

Czech FAIR Data Infrastructure

Luděk Matyska

CESNET & Masaryk University



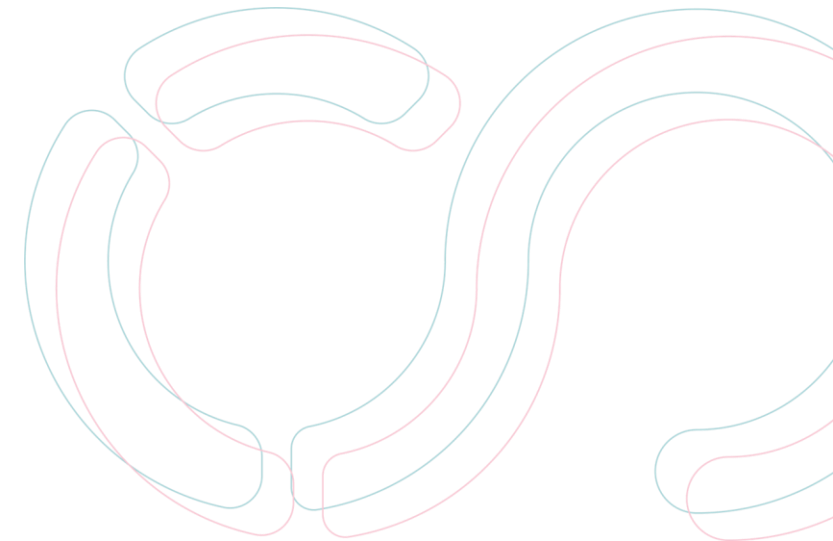
Co-funded by
the European Union



MINISTRY OF EDUCATION,
YOUTH AND SPORTS

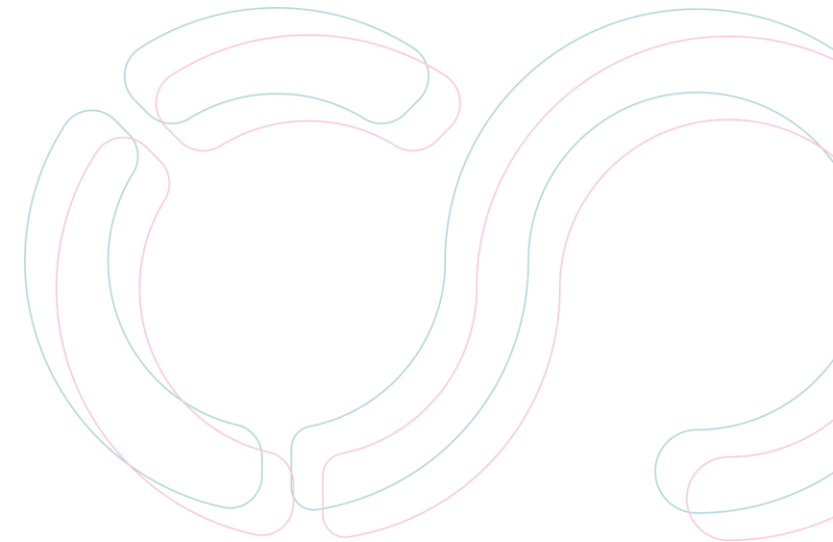
About me

- Principal Project Manager of the NRP for Research Data systemic, strategic project within the Open Science call
 - Building the NRP
 - <https://www.eosc.cz/>
- Vice-director of ICS MU
- Director @ 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/>
- ELIXIR CZ Board Vice-Chair
- Research
 - Large scale distributed infrastructures
 - Distributed Security/AAI
 - Sensitive/Human Data Processing Environments



In this presentation

- Some background
- What
- Why
- How
- In an European context



Background

- EOSC in development since 2016
 - More than half a billion Euros invested
- Time to have something tangible before the 10th anniversary
- Originally, EOSC presented as a concept, not an infrastructure
 - Created a lot of misunderstanding
 - What we are actually trying to achieve/build?
- Part of the Open Science
 - A change how we do science – Really?
 - Mindset change – Why? Who benefits?

