## EOSC Initiative in the Czech context

Matej Antol, Ph.D.







### About me

- Principal Project Manager of the IPs EOSC CZ project
  - systemic, strategic project within the Open Science call
  - initiative creating national environment for FAIR research data
  - https://www.eosc.cz/
- CEO @ CERIT-SC, ICS MU
  - one of three partners of the national e-infrastructure e-INFRA CZ
  - www.cerit-sc.cz/
  - https://www.e-infra.cz/
- Research background
  - Analysis of complex, unstructured data
  - Faculty of Informatics, Masaryk university, CZ
  - <a href="https://disa.fi.muni.cz/complex-data-analysis">https://disa.fi.muni.cz/complex-data-analysis</a>





# In this presentation

- On value of research data
- EOSC CZ two years in the making
- Czech National Data Infrastructure
- Glimpses at the first operational services
- What's next





### On value of research data







# Chemistry, structural biology and life

- PDB (protein data bank) was established in 1971
  - 60.000 depositors
  - 500.000 data entries: experimentally obtained protein structures
- A couple of weeks ago, the Nobel Prize for Chemistry was awarded for computational protein design and protein structure prediction - AlphaFold.
  - We now have 200.000.000+ protein structures available
- Both PDB and AFDB now used by millions for literally endless applications such as decomposing plastics and antibiotic resistance.



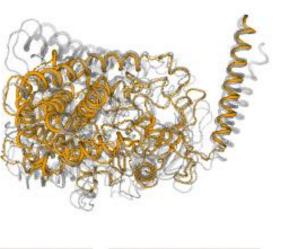
David Baker

Prize share: 1/2



Demis Hassabis

Prize share: 1/4



III. Niklas Elmehed © Nobel Prize

John M. Jumper Prize share: 1/4

The Nobel Prize in Chemistry 2024 was divided, one half awarded to David Baker "for computational protein design", the other half jointly to Demis Hassabis and John M. Jumper "for protein structure prediction"



# Astronomy, physics and the universe

- Since 2000, The Sloan Digital Sky Survey (SDSS) collects data of galaxies
  - More than 500 contributors from 13 countries
  - Millions of data entries: 2.5 million galaxies and 400.000 quasars
- Its data helped prove the accelerating expansion of the universe, leading to the Nobel Prize in Physics in 2011 for the discovery of dark energy.
- Currently, it is a resource for over 10.000 scientific papers, and contributed to the discovery of over 500.000 new galaxies and 20.000 asteroids



© The Nobel Foundation. Photo: U. Montan Saul Perlmutter

Prize share: 1/2



© The Nobel Foundation. Photo: U. Montan Brian P. Schmidt

Prize share: 1/4



© The Nobel Foundation. Photo: U. Montan Adam G. Riess

Prize share: 1/4

The Nobel Prize in Physics 2011 was divided, one half awarded to Saul Perlmutter, the other half jointly to Brian P. Schmidt and Adam G. Riess "for the discovery of the accelerating expansion of the Universe through observations of distant supernovae"





# Art, provenance, culture and history

- Getty Provenance Index (GPI) exists since 1980
  - Data from more than 50 institutions worldwide, including museums, galleries, and auction houses, as well as individual collectors and research scholars.
  - Over 1.5 million records: over 300,000 individual works of art, 270,000 owners and 200,000 auction sales.
- Significant impact on restitution efforts, particularly for artworks looted during World War II, just as the painting called "The Polish Girl" by Jean-Baptiste Greuze
- Big impact on Art History and Provenance Research, supporting over 1.000 scholarly publications, spanning fields like art history, economics of the art market, cultural heritage studies, and legal research on restitution.





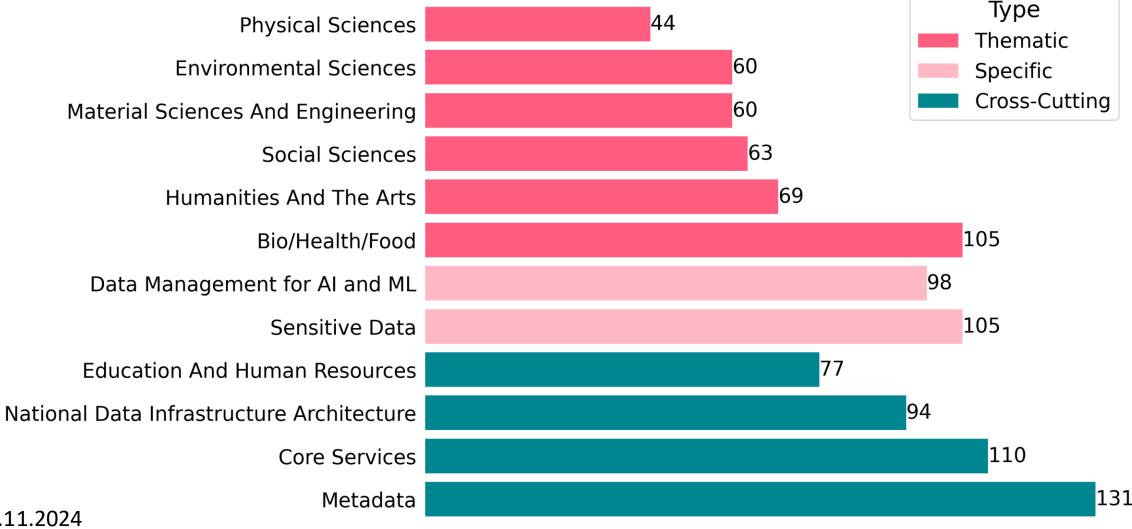
# EOSC CZ two years in the making



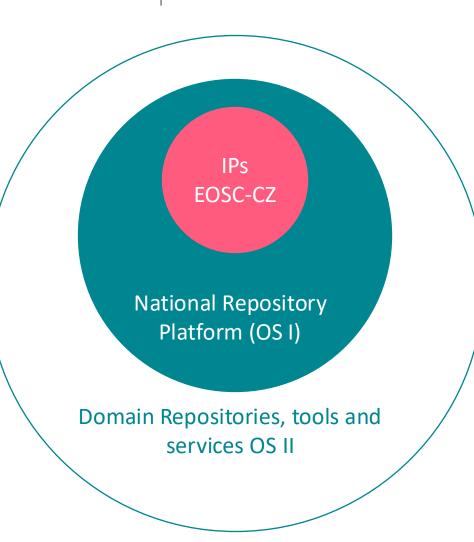




# 450+ people in EOSC CZ working groups



26.11.2024



#### IPs EOSC-CZ (since 2023) – Fundamentals for EOSC implementation in CZ

- Organizational (**Secretariat**) <a href="https://www.eosc.cz/en/secretariat">https://www.eosc.cz/en/secretariat</a>
- Technical (National Metadata Directory) -- https://nma.eosc.cz/
- Knowledge and skills (**Training Centre**) -- <a href="https://www.eosc.cz/en/training-centre">https://www.eosc.cz/en/training-centre</a>

#### National repository Platform (OS I, since 2024) - "technical core"

- Repository platforms (dspace, cesnet invenio, asep arl) (50+ PB user capacity)
- First exemplary repositories
- Core services (PIDs, DSW, licenses, ...)
- **Compliance** and UX (cybersecurity, ServiceDesk, ...)
- **Training** technical side of things

