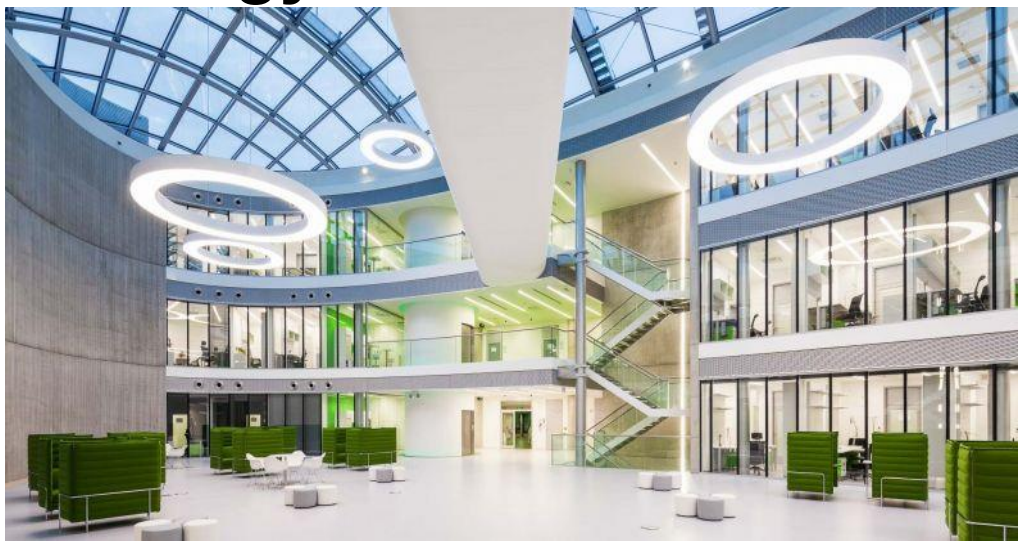
Enhancing Data Accessibility: Strategies and Tools for Effective (Meta)data Management *in life-sciences*

Adrián Rošinec, Tomáš Svoboda

IBERGRID 2023, 28.9.2023, Benasque

CEITEC

- **Central European Institute of Technology**
- Established 2011
- Fields of science centre
  - Life sciences
  - Advanced materials
  - Nanotechnology
- Establishing institutions (Brno universities/institutes)



CEITEC in numbers (2022)

- 28 research groups
- **13 core facilities (CF)**
  - 720 users (from 22 countries)
- 270 research FTEs
- 241 publications (108 Q1, 24 T5)
- Ph.D. program (212 students)

Core facilities

MUNI MU

Cryo-Electron Microscopy and Tomography Core Facility


Jiří Nováček, Ph.D.

MUNI MU

Josef Dadok National NMR Centre


Radovan Fiala

MUNI MU

Proteomics Core Facility


Prof. Zbyněk Zdráhal

MUNI MU

Biomolecular Interactions and Crystallography Core Facility


Josef Houser, Ph.D.

MUNI MU

Nanobiotechnology Core Facility


Jan Přibyl, Ph.D.

MUNI MU

Multimodal and Functional Imaging Laboratory


Michal Mikl, Ph.D.

MUNI MU

Genomics Core Facility


Boris Tichý, Ph.D.

MUNI MU

Cellular Imaging Core Facility


Milan Ešner, Ph.D.

MUNI MU

Plant Sciences Core Facility


Natallia Madzia Valasevich, Ph.D.

MUNI MU

Bioinformatics Core Facility


Vojtěch Bystrý, Ph.D.

MUNI MU

Biological Data Management and Analysis Core Facility


Radka Svobodová, Ph.D.