EOSC targets

- Research data are valuable
 - And large part is lost after initial processing
- Only properly annotated data are valuable (even in mid term)
- The value goes beyond groups that create the (primary) data
- The data are critical for the research reproducibility
- Not only data, but the processing tools and environments must be maintained (the other digital artefacts)

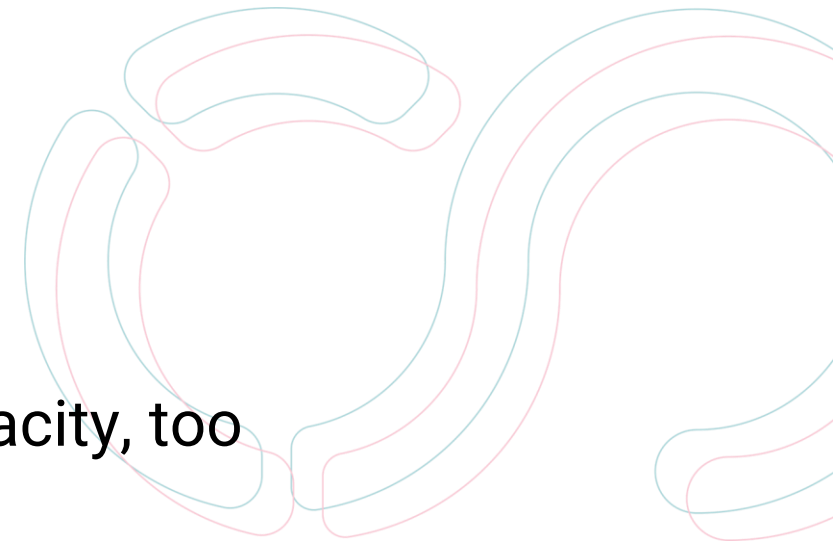
- EOSC as a web of data and services and tools

EOSC Ecosystem

- EOSC (sometimes) presented as the solution of the still unsolved problems with data
 - Overlooking decades of work in many scientific disciplines
 - Overlooking/diminishing role of ESFRI in care of data
- => There is a lot to build on, “just” to find a proper way
- Different national solutions
 - As a reaction to all these uncertainties
- Increase risk of loosing focus
 - Unclear what to focus to

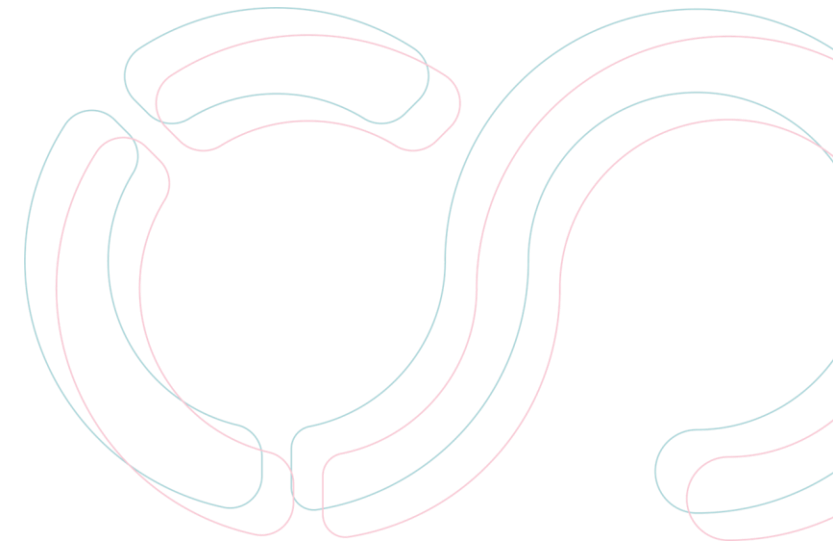
What (we are building)?