#### OS II (since 2025) - "domain specifics"

- Under preparation, content not clear yet
- Based on expertise of the 8 thematic / discipline workgroups
  - Bio/Health/Food, Matech, Al & ML, Social Sciences, Physics,
     Humanities & Arts, Enviro, Sensitive Data
  - see <a href="https://www.eosc.cz/en/working-groups">https://www.eosc.cz/en/working-groups</a>



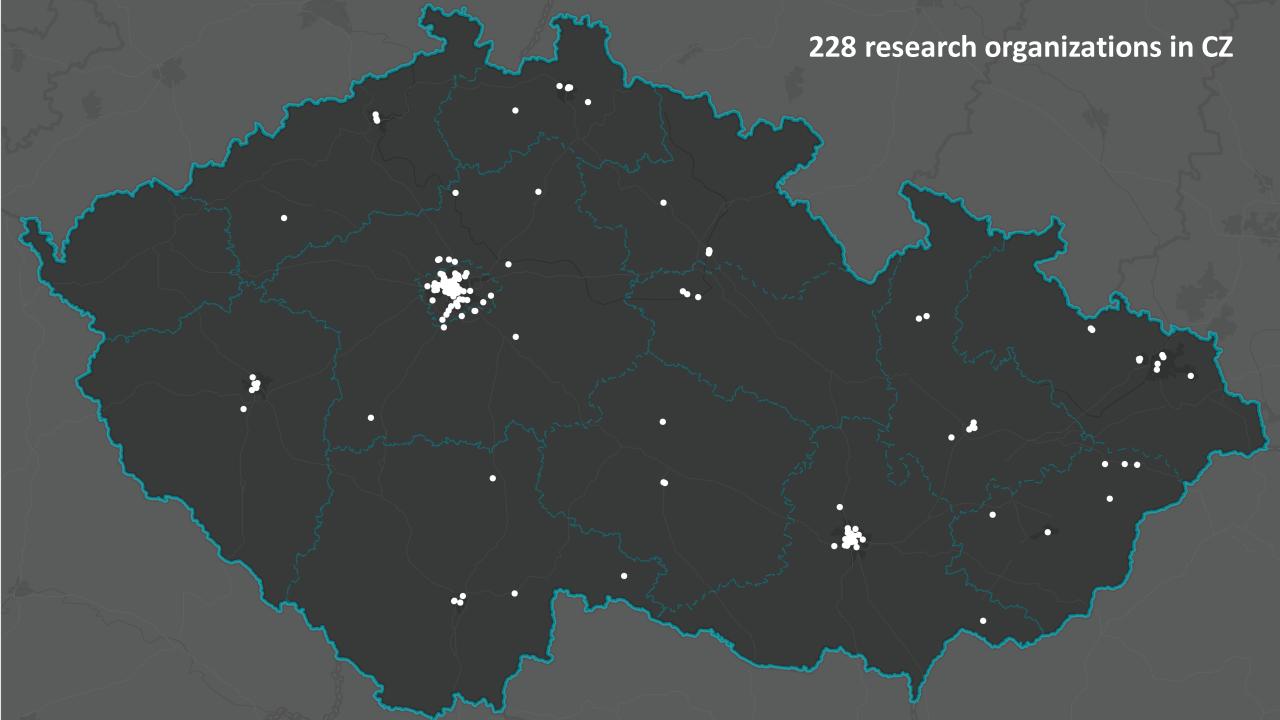
#### Conferences, workshops and community meetups

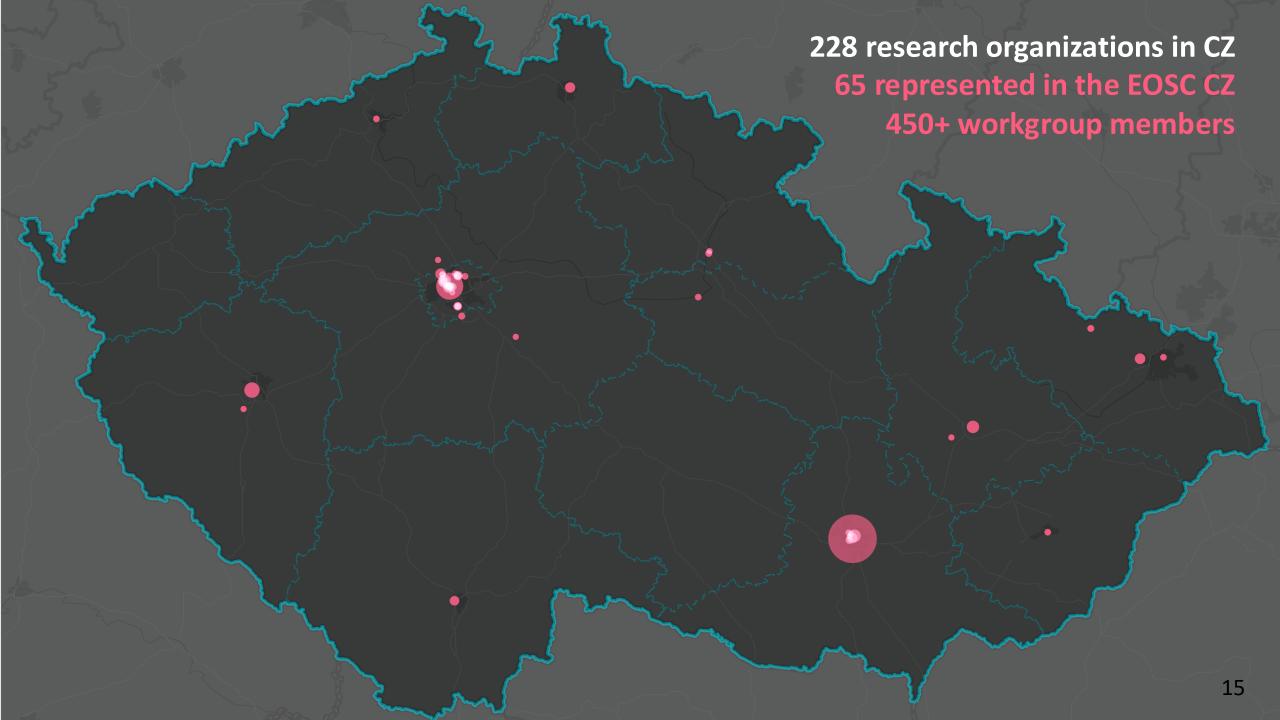
- 21 trainings and workshops
  - ca **3.000 registrations** from ca 90 research institutions
- 3 Conferences with hundred+ attendees
  - 150 visitors of this conference present + more than 150 online
- 31 Physical meetings of communities and working groups
  - 3 roadshow in Prague, Brno and Ostrava
  - 5 EOSC CZ Networking events bringing EOSC CZ to the universities in České Budějovice, Liberec, Olomouc, planned trips to Zlín and AV ČR
  - + over 100 online working group meetings
- Thousands people continuously in touch
  - 300+ followers on social networks
  - 300+ EOSC newsletter subscribers
  - 3.000+ views of content on YouTube

