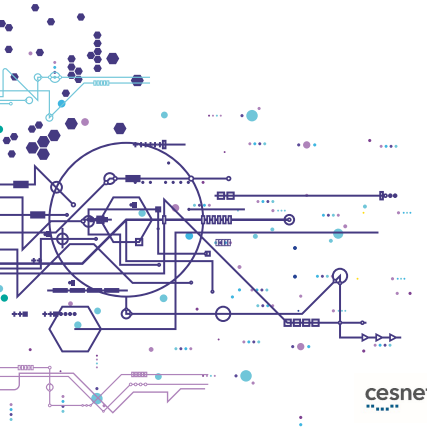
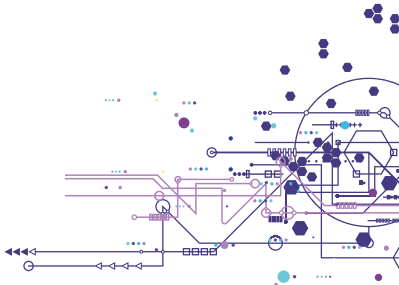
TECHNICAL ARCHITECTURE OF CZECH NATIONAL DATA INFRASTRUCTURE

David Antoš

12 September 2023





Agenda

- context of Czech national e-infrastructure
- EOSC CZ
- means to reach a “national agreement”
- technical composition of the systems
  - repository platform
  - supporting services
  - relation to other infrastructure services
- actors, projects, plans

Our Approach

- EOSC as an opportunity to elevate infrastructure services
- and bring focus on research data
  - integrating state-of-the-art practices of handling data into the infrastructure
  - supporting new requirements thrown onto the scientific community
    - e.g. push towards openness and FAIR

Context: Czech National e-Infrastructure

- large national research e-infrastructure
  - a consortium of three originally independent members
  - in the process of merging (having common projects)
- CESNET
  - legal body since 1996
  - owned by all public universities and the Academy of Sciences
  - originally NREN, expanded into Grid and Cloud computing and related services
- CERIT-SC at the Masaryk University
  - e-Science like centre with focus on cloud and other technologies and broad scientific collaboration
- IT4Innovations at the Technical University Ostrava
  - standard supercomputing centre

e-INFRA CZ Services

- network (backbone, edge connectivity of universities)
- computation—MetaCentrum (low-barrier open access, users with results getting better share)
- supercomputing—acting as a grant agency
- collaboration environment
- data storage—filesystem, object storage, EFSS
  - unstructured data
  - about 150 PB total physical capacity
  - use cases: backup/archives/sharing
  - paradigm shift necessary

(mostly) unified with AAI

Focus towards EOSC

- work plan for the whole e-INFRA CZ consortium
- focus on research data
  - data as first-class citizen
  - not just appendix to a publication
- aligned with developing *all infrastructure services*
- EOSC understood as a federation of FAIR data and related services
  - and it's up to us how to develop the idea
- EOSC implementation has been discussed since 2021
  - under the auspices of the Ministry of Education, Science and Youth (MEYS)

Searching for a National Agreement

- too many players with various interests
- working groups discussing requirements and implementation
  - open platform established during autumn of 2021
- 4 foundation working groups
  - Metadata, Architecture, Core Services, Education
- 7+1 thematic working groups
  - Bio/Health/Food, Enviro, Physics, Material Sciences and Technology, AI&ML, Social Sciences, Humanities
  - Sensitive Data

Repository

- storing data with appropriate descriptive metadata
- what is data
  - files or their collections (data sets)
  - taking into account that the line between data and publications is blurry
    - let’s not get religious about that
- store the data for “one forever” (or 5 years)

Repository in EOSC (CZ) Context

- storing data with metadata
- supporting FAIR principles
- web interface and API (= machine readable)
- organisational view: repo is responsible for its data
- should contain citable data sets
  - *ensure their immutability and consistency*
- Core Trust Seal certifiable
  - cf. `https://www.clarin.eu/content/checklist-clarin-b-centres`

National Data Infrastructure

- NDI consists mainly of
  - National repository platform (NRP)
  - National metadata catalogue
  - National repository catalogue
  - storage in the infrastructure
  - supporting systems

NDI components I

- National Repository Platform: later in more detail
- National Metadata Catalogue
  - metadata aggregator
  - generic user interface for data searching
  - actually a repository instance in the NRP
  - ongoing discussions about its metadata model
- National repository catalogue
  - just a list of repositories
- PIDs: technical implementation + support/consortia
  - assigning PIDs to stored objects
  - DataCite consortium, Orcid expected
  - more to come