What is role of CF

- Almost 700 special instruments
  - Cutting-edge technologies and instruments
  - MRIs, CryoEM, Xray Crystallography, …
- Provide services to research groups, scientists
  - **Conduct experiments and provides data**
  - accessible to internal/external users from academy and industry at national and international levels

Data management in CEITEC

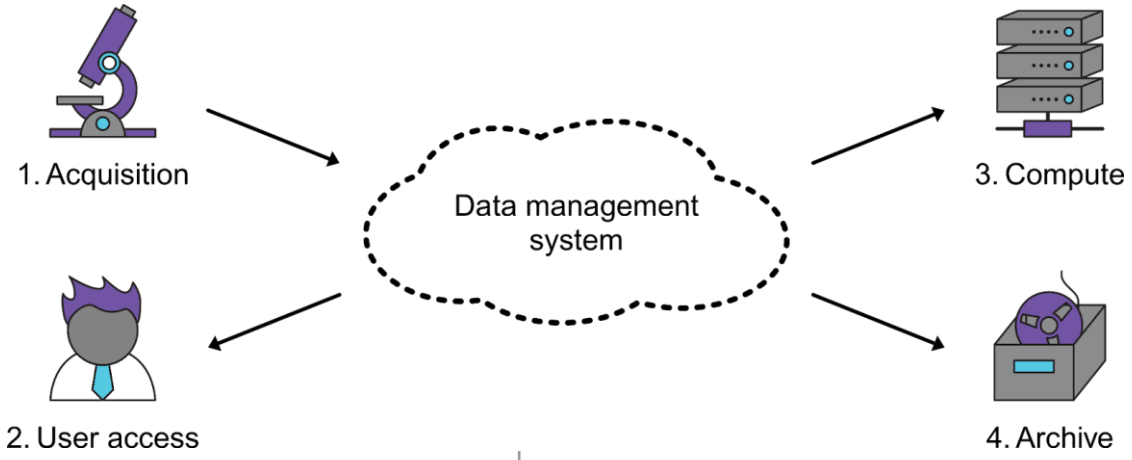
- Wide spectrum of produced data types
  - Big – CryoEM images, very small – CSVs, sensitive data, …
- Currently no unified solution
  - No central data repository or storage facility
  - CFs/Labs may have large storage solutions or NAS or only laptop disk as storage solution
- Some CFs don’t manage data
  - Passing data to the users using USB disks, CDs, random sharing services
- Institutional agreement to provide *unified solution across all CFs*