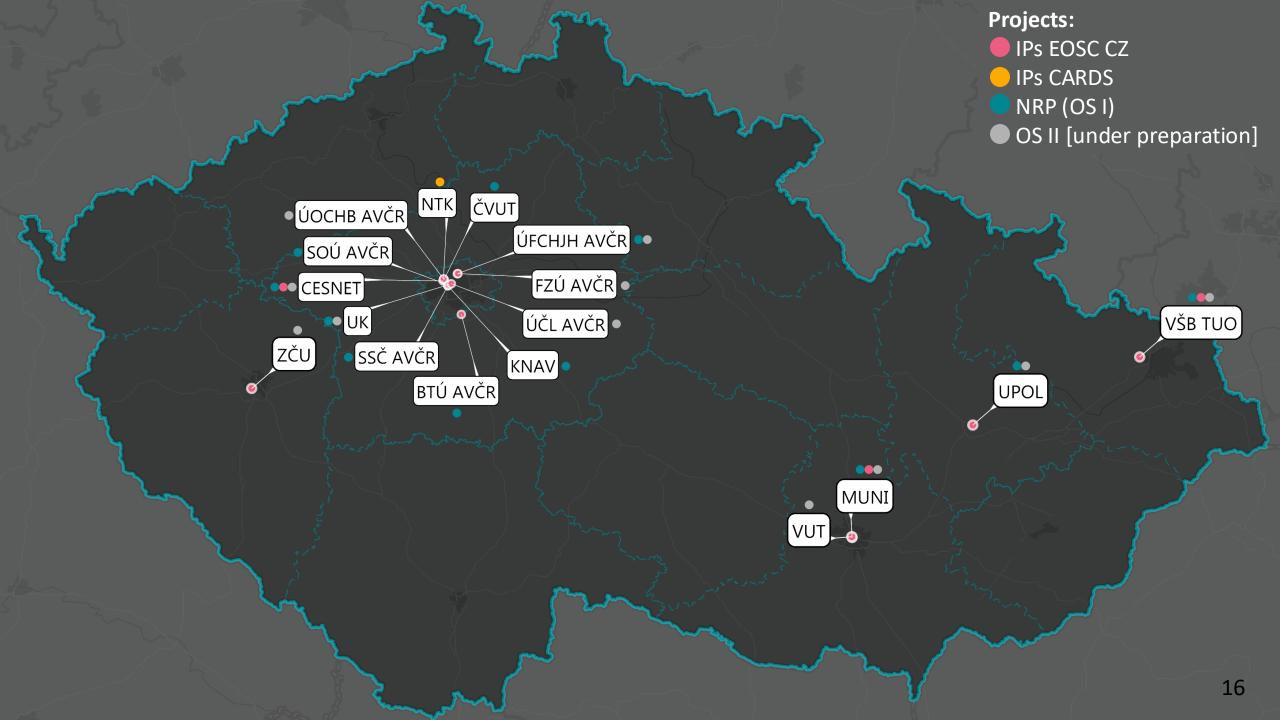


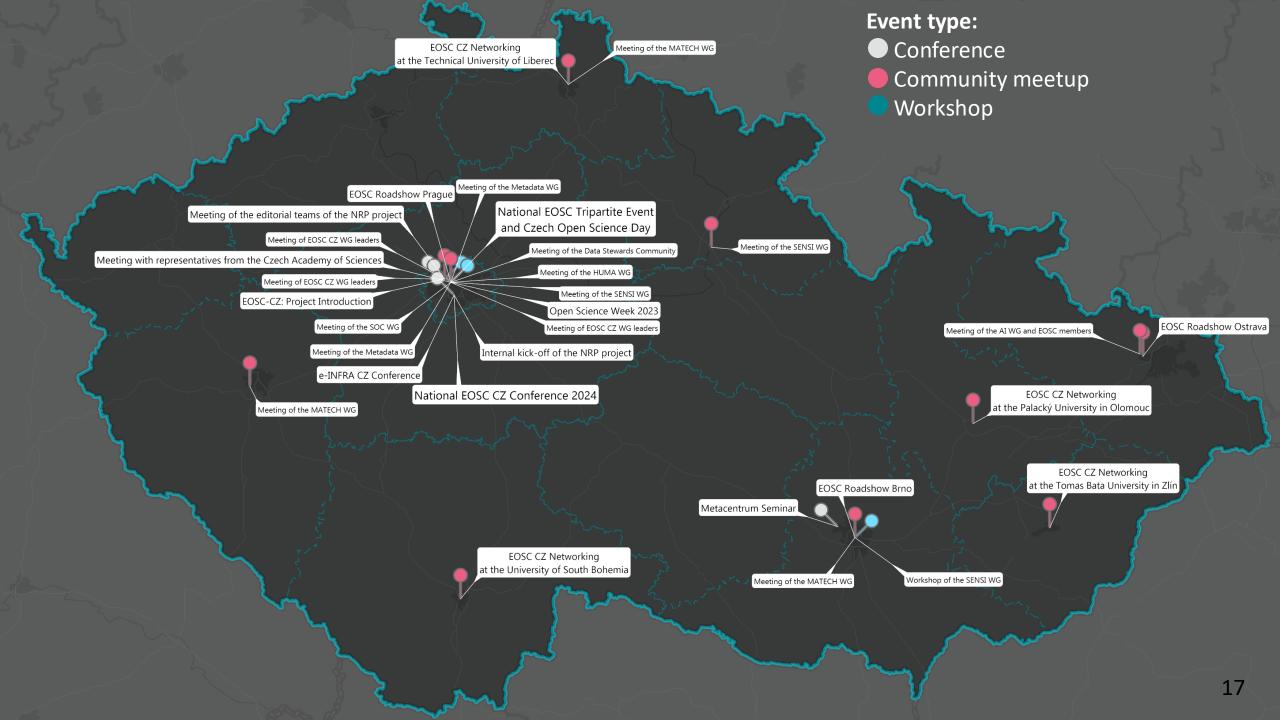














# Shaping the European development

- Notable engagement of Czech professionals
  - 14 Czech members in the EOSC Association Task Forces
    - 2 Czech Co-Chairs: Jiří Marek and Petr Holub
  - Members in the Opportunity Area Expert Groups: Miroslav Ruda, Marek Cebecauer and others
  - EOSC Executive Board (2019 2020), EOSC Steering Board Vice Co-Chair (since 2023) and Tripartite group: Jan Hrušák
- Development of the infrastructure
  - CESNET subcontractor in LOT 1 and LOT 3 developing EOSC EU Node
  - Regular contribution and consultations to the EOSC Association activities:
     SRIA and MAR, Federation Handbook, EOSC Nodes survey, EOSC Symposium, etc.
  - Participation in the INFRAEOSC projects (Horizon Europe), OSCARS, BEYOND, ENTRUST, AARC TREE, CRAFT-OA (EOSC United submitted)













### The Czech National Data Infrastructure





# Infrastructure components

#### Repository platform

- Number of platforms -- CESNET Invenio, Clarin-DSpace, ASEP ARL
- Total of 50+ PB of user data storage capacity
- Offered to research communities to create and operate specific repositories

