MUNI
ICS

Podrobnější představení implementace ukládání dat v jednotlivých CEITEC CF

Tomáš Svoboda

O mně

- Studium FI MU (2009 – 2014)
- Studium ÚSI VUT (2014 – 2016)
- Centrum dopravního výzkumu (2014 – 2017)
- CESNET – MetaCentrum (2015 – 2023)
  - Podpora uživatelů - výzkumné spolupráce
- ÚVT MU (od 2017)
  - Vývojář IT
- Doktorské studium PřF MU (od 2023)
- CESNET – Datová úložiště (od 2023)
  - Systémy pro správu dat

Jak to začalo?

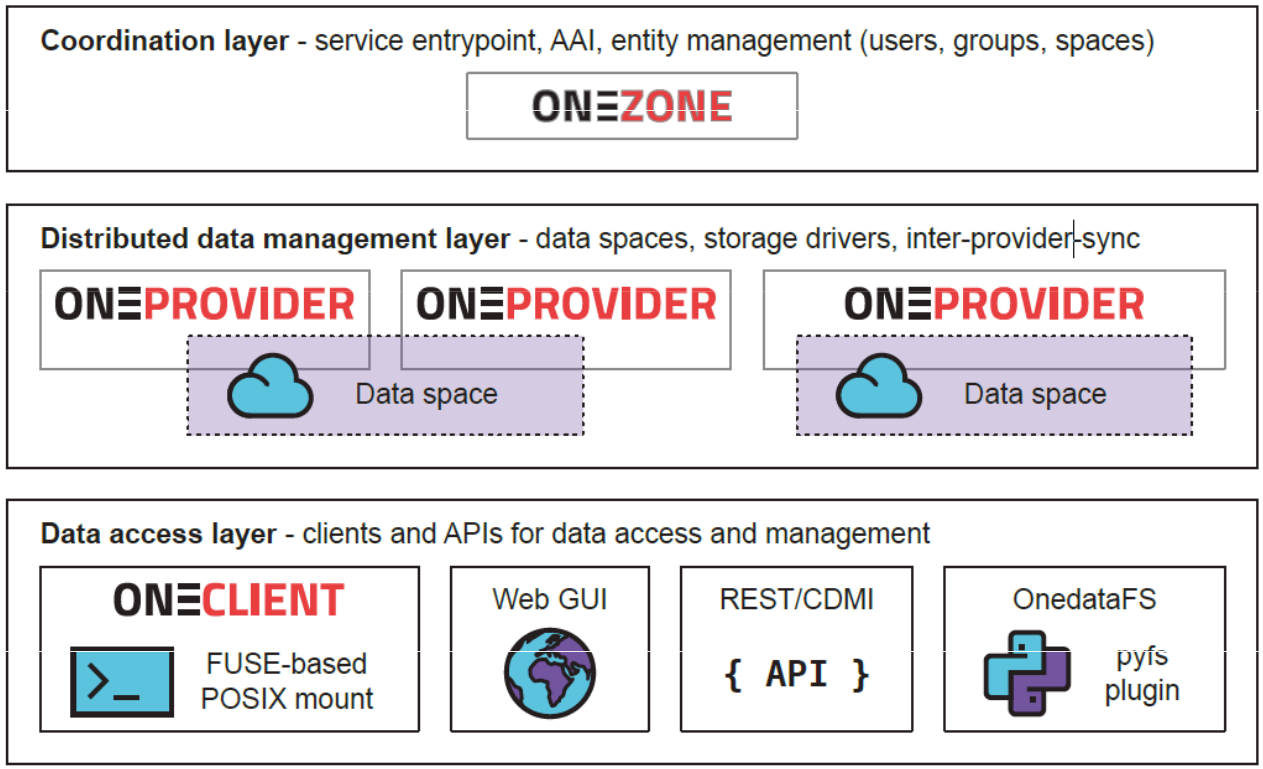
CNB
CENTRO NACIONAL DE BIOTECNOLOGIA

CSIC
CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS

CEITEC

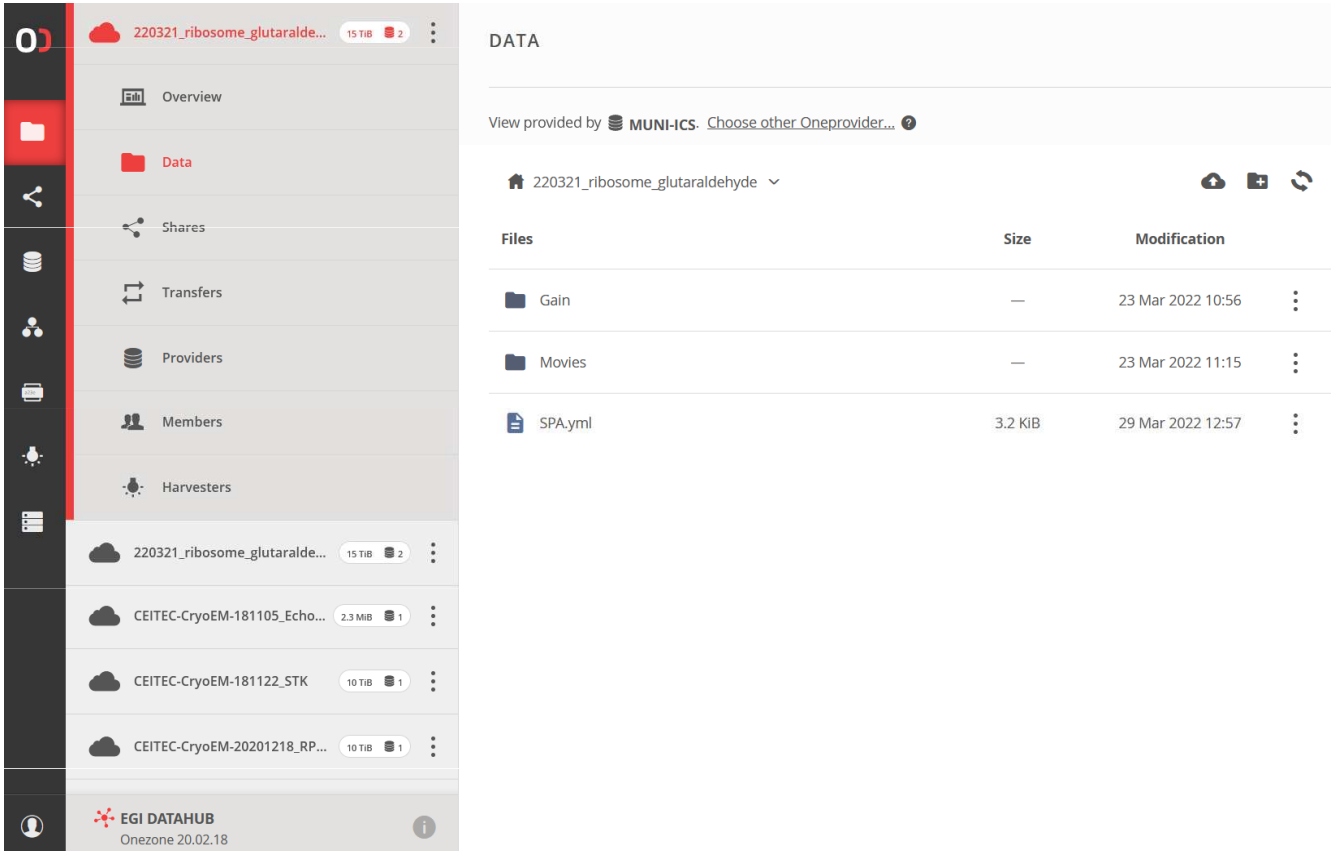
Onedata

- Systém pro globální správu a přístup k datům
  - ukládání, sdílení, archivace a publikace
- Přístup k datům jako známá cloudová řešení (Dropbox, …)
- Uzpůsobeno pro vědecké prostředí
  - Podpora FAIR principů
  - Podpora HPC prostředí
- Úložné technologie
  - POSIX, S3, Ceph, …