MUNI | CEITEC

Data management system

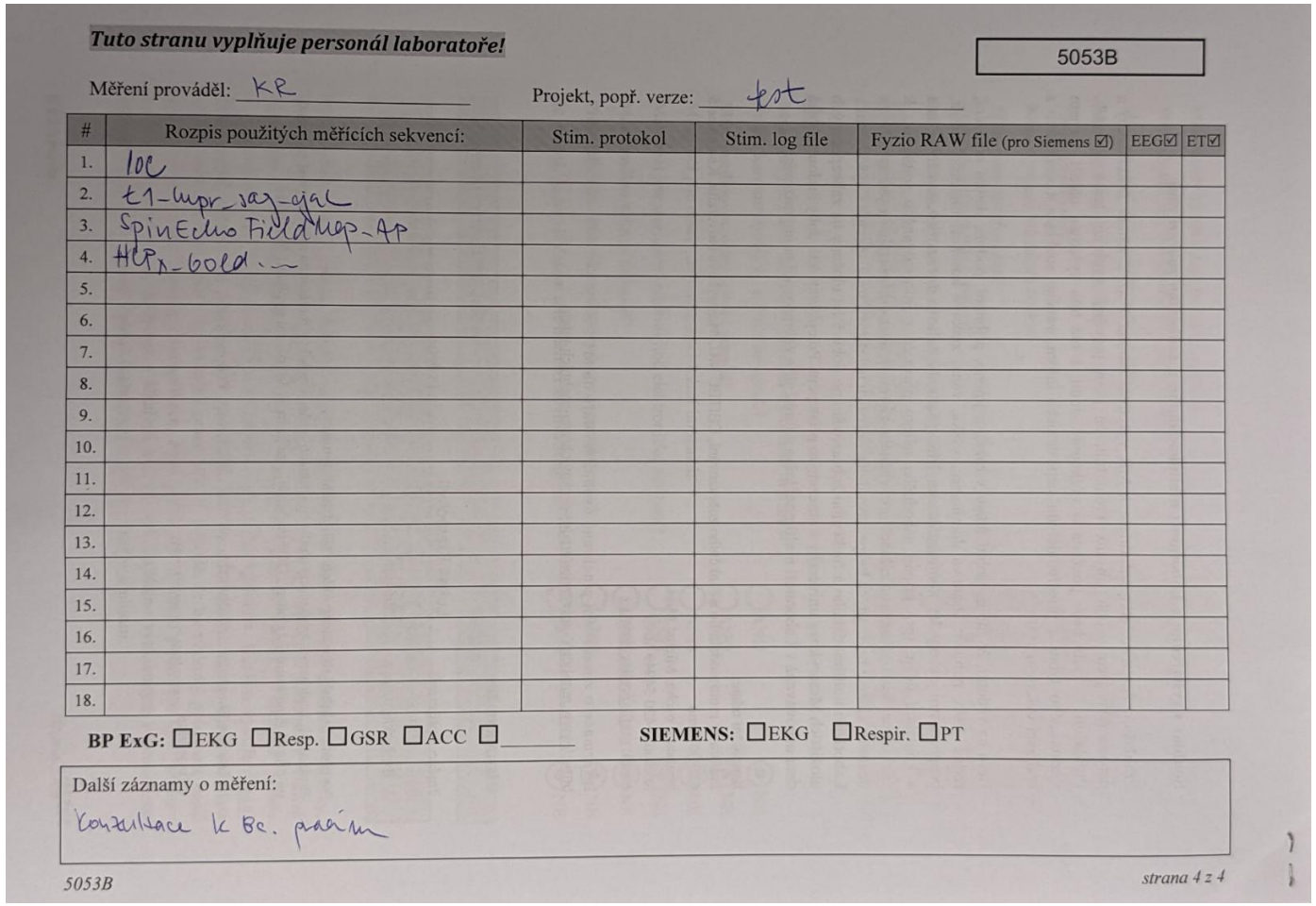
- Requirements for DMS
  - Handle data acquisition – from instruments
  - Run data workflows (metadata ext., preprocessing)
  - Ability to provide user access and sharing
  - Ability to prepare dataset for publication
  - Guarantee provenance
  - Archiving and long-term preservation
  - Mounting to processing / collaboration tools
- All this should be as

  “automated” as possible



MUNI | CEITEC

Metadata - integral part of each dataset



Tuto stranu vyplňuje personál laboratoře!

5053B

Měření prováděl: KR Projekt, popř. verze: test

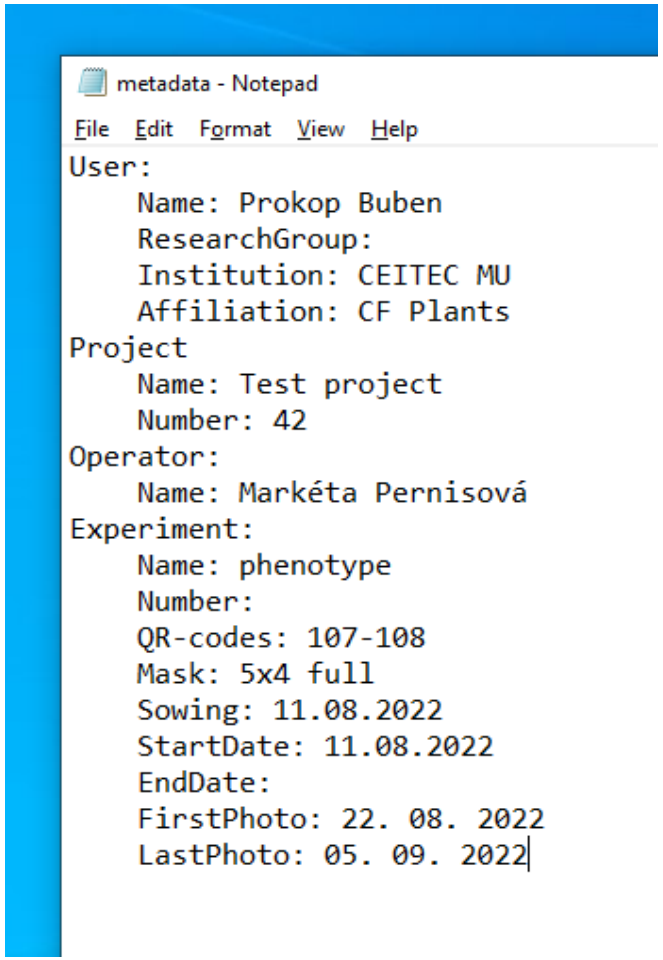
#	Rozpis použitých měřících sekvencí:	Stim. protokol	Stim. log file	Fyzio RAW file (pro Siemens ☑)	EEG☑	ET☑
1.	loc					
2.	t1-mpr_sag-gar					
3.	SpinEcho FieldMap_AP					
4.	HCP_A-bold_...					
5.						
6.						
7.						
8.						
9.						
10.						
11.						
12.						
13.						
14.						
15.						
16.						
17.						
18.						