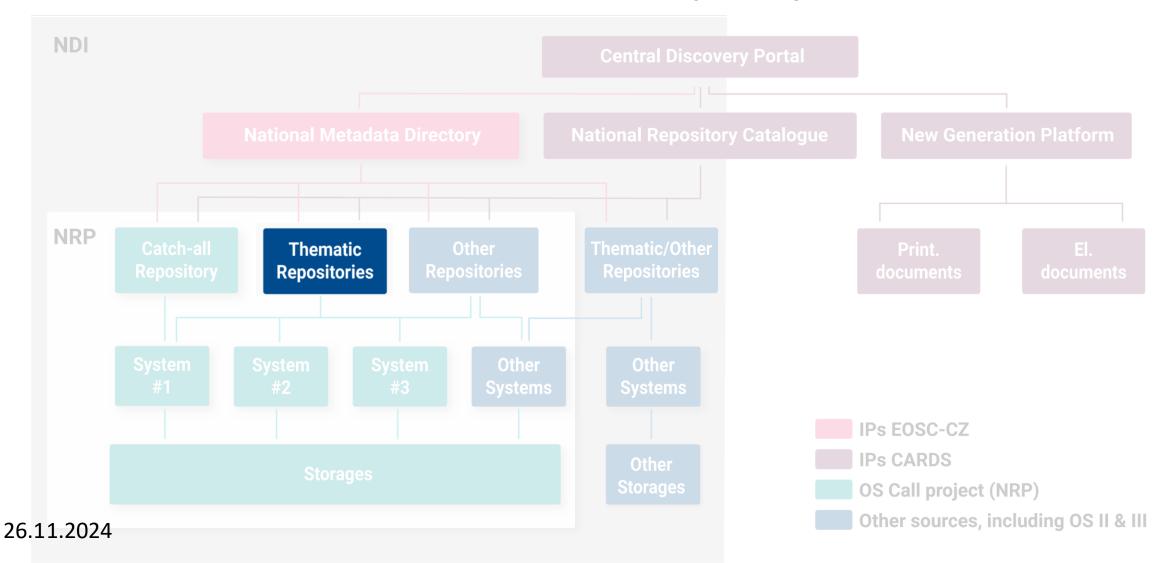
#### Services

- Support for Data Management Planning
- Support for persistent identifiers
- AAI
- FAIRificator
- Interfaces to computing environments for analysis
- Data and objects search and discovery
- Monitoring
- Support for metadata