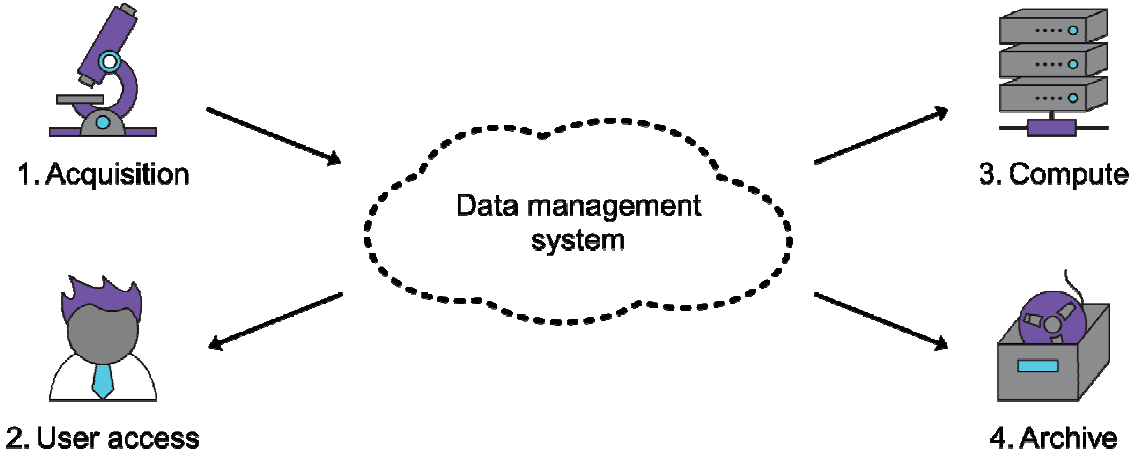
Zkušenosti

- Negativní proroci
- Instalace
- Dokumentace
- Navázání kontaktu s autory

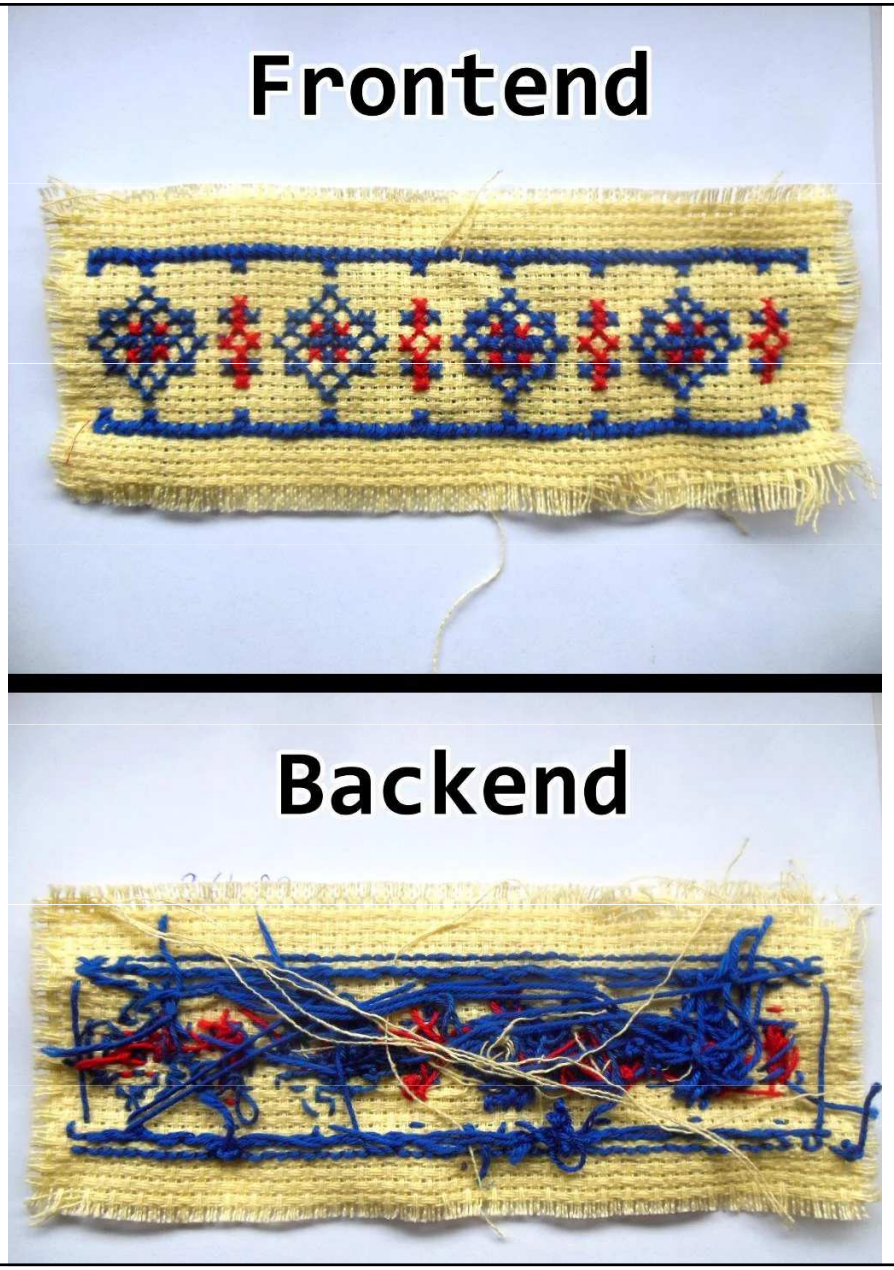


Aplikace fs2od

- Automatizovaný workflow životního cyklu datové sady
- Konfigurovatelné možnosti
  - Načítání metadat
  - E-mailové notifikace
  - Replikace (přiblížení dat k výpočtům)
  - Archivace (pro dlouhodobé uložení)
  - Mazání (z primárního úložiště)

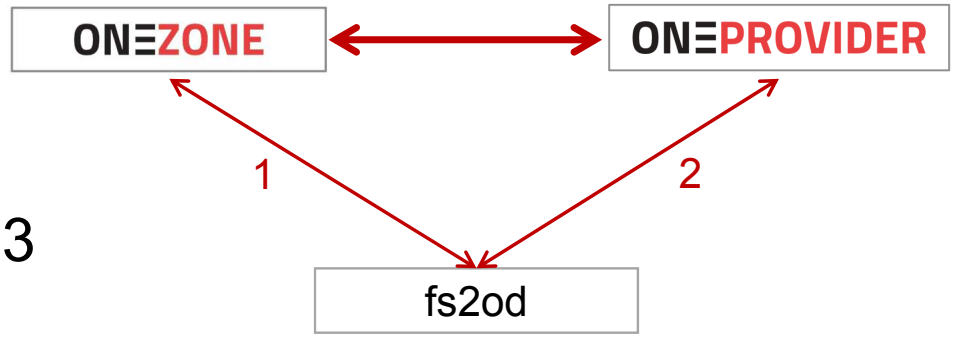


Realita nebyla tak růžová



Zádrhel #1

- Vzrůstající pravděpodobnost výskytu chyby při alokaci úložiště pro datovou sadu (300+)
- Debuggování
- Nedůvěra, ze strany autorů
- Problém při interní komunikaci v rámci Onedata
- Nelineární složitost ověřování konzistence dat v Onezone



- Fix leden 2023

Zádrhel #2

- Náhodné zaseknutí registrace existujících souborů (+1000)
- Space ID (`8cd781e98253ccb7f632b1b952314937ca2097`)
  - „Náhodně generované“
  - V interním plánovači se pracuje jen 8 B prefixem, což je ale ve skutečnost 4 B
  - Při kolizi zastavena akvizice dat pro všechny datové sady
  - Nástroj tmate