NDI components II

- user management/AAI
  - Perun based + ProxyIdP
  - integrated with e-INFRA CZ
- data transfer tools
- monitoring of the infrastructure
- generic data storage
  - file systems coupled with computation resources
  - Ceph (S3, RBD, CephFS) for large data
    - for unstructured data: planned to maintain, but no extensions

Management of the Infrastructure

- organisational view
- all components except the NRP
  - managed by the e-infrastructure
  - i.e. CESNET/CERIT-SC/IT4Innovations
- NRP: emerging consortium
  - e-INFRA
  - selected universities
  - selected institutes of the Academy of Sciences

Repository in the NRP

- what is “a repository in the NRP”
  - standard repository
    - e.g. an Invenio or DSpace installation
  - in a specific configuration
    - URL
    - visual appearance
    - lists of deployed metadata models
    - support for specific metadata (e.g. visualisation)
  - with user access control
  - integrated with other systems in the NDI
    - mainly exporting metadata, registered, …
    - not using an available component should have a good reason
- a repository is typically discipline specific or institutional
  - user group requesting a repo must cooperate to set it up (admin, curator)

NRP Software Stacks

- software stacks (expected)
  - CESNET Invenio
    - fork heavy in group and record lifecycle management
  - CLARIN Dspace
    - fork by the Institute of Formal and Applied Linguistics, Charles University
  - ARL
    - used by the Library of the Academy of Sciences
    - commercial system developed by Cosmotron (company operating in CZ/SK)

NRP

- available tools (no development from scratch)
- distributed, multiple implementations
- repository instances should be created “just by configuration”
  - definitely not a separate bare-metal installation
- NRP must contain mechanisms to update repositories
  - esp. security
- NRP implementation must contain
  - ability to create repositories on request
  - documentation for repo admins and common parts for the users
  - support for repo admins, “3rd level” support for users

Resources in the NRP

- storage based on Ceph and its S3 interface
- clusters to run containers (Kubernetes)
- some partners participate on operations of these layers
  - some just operate their software stacks
- detailed relations and responsibilities being formed
- expected 5 major sites (≈ 16 racks each)
  - at least 2 secured enough for sensitive data
    - optimally 3 out of 5
  - “medical records” level

Other NRP Implementations

- user groups with established tools?
  - we can offer at least storage+cluster resources
- the tool should “preferably be a repository”
  - even though it is often called a “database” or whatever
  - it should provide guarantees for marked records to be immutable and citable
- integrated to supporting systems
- must have a team to keep it running and to support it
- may/may not come with their own HW resources

Users

- repository end user
  - managed by the repository administrator
- repo admin
  - partner for the infrastructure
- ≈ virtual organisation admin
  - “members and administrators”

Projects I

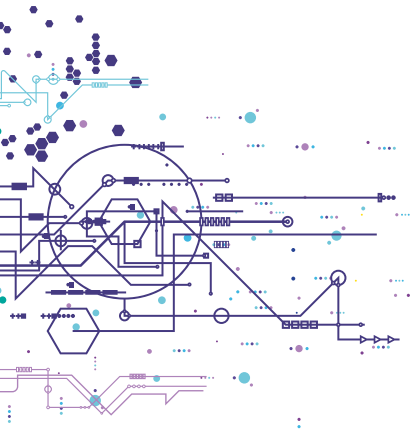
- Czech Academic and Research Discovery Services (CARDS)
  - National Library of Technology
  - metadata support, PID consortia
  - library catalogues (out of scope for us)
  - running
- EOSC CZ
  - e-INFRA CZ/CESNET
  - approx. 18M EUR, 6 years
  - centralised services
    - metadata catalogue, AAI, monitoring, …
    - EOSC secretariat
  - running since January

Projects II

- National Repository Platform
  - consortium coordinated by CESNET
  - expected 8-10 universities + academy of science institutes
  - approx. 45M EUR/6 years
  - expected about 12M to hardware
  - to be submitted in November
  - covering
    - infrastructure, running the NRP
    - pilot user groups
    - service development and integrations (e.g. data FAIRification, deposition automation, workflow integration, license management, AAI, …)
    - education (preparing courses on data stewardship, …)
- separate call for e-INFRA CZ services
  - heavily reduced, though
  - to be submitted in November

Projects III

- expected another call in 2 years: “Open Science II”
  - supporting user communities
  - covering their specific needs
  - FAIRification of various data resources/databases
  - (wild guess) to kick-start transition of the environment into higher level of integration
    - including computation, live data processing
    - building actual virtual research environments



Contact
info@einfra.cz

e-infra.cz