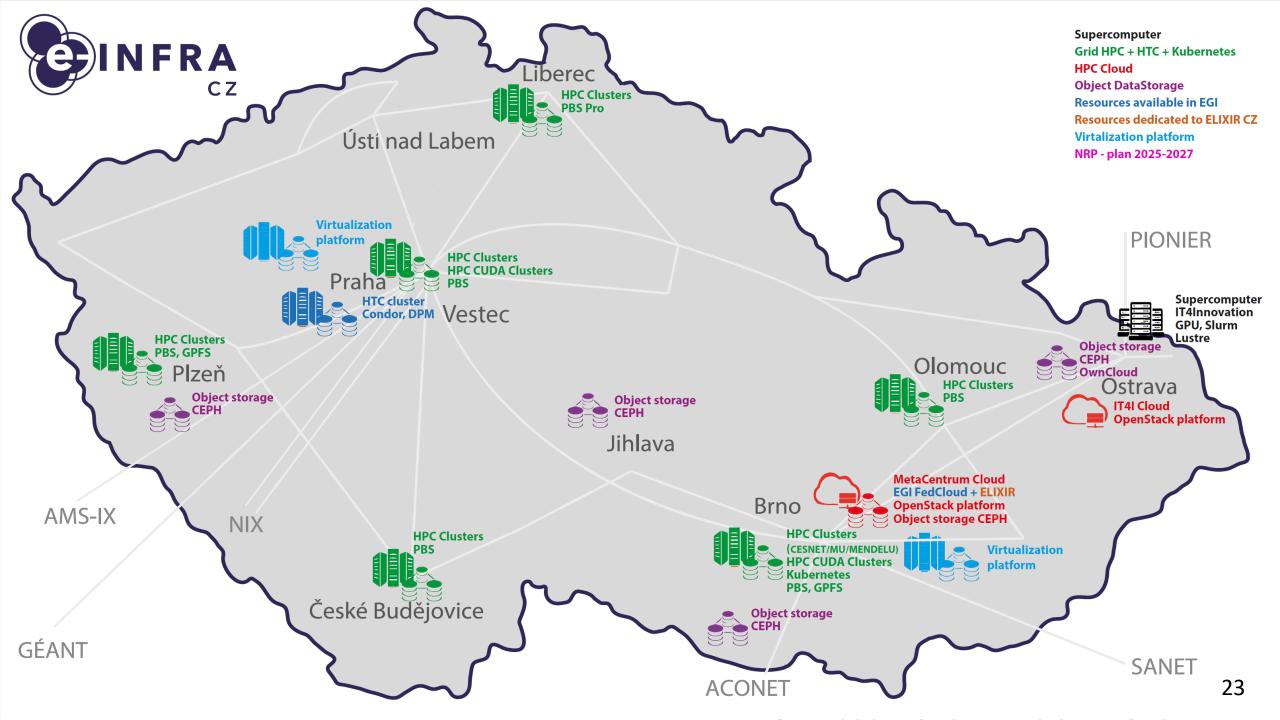


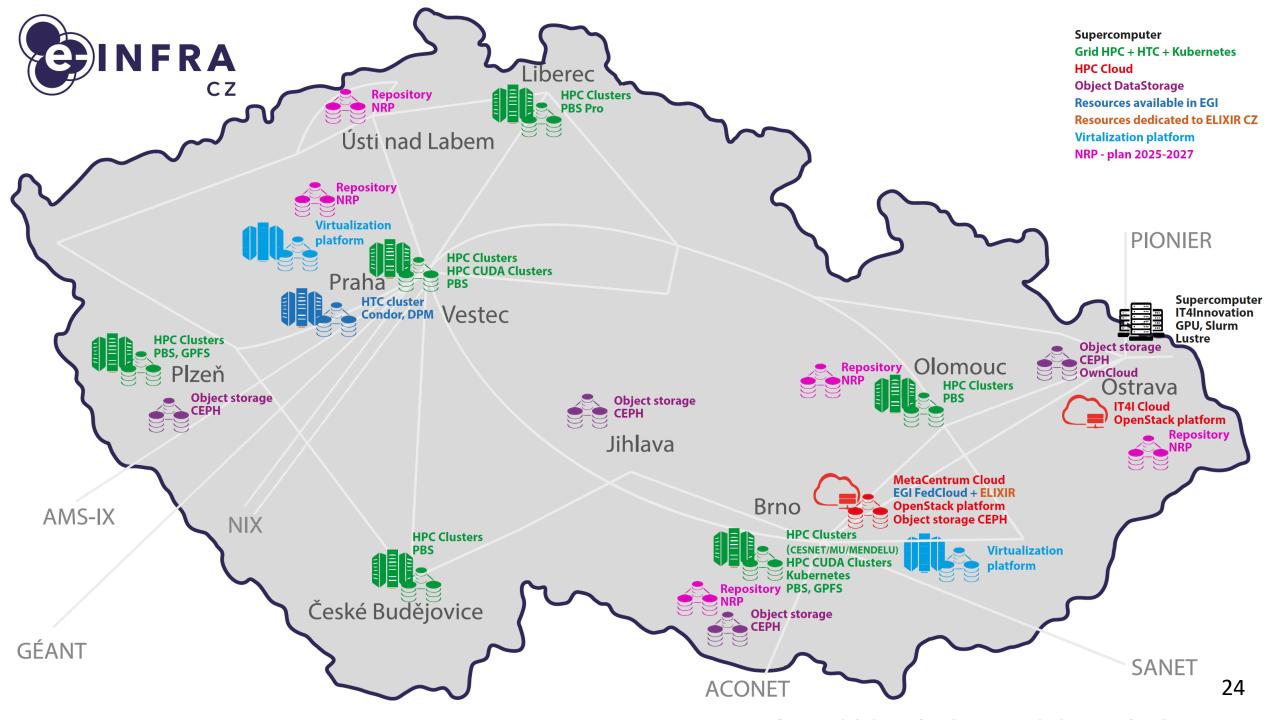


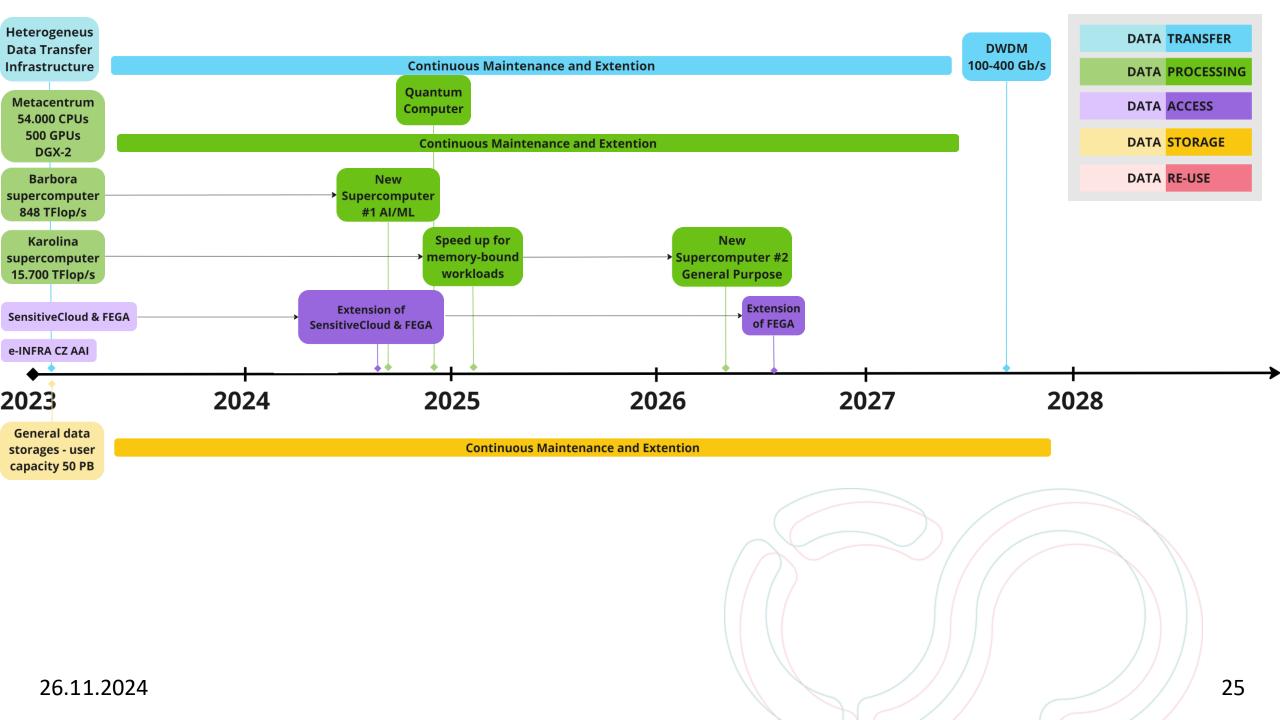
# National Data Infrastructure (NDI)

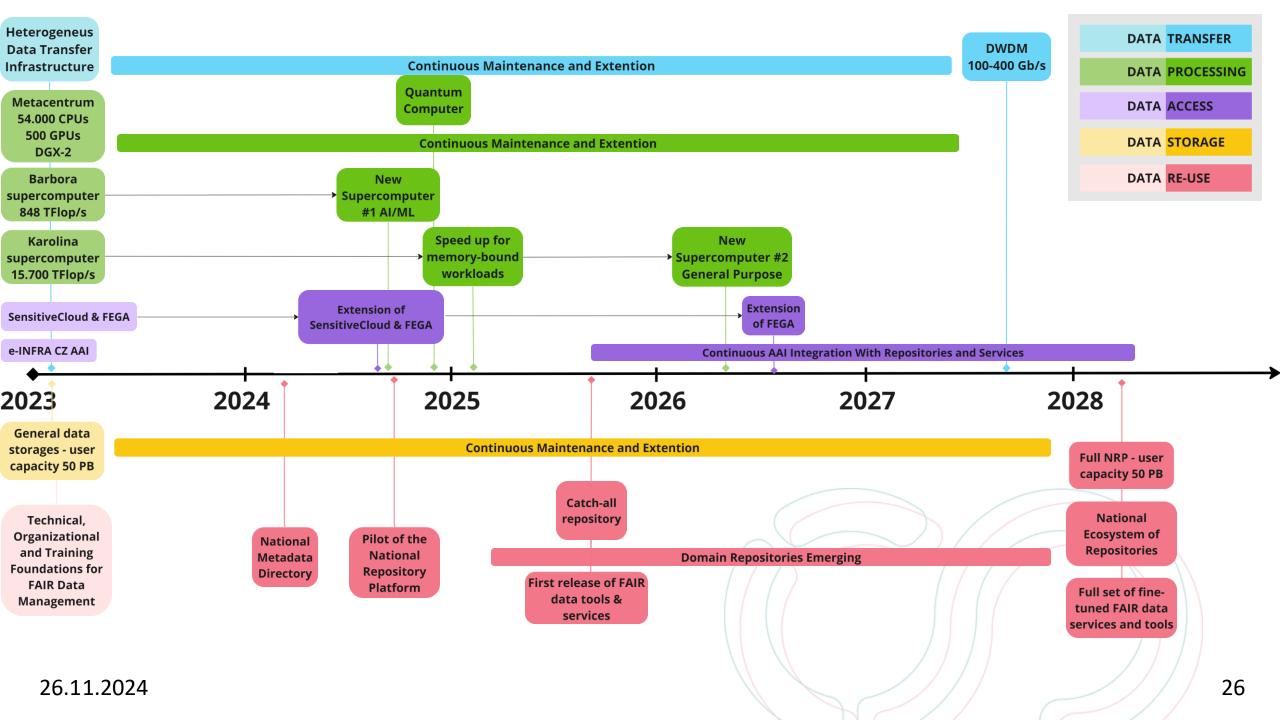








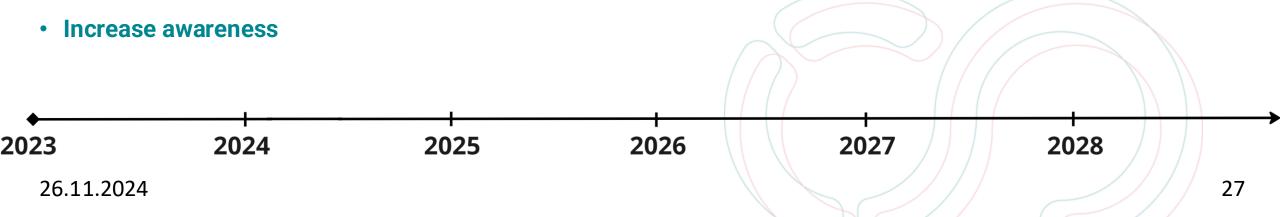




# Nearing the end of the first phase

#### Phase 1

- Engage Czech researchers and professionals
- Organize the work groups and communities
- Kick-off 3 major projects
- Design the core infrastructure and services
- Release first services