- An ecosystem for complex support of dealing with FAIR data
 - National Data Infrastructure
- National Data infrastructure
 - Czech contribution to the “What EOSC should be”
- Not reinventing a wheel
 - Extensive involvement of large research infrastructures
 - Used to think in a long term perspective
 - Institutions
 - And researchers themselves
- Not only some technology and services, but human capacity, too



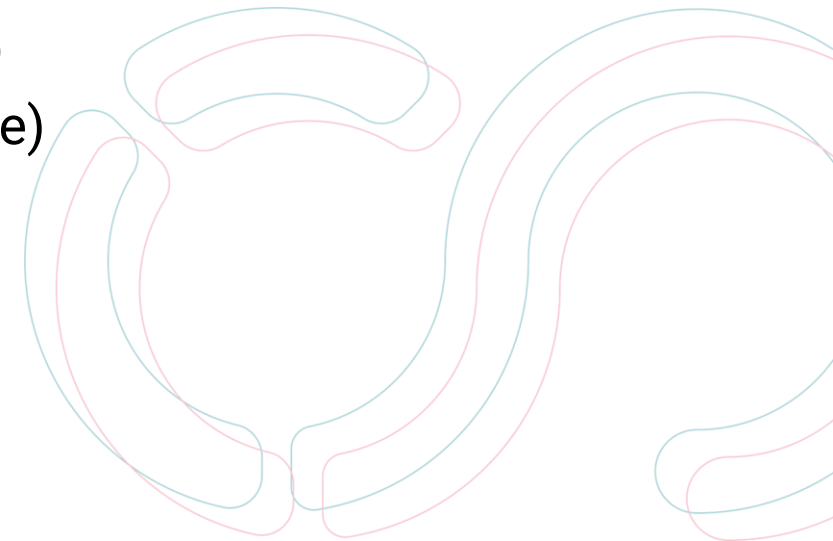
Why FAIR Data

- **F**indable, **A**ccessible, **I**nteroperable, **R**eusable



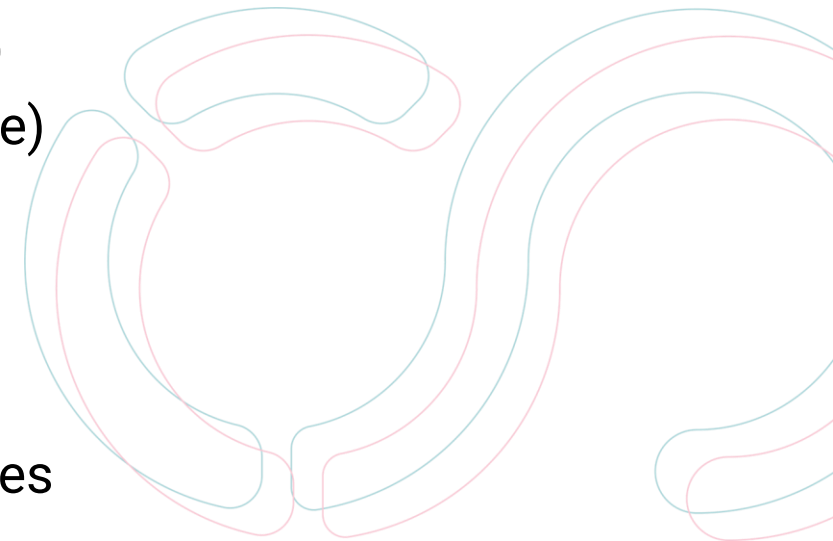
Why FAIR Data

- **F**indable, **A**ccessible, **I**nteroperable, **R**eusable
- **B**enefits:
 - You don't lose your data (Findable)
 - You will know where they are and how to get them (Accessible)
 - You have your data properly described/annotated (Interoperable)
 - You will be able to use them again (Reusable)