BP ExG: ☐EKG ☐Resp. ☐GSR ☐ACC ☐_____ **SIEMENS:** ☐EKG ☐Respir. ☐PT

Další záznamy o měření:

Konzultace k Bc. práci

5053B *strana 4 z 4*



MUNI | CEITEC

Why to bother with metadata?

- Metadata help to understand why datasets was created
  - context/equipment/method/source – patient
- Right metadata ontology helps with interoperability
  - Datasets might be reused within the community
- Finding datasets
  - Functions like searching/filtering/categorizing/identification
- Provenance – how the dataset was created
- Licensing

Obtaining metadata

- Manual annotations
  - Human labor, error-prone, but very popular
  - Goal to „ make annotating work enjoyable“ – nice GUIs, autocomplete, validators, context recognition
- Automated annotations
  - The base are metadata from instrument settings, administrative ones (owner of dataset), information about experiment, ...
  - e.g. image recognition, capture/filter outputs from analytics tools and simulations
- Lab entries and lab notebooks

Example: Extraction of metadata from molecular dynamics

- Created parser of output from `gmx dump`
- Could be placed as a step of data workflow
  - Simulations ends -> output is stored on storage -> took by the DMS -> run preprocessing -> creating dataset
- Parsing on specific keywords and values important for dataset findability
  - based on the specification from the subset of the community
  - this community need accepted standard
- Results are outputted in json/yaml format stored in DB