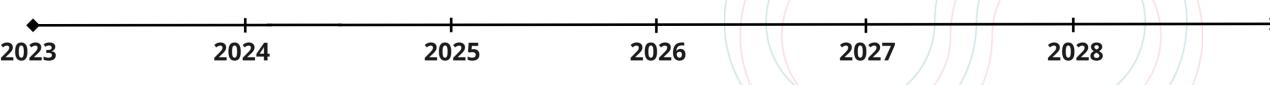
# Beginning the second phase

#### Phase 1

- Engage Czech researchers and professionals
- Organize the work groups and communities
- Kick-off 3 major projects
- Design the core infrastructure and services
- Release first services
- Increase awareness

#### Phase 2

- Deploy the infrastructure
- Deploy all the core services
- Integrate and deploy first repositories
- Support the uptake of data management skills and know-how
- Integrate the national and international ecosystem
- Propose the sustainability model





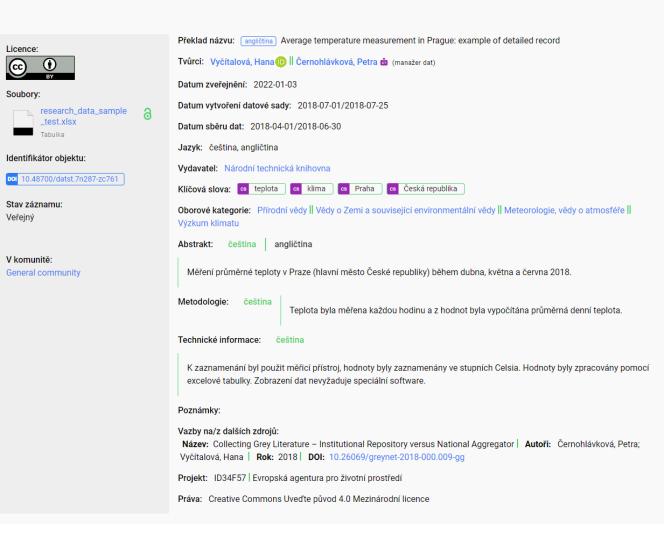
# Glimpses at the first operational services for FAIR research data management

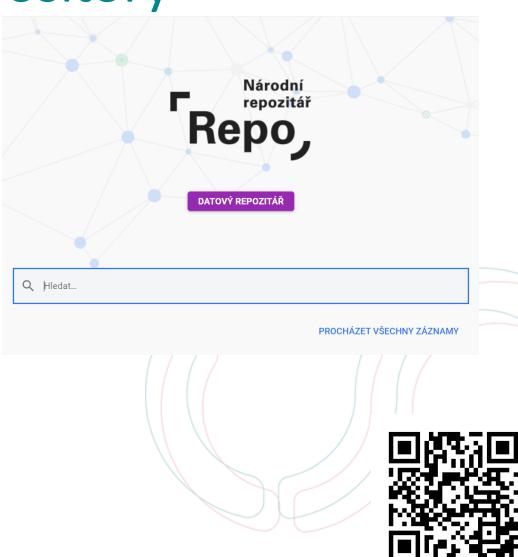




# National "catch-all pilot" Repository

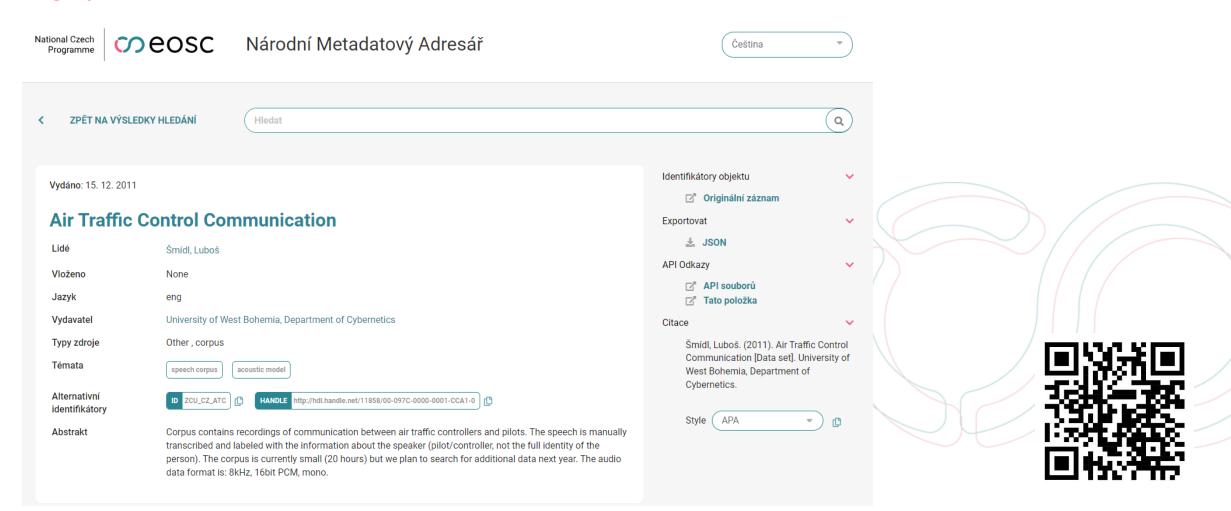
detailní záznam Měření průměrné teploty v Praze: vzorový detailní záznam





# National Metadata Directory

• Single point of contact for research data – uniform format and metadata



### Identifikátory CZ – Portal for Persistent Identifiers

#### didentifikátory.cz

Persistent Identifiers



Home / Persistent Identifiers

#### Persistent Identifiers

Learn more about each persistent identifier (PID). Persistent identifiers are tools that are used to uniquely identify people, organisations, and other objects (e.g., books, articles, datasets) in a scholarly communication system.