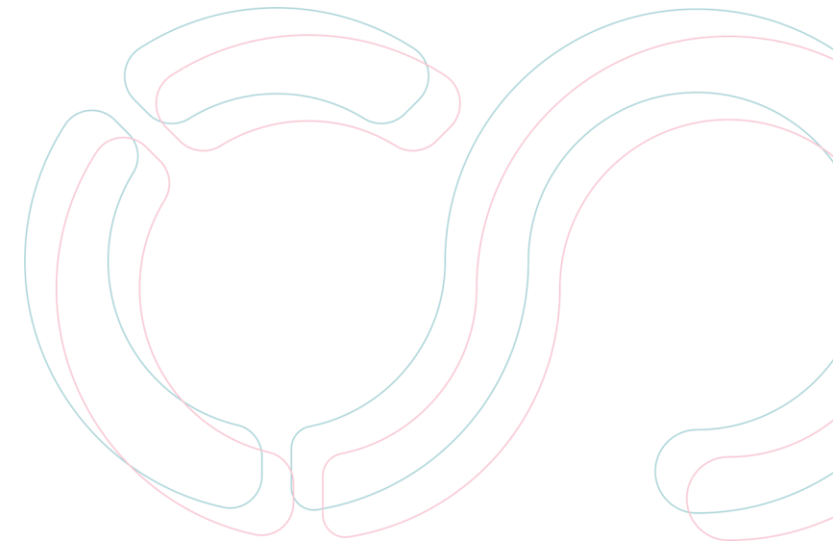
Why FAIR Data

- **Findable, Accessible, Interoperable, Reusable**
- **Benefits:**
 - You don't lose your data (Findable)
 - You will know where they are and how to get them (Accessible)
 - You have your data properly described/annotated (Interoperable)
 - You will be able to use them again (Reusable)
- **Additional benefits**
 - You can share such data without (much) work on your side
 - You can (easily) combine such data with data from other sources



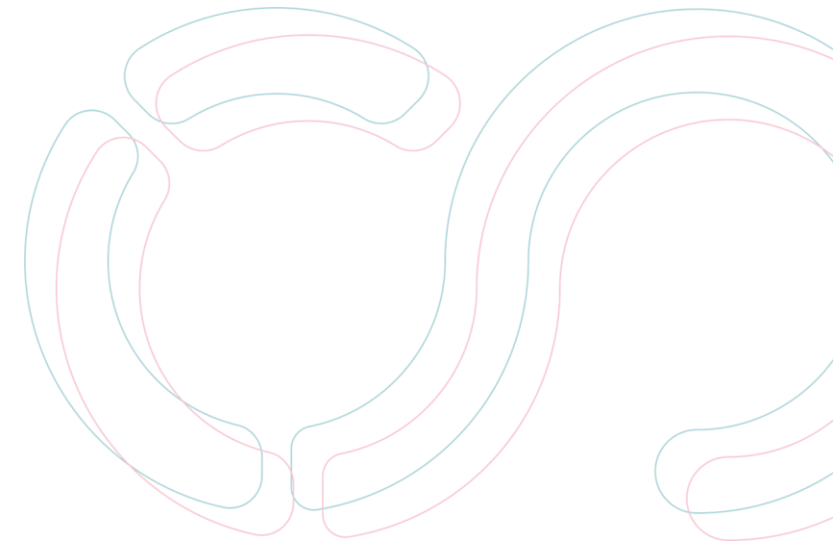
Why FAIR Data

- **Findable, Accessible, Interoperable, Reusable**
- **Benefits:**
 - You don't lose your data (Findable)
 - You will know where they are and how to get them (Accessible)
 - You have your data properly described/annotated (Interoperable)
 - You will be able to use them again (Reusable)
- **Additional benefits**
 - You can share such data without (much) work on your side
 - You can (easily) combine such data with data from other sources
- **And you can publish your data**



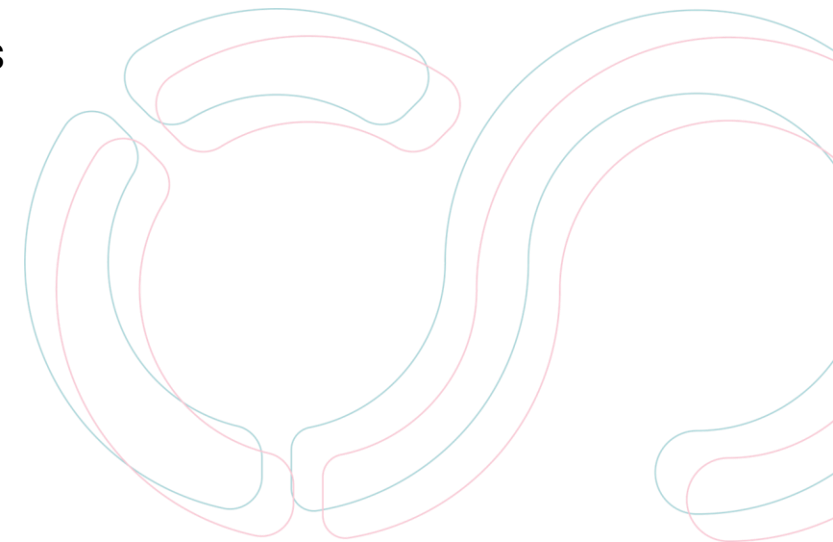
How is FAIR Data Management Supported?