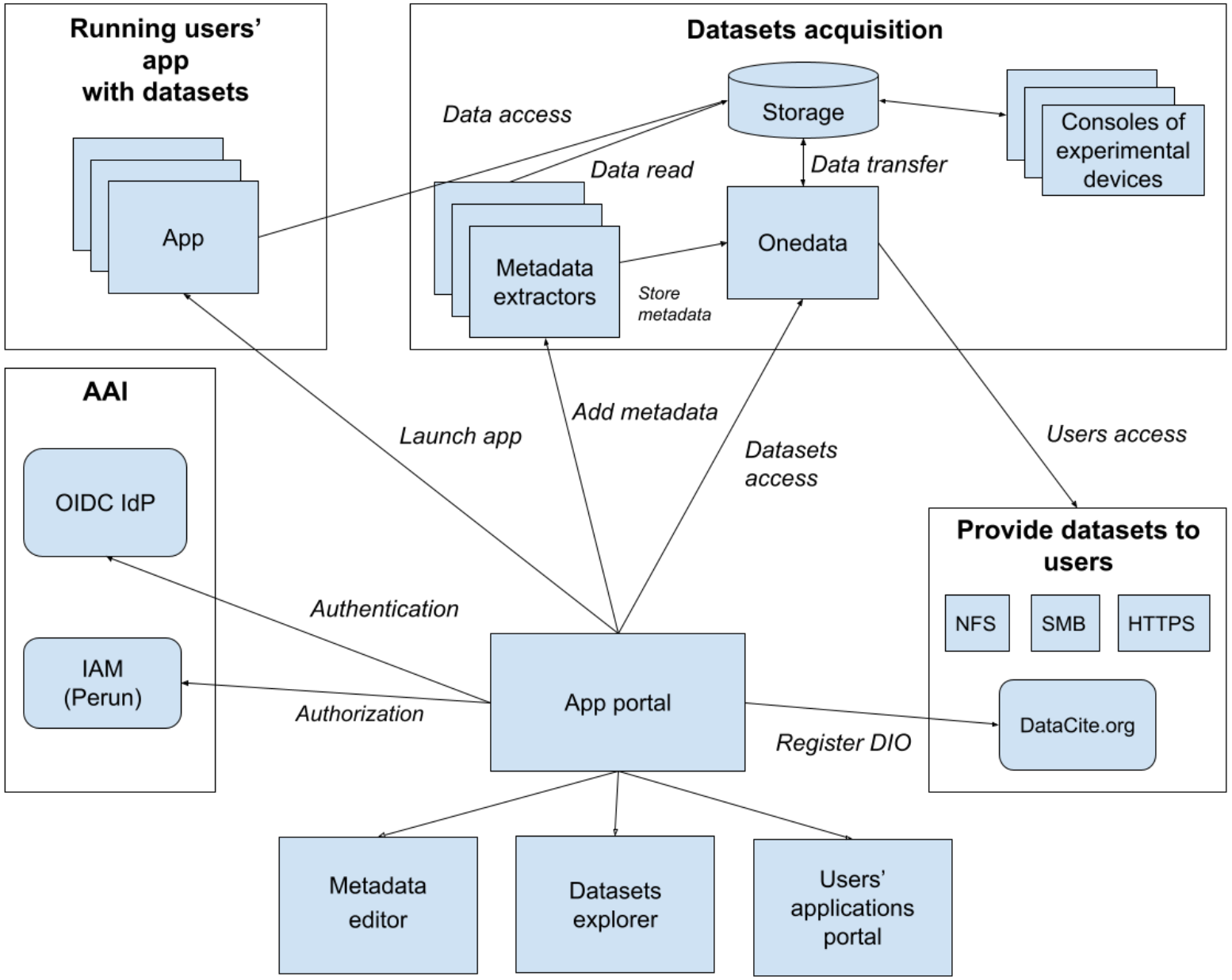
Metadata standards and ontologies

- Making CEITEC compliant with the world
- Community schemas
  - Some communities have rich ontologies – EDAM, Open Microscopy, EMPIAR
  - Some communities don’t have anything
    - e.g. Molecular dynamics, plants/crops experiments
    - Motivate and provide support to establish partnership across CZ and for now Max Planck Institute
- Administrative and technical schemas (for CEITEC/National e-infrastructure)
- Institutional and national requirements for datasets
  - Administrative metadata for funders (project IDs, …)
  - Part of EOSC CZ initiative

Data management platform

- Aim to provide user interface to support data management life-cycle
- Ability to manage datasets self-service in a transparent manner
  - hide complexity of backend services
- Interconnecting many backend services
  - Onedata
  - Raw storage system
  - Metadata catalogue
  - Project control and IAM
  - DataCite
  - Public (meta)data repositories
  - Metadata schemas repositories and editors
  - FAIR checkers

MUNI | CEITEC


Running users' app with datasets
App
Datasets acquisition
Storage
Consoles of experimental devices
Data access
Data read
Data transfer
Metadata extractors
Onedata
Store metadata
AAI
OIDC IdP
IAM (Perun)
Launch app
Add metadata
Datasets access
Users access
Authentication
Authorization
App portal
Provide datasets to users
NFS
SMB
HTTPS
DataCite.org
Register DIO
Metadata editor
Datasets explorer
Users' applications portal

Challenges

- Make the solution accepted at least nation-wide
  - We have support of national e-infrastructure
- Development costs and sustainability of a project
  - Making use of what's available, minimum coding