### SensitiveCloud

#### **Environment for processing sensitive data**

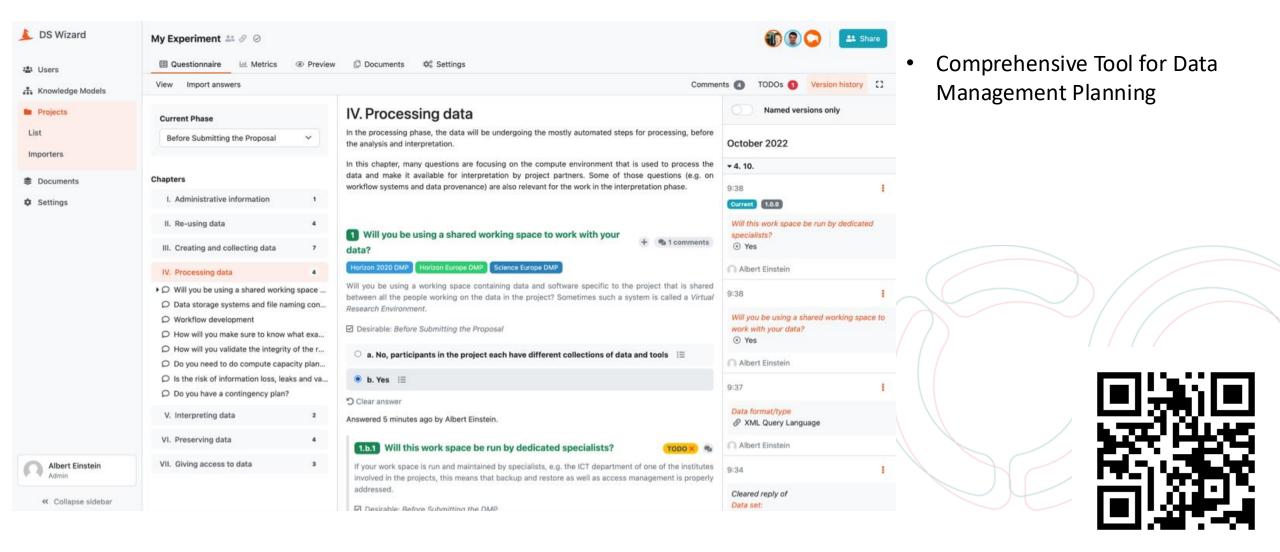
- Virtual desktop
- Computing resources
- Secure applications
- Storing, sharing and cooperation
- VPN, Kubernetes





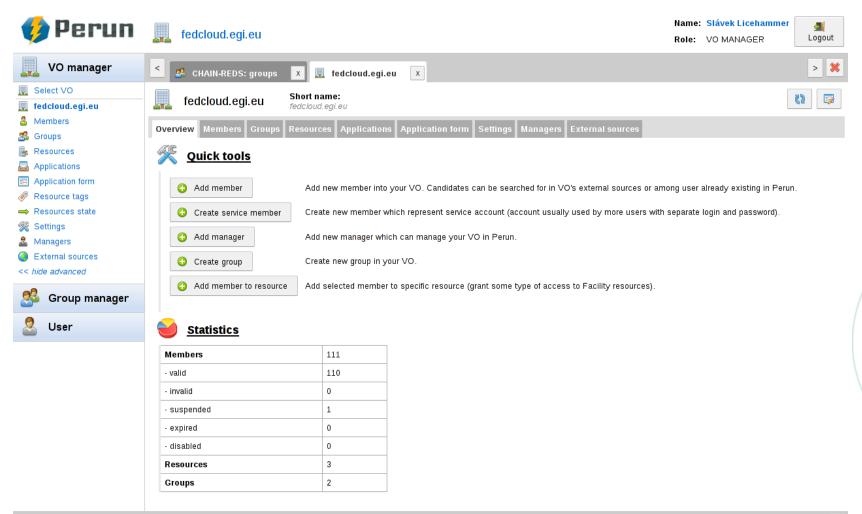


# Data Stewardship Wizard





### Authentication and authorization infrastructure (AAI)



To allow users from different institutions easy access to data and services.

- Access and identity management
- Management of groups and roles
- Rights' delegation
- System integration

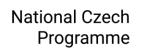
Settings:





## What's next







### In 2025...

#### Infrastructure

- Installation of first hardware for repositories
- Emergence of first and pilot repositories
- National Repository Catalogue

#### **Services**

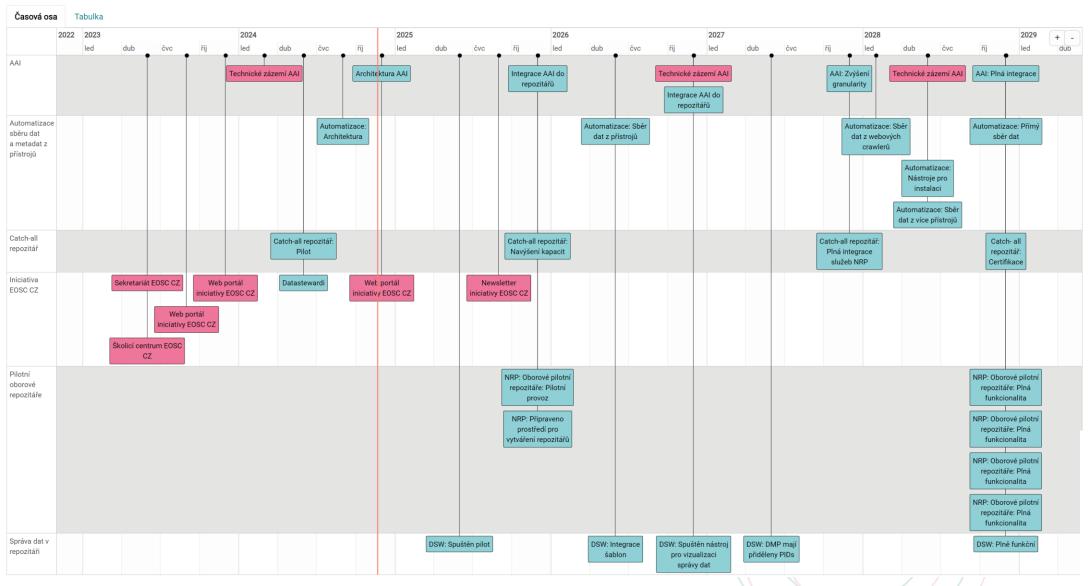
- First versions of all major core services, including
  - Repository platforms
  - FAIR Implementation Profile Wizard
  - Data Stewardship Wizard
  - License management
  - AAI for repositories
    - ... and more

#### **Events**

- National Tripartite Event
- Data Steward Summer School
  - ... trainings, workshops, seminars, conferences, ...









https://www.eosc.cz/projekty/narodni-repozitarova-platforma-pro-vyzkumna-data-os-i-nrp/nrp

#### National Czech Programme





visit https://www.eosc.cz/en

#### EOSC CZ: Towards the development of Czech national ecosystem for FAIR research data

Matej Antol O, Jiří Marek O, Michaela Capandová O, Jaroslav Juráček O, and Luděk Matyska O

Abstract—This short paper presents a compact overview of Of Sciences and Research Infrastructures. Examples are enthe Czech approach to implementing the European Open Science vironments such as LINDAT/CLARIAH-CZ [I3] for natural Cloud and plans for developing a Czech national infrastructure for FAIR research data. Its purpose is to provide an allencompassing summary of the near future of research data management in Czechia As such we deliberately attempt to explain complicated concepts in minimum words, sacrificing the precision of expression for compactness.

Index Terms-EOSC, EOSC CZ, FAIR data, National Data Infrastructure, National Repository Platform, Open Science

#### I. INTRODUCTION

HE importance of data in research is continuously rising. while approaches to store, manage and share these data seem to fall behind. The value of the data is reduced by their considerable heterogeneity and lack of structure, which leads to low reproducibility and hinders scientific progress. Open Science (OS) [1] seeks to address some of these current issues, focusing on data availability and sharing, urging for more collaboration and emphasising research integrity. European Open Science Cloud (EOSC) [3]. [4] is an international initiative that builds on the Open Science principles. EOSC seeks to create a common European research environment [5] to store, share and re-use research data and other digital objects without barriers. We call such data and objects FAIR [2] (Findable, Accessible, Interoperable, Reusable)

#### II. EOSC CZ - INFRASTRUCTURE AND SERVICES FOR FAIR RESEARCH DATA

The establishment of fundamental principles for the Czech national FOSC implementation took place in 2021, resulting in the document called Architecture of FOSC implementation in the Czech Republic [6]. The document represents the official start of the EOSC CZ initiative [7]. The primary tangible outcome of this initiative will be a National Repository Platform (NRP) - a core component of the National Data Infrastructure (NDI). NRP will be a federated ecosystem of distinct technological layers (see Fig. []) and associated services (see below).