- **Data Management Plans**
 - Creation and manipulation supported through Data Stewardship Wizard
 - We aim towards the actionable DMPs
 - These will directly manipulate the infrastructure to simplify data management
- **National Metadata Directory**
 - Starting point for data search
- **Access Management**
 - AAI to help defining who can have access and under which conditions
 - So not only “fully open” data supported
- **Licensing**
 - The list of potential licenses and selection support tool
 - Actionable selected licenses – the AAI will check the conditions



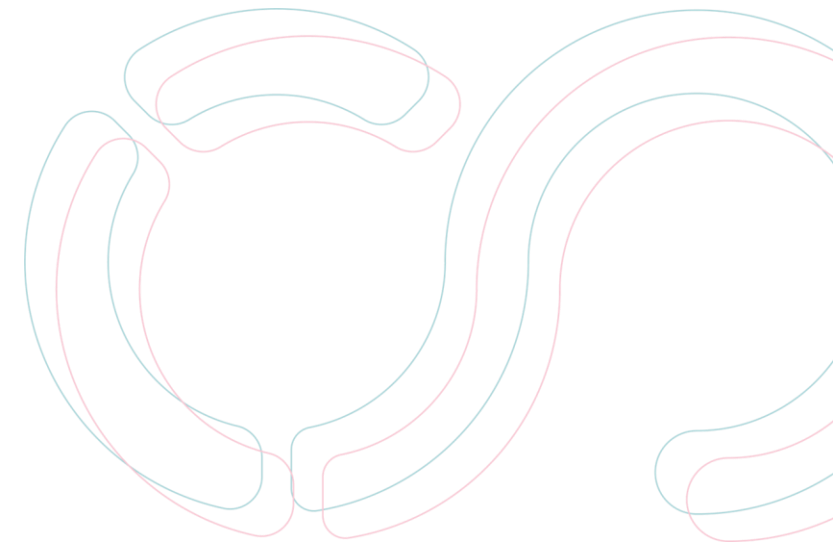
How is FAIR Data Management Supported?

- FAIRification tool(s)
 - To help make data FAIR
 - To help check what is still missing
- Repositories in the NRP
 - Storage for your data, with automatic connection to other components
 - You can either use a repository in NRP to store your datasets
 - Or you can create and maintain your own repository
 - Guaranteed immutability of individual datasets, protection from loss
 - Basis data processing integrated with NRP (containers, Kubernetes)
- Connection to instruments
 - Allow fast direct transfer from instrument into a repository



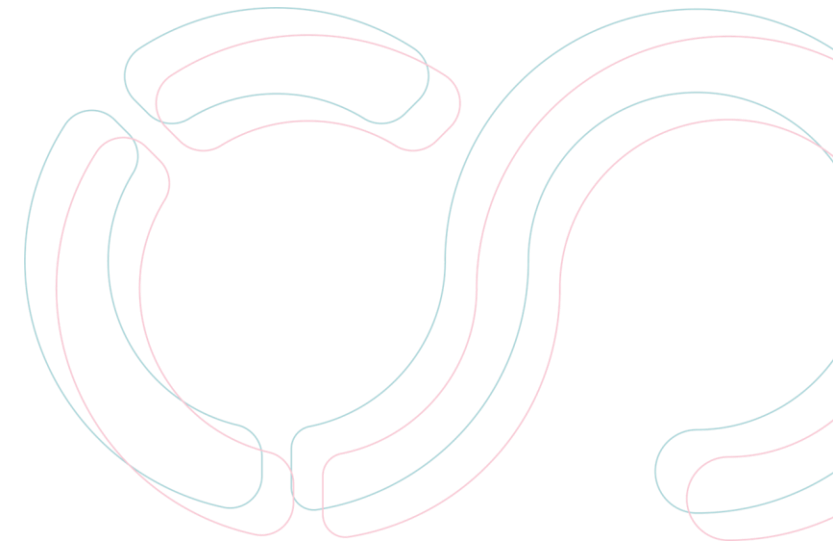
How is FAIR Data Management Supported?