- Fix srpen 2023

Zádrhel #3

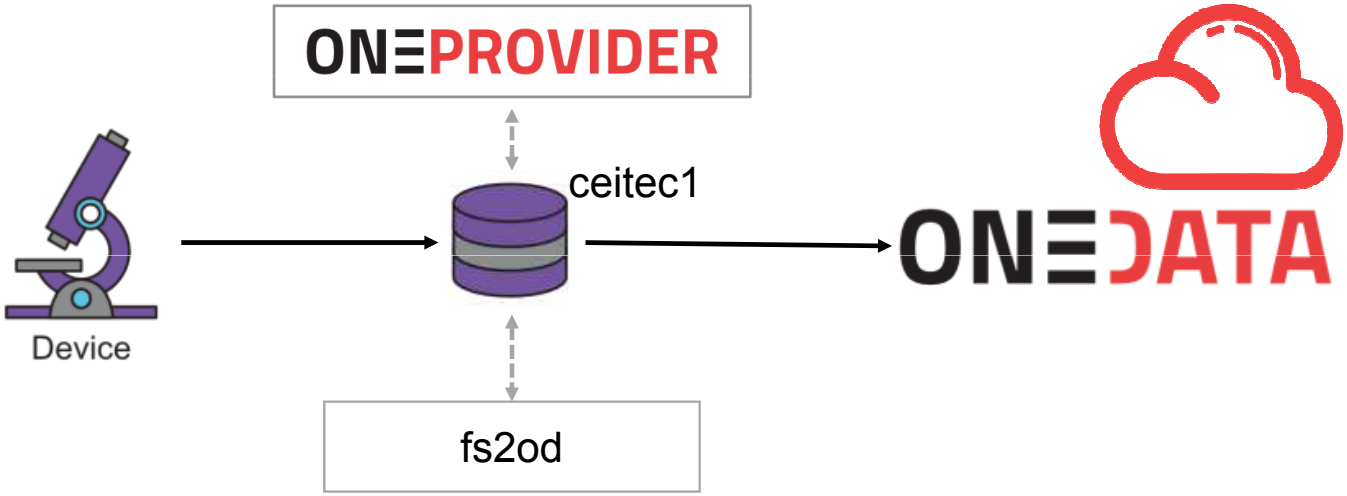
- Webové GUI nepřipravené pro tak velké množství datových sad (800+)
- Pracujeme přes API, ale GUI se používá při ručních zásazích a kontrole
- Při velkém počtu datových sad selže načtení celé stránky

- Fix květen 2023 (částečně)

CEMCOF

CEMCOF

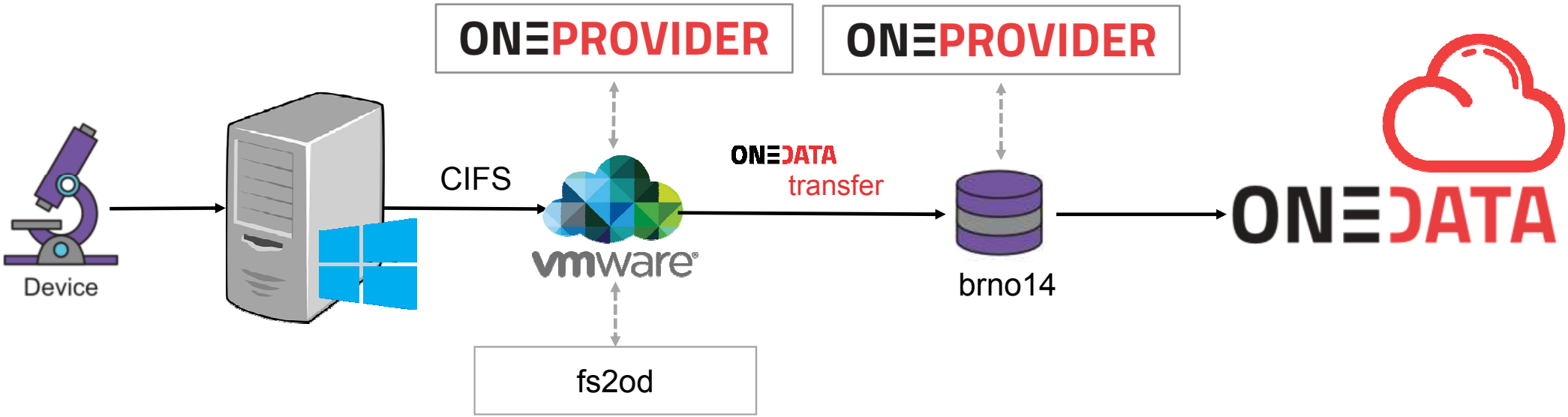
– Vlastní velké diskové pole, Oneprovider přímo na čelním uzlu