The data infrastructure will complement the existing Czech national e-infrastructure e-INFRA CZ [11] with all its services. NDI will be fully integrated at the European level [12]. NRP will interconnect with the already running parts of NDI: data repositories and services held at universities, Czech Academy

Authors are with the CERIT-SC centre, Institute of Computer Science, Masaryk University, 60200 Brno, Czechia. Contact at info@eosc.cz. Manuscript published February 20, 2024

language processing, Czech-BioImaging [14] for biological and medical imaging or EIRENE RI [15] for human exposome.

Next to the repositories themselves, the initiative plans deploy and integrate several FAIR data-related services designed for NDI users. Notably: Central Discovery Portal (CDP) integrated into the New

- Generation Platform (PNG) will ensure the searchability and availability of all types of resources (electronic digitized and printed) and research results. National Metadata Directory to search in NDI metadata
- · Single Authentication and Authorization Infrastructure (AAI) solution Perun [16] to guarantee data accessibility. Support for data management planning via Data Stewardship Wizard [17]
- · Support for Persistent Identifiers (PIDs) [18].
- · Support for data FAIRification.
- Data mgmt. tools such as OneData [19] or iRODS [20].
- . Training [21] and university courses on data management.

#### III. ACTIVE COMMUNITIES AND HOW TO PARTICIPATE

Researchers' engagement is vital for the EOSC CZ's success. Since 2021, as a reaction to the EOSC CZ Architecture document, 12 EOSC CZ working groups [22] have been established through a self-organizing community effort. These groups will be operational during the entire EOSC CZ initiative, and registration is continuously open to new potential members. A list of their members is publicly available. Currently, the initiative is in its initial implementation phase, and the active participation of scientists in the working groups

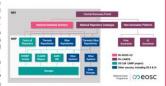


Fig. 1. NDI and NRP blueprint with five abstraction layers. Bottom-up: hardware infrastructure dislocated across Czechia; three initial repository systems – CESNET Invenio [8], CLARIN-DSpace [9] and ASEP/ARL [10];

s the main guarantor for the NDI ecosystem to encompass and support all relevant research data management needs of research communities.

Onen Science coordination team within the National Library of Technology On top of that, collaboration is being established with the already existing national Open Science communities:

- · Open Science working groups of the Association of Libraries of Czech Universities,
- · national Data Steward Community and
- · members of the institutional Open Science centres within Czech academic institutions.

IV. How to benefit from the EOSC CZ outcomes

The NDI's ecosystem of services will be offered to the whole research community regardless of their active participation in the EOSC CZ initiative. The EOSC CZ Secretariat [23] and Training Centre [21] are already operational, providing consultancy, seminars and workshops for the Czech research cosystem. The National Metadata Directory will be deployed. in 2024, followed by the NRP with a portion of core services in 2025. By this time, the first domain and other repositories should also be emerging. This first phase will be completed in 2026, with an entire NRP and its services available. The initiative will concurrently foster the development of data management and other related skills for all Czech academia members. It will also encourage the systemic formation of data steward and curator roles across the academic ecosystem.

With this infrastructure, any reasonably interested Czech scientist should have sufficient information, know-how, skills, institutional support, and services to store, share, and reuse research data efficiently. These ambitions summarize the main objective of the EOSC CZ initiative.

#### ACKNOWLEDGMENTS

The EOSC CZ initiative has active collaborators who significantly exceed the authors of this paper. Out of these, we would namely like to acknowledge the contributions of Radka Římanová, Klára Slanařová, Petra Černohlávková, Martin Svoboda, Miroslav Bartošek, David Antoš and Michal Růžička.

APPENDIX: FINANCIAL SUPPORT FOR EOSC IN CZECHIA Czech Ministry of Education, Youth and Sports (MEYS) supports the EOSC CZ initiative [24] via two systemic projects and three open science calls

- · Individual Systemic Project (IPs) EOSC-CZ, coordinated by Masaryk University with two additional partners. supported with 18 mil. EUR to provide a fundamental organizational, technical, and training environment.
- . IPs CARDS, coordinated by National Library of Technology, supported with 56 mil. EUR, to provide support for PIDs, research data description, and deliver the PNG. · OS Call I, with an allocation of 50 mils. EUR, to create
- the NRP, its core services and related training. · OS Call II, with an allocation of 36 mil. EUR to support domain-specific data management, repositories and
- related services over the NRP. OS Call III, scope of which is currently under discussion.

#### REFERENCES

- 111 Munafo, M., Nosek, R., Rishon, D. et al. A manifesto for reproducible science. Nat Hum Behav 1 (2017). doi.org/10.1038/s41562-016-0021

  Wilkinson, M.D. et al. The FAIR Guiding Principles for scientific data The initiative is also closely connected with the National management and stewardship. Scientific data, 3(1), pp.1-9. (2016)
  - https://eosc-portal.eu/
  - [6] https://www.msmt.cz/uploads/311/Architektura\_implementace\_EOSC\_ v CR ndf

  - 101 https://asep-portal.lib.cas.cz/basic-information/dataset-repository/

  - https://www.e-infra.cz/en https://eosc.eu/tripartite-collaboration/czech-republic/
  - https://lindat.cz/
     https://www.czech-bioimaging.cz/
  - [15] https://www.eirene-ri.eu/

  - 81 https://identifikatory.cz/en
  - https://www.cerit-sc.cz/management-of-data-workflows

  - https://www.eosc.cz/en/working-groups https://www.eosc.cz/en/secretaria
  - seznam-operaci-(prijemcu)



Matei Antol is the principal project manager of the IPs National Auton is the principal project manager of the 175 EOSC-CZ. He is also the integration manager of the Czech e-infrastructure e-INFRA CZ and an executive director of one of its three partners, the CERIT-SC infrastructure. He has a long background in IT and research projects. His research activities focus on man aging and analysing complex, high-dimensional data.



Jiri Marek is the General Secretary of the EOSC CZ initiative and head of the FOSC CZ Secretariat. He holds the role of the Open Science manager at Masaryk University and serves as a head of the CZARMA Open Science Task Force. He is also involved with activities regarding digitization of the public sector via open echnologies (Open Cities, etc.)



CZ Working Groups Metadata and Materials Sciences and Engineering. Her research in the biomedical field focused on the development of cellular elements d biomaterials for lung tissue engineering. She loves trospinning and scanning electron microscopy.



Jaroslav Juracek is the secretary to the EOSC CZ Working Group Bio/Health/Food. Beyond that, he takes part in building the European Genomic Data Infrastructure and related activities at the national level. His focus is set on advancing open science initiatives and access to and utilization of genomic data for research



Ludek Matyska is a full professor at the Faculty of Informatics, Masaryk University, with a long track in developing national and European research infrastruc-tures. He is the director of the CERIT-SC, one of three members of the e-INFRA CZ steering board the principal project manager of the NRP project, and

#### read https://arxiv.org/pdf/2402.13343

26.11.2024 39

# Thank you for your attention