- Provenance support (also 3rd A in **AAAI**)
 - Keep track what's happening, reproducibility support
- Connection with the e-INFRA CZ hot data, compute and analysis capabilities
 - Fast data transfer and use of data in repositories
 - Storage for data combination and processing
 - Workflow support
- Processing of sensitive (human) data
 - SensitiveCloud and other Trusted Research Environments
- Secure and compliant ecosystem
- Support, Training, Education



Collaboration – another dimension of How

- **Czech EOSC implementation supported by a series of collaborative projects**
 - A building of an infrastructure is not purely competitive
 - What to do with losers?
 - Also Open Science is about collaboration and knowledge sharing
- **Large, reasonable long projects**
 - To support cooperation
 - To create trust among partners but especially for users
- **Does not mean there is no excellence**
 - A bottom up approach to build an ecosystem

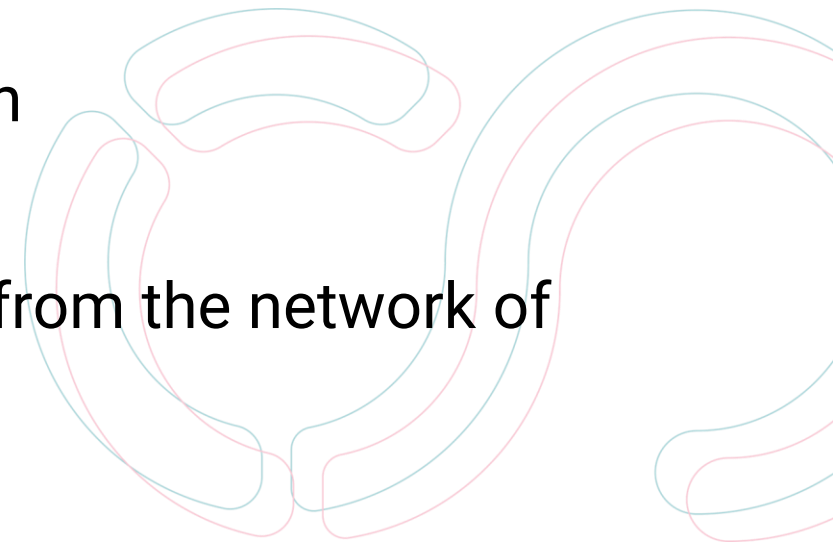


European Dimension

- EU started with EOSC around 2016
- Dozens of projects
 - Also EOSC partnership within HEU
- Formal setup – EOSC Tripartite
 - European Commission representatives (DG CONNECT and DG RESEARCH)
 - EOSC Steering Board (Ministry nominated national representatives)
 - EOSC Association (to represent the community of users and providers)
- Strategic Research and Innovation Agenda (SRIA) – version 1.2
 - Multi-Annual Roadmap (MAR)

EOSC Federation

- Recent development
- EOSC EU Node
 - Result of procurement by EC
 - Offer hub/like services (e.g. centra AAI hub)
 - Limited storage and processing capabilities (e.g. Jupyter Notebooks)
- Other EOSC nodes currently under discussion/evaluation
 - National Nodes – resembling the EOSC EU Node
 - Thematic Nodes – providing access to data and specific tools
- The idea is to create an EOSC infrastructure composed from the network of EOSC Nodes
 - Remember “EOSC is a web of data and services”



EOSC Federation – New Challenges

- Not yet properly defined interaction with large research infrastructures
 - Should any ESFRI become an EOSC (Thematic) Node?
 - How to do with branding?
- Unclear interaction with resource providers
 - E.g. should EuroHPC centers become specific EOSC Nodes?
 - What about EGI.eu, GEANT, ... (and at national level MetaCentrum)?
- An EOSC Handbook under preparation to look for answers

Summary

- EOSC Implementation
 - An opportunity as well as a challenge
- Czech implementation deliberately narrow
 - Focus on FAIR data management support (in an extensive way)
 - Strong relationship/integration with e-INFRA CZ
 - Emphasis on training and general awareness/knowledge transfer
- Wider involvement and support by researchers
- An opportunity to improve how we deal with research data -> improvement in science we do



Thank you for your attention

