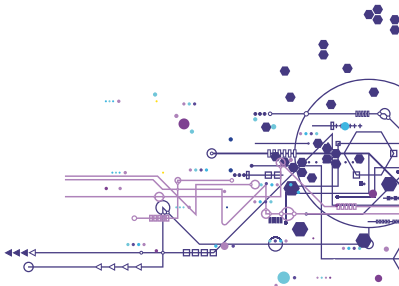
HANDLING SCIENTIFIC DATA IN THE NATIONAL DATA INFRASTRUCTURE

David Antoš

29 April 2024





Overview

- generic data storage services in the infrastructure
  - their role in the infrastructure shifting
- National repository platform (NRP)
  - architecture
  - state of its development
  - 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
    - 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

Current Storage Facilities and Their Renewal

- we still operate
  - hierarchical system with a tape library
  - disk array
  - will be decommissioned
- operating object storage clusters CL1–CL5
  - approx. 121 PB physical capacity
  - planned „renewal“ CL1 (2024), CL2 (2025), CL3 (2026)
    - gradual replacement of the physical infrastructure
    - should go unnoticed by the users

Data Storage Services

- built upon Ceph
  - user interfaces S3, RBD
  - possibly CephFS if absolutely necessary
- note: MetaCentrum and IT4I still operating classic file-system access
- FileSender—`https://filesender.cesnet.cz`
  - temporary storage for file transfer
- ownCloud—`https://owncloud.cesnet.cz`
  - sync'n'share will be available
  - we are considering to switch to an alternative
  - sustainable for conservative growth of the user community

Supporting Scientific Data

- we used to identify basic use cases:
  - backup, archives, data sharing
- new requirements for
  - data retention
  - data FAIRness
- use cases of big storage facilities must make sense wrt. the National Repository Platform

⇝ redefining the role of generic data storage

Storage Use Cases I

- role of generic/unstructured data storage in the infrastructure:
  - data storage facilities for big scientific data:
    - used for computation tasks and exceeding standard disk arrays in size
    - “to be FAIRified”
    - too big to be stored directly into repositories
    - shared among users
  - buffer for users until the National repository platform is ready
    - but not as a final resting place of an unstructured mess

Storage Use Cases II

- note: everything is an evolution, not a revolution
- majority of the data must be scientific
- archives: this function will be taken over by repositories
  - “files with no metadata in folders” can no longer be considered an archive
  - user transition to repositories will take years
  - education of user communities is the key
- backups
  - usage patterns must fit into the infrastructure
  - eg. temporary backups of scientific measurements may be OK (depending on the whole workflow)
- sharing scientific data
  - yes, using our excellent AAI tools

NRP

- National Repository Platform
  - distributed, multi-tenant system for repository instantiation
- types of users:
  - repository end-user
    - searches data, downloads, deposits data
    - is typically interested in a particular repository
  - repository administrator/curator
    - needs 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
- note: there are no computational resources in the NRP, those are in e-INFRA CZ

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
- potentially CTS certifiable
  - cf. `https://www.clarin.eu/content/checklist-clarin-b-centres`
- it should contain “citable data sets”
  - ensuring their immutability
- a repository is a technical, personal, and process solution for long-term storage and publication of citable digital objects
- not just an archive

Repository in a Scientific Workflow

- when should data be stored into a repository
- TL;DR: it depends
- aspects to balance
  - as soon as possible
    - when the data is fixed
    - soon deposition makes tracking metadata easier
  - but not sooner
    - e.g. big primary data that is strongly decimated
- staging data to computations from a repository is similar to standard object storage
- data in the repository ≠ published!
  - publication always under user's control

Implementations of NRP Repositories

- CESNET Invenio (CESNET)
- CLARIN DSpace (Charles University)
- ASEP/ARL (Academy of Sciences)
- others possible
  - they have to be “repositories”
  - piloted in the project
  - the infrastructure offers S3 storage and Kubernetes containerisation as a service

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

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
  - service running&harvesting
  - hardware procured

Main Milestones of the NRP

- Installation of S3 and Kubernetes to run repositories: mid 2025
- “Repository as a Service:” 2025
- First dedicated hardware resources for the NRP: Q2/2025
  - Catch-all repository, other repositories moved there
- 3 geographic locations: Q4/2025
- Continuous integration of project results into the infrastructure
- Full capacity of the infrastructure: 2028

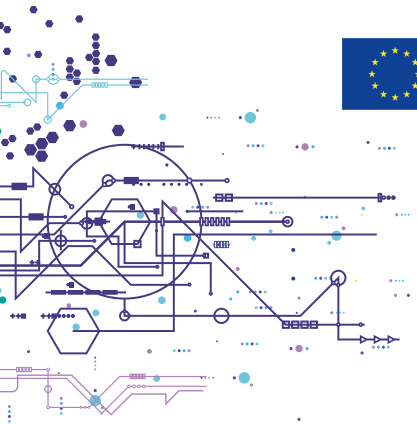
Where to Seek Documentation and Support

- `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`

Summary

- shifting the role of generic storage facilities in the infrastructure
  - emphasis on working with scientific data
  - archival functionality ⇝ NRP
  - tighter coupling to data processing
- National Repository Platform will build a pillar of the National Data Infrastructure
  - supporting new needs of scientific communities

e-INFRA CZ



Co-funded by the European Union

MINISTRY OF EDUCATION, YOUTH AND SPORTS

e-infra.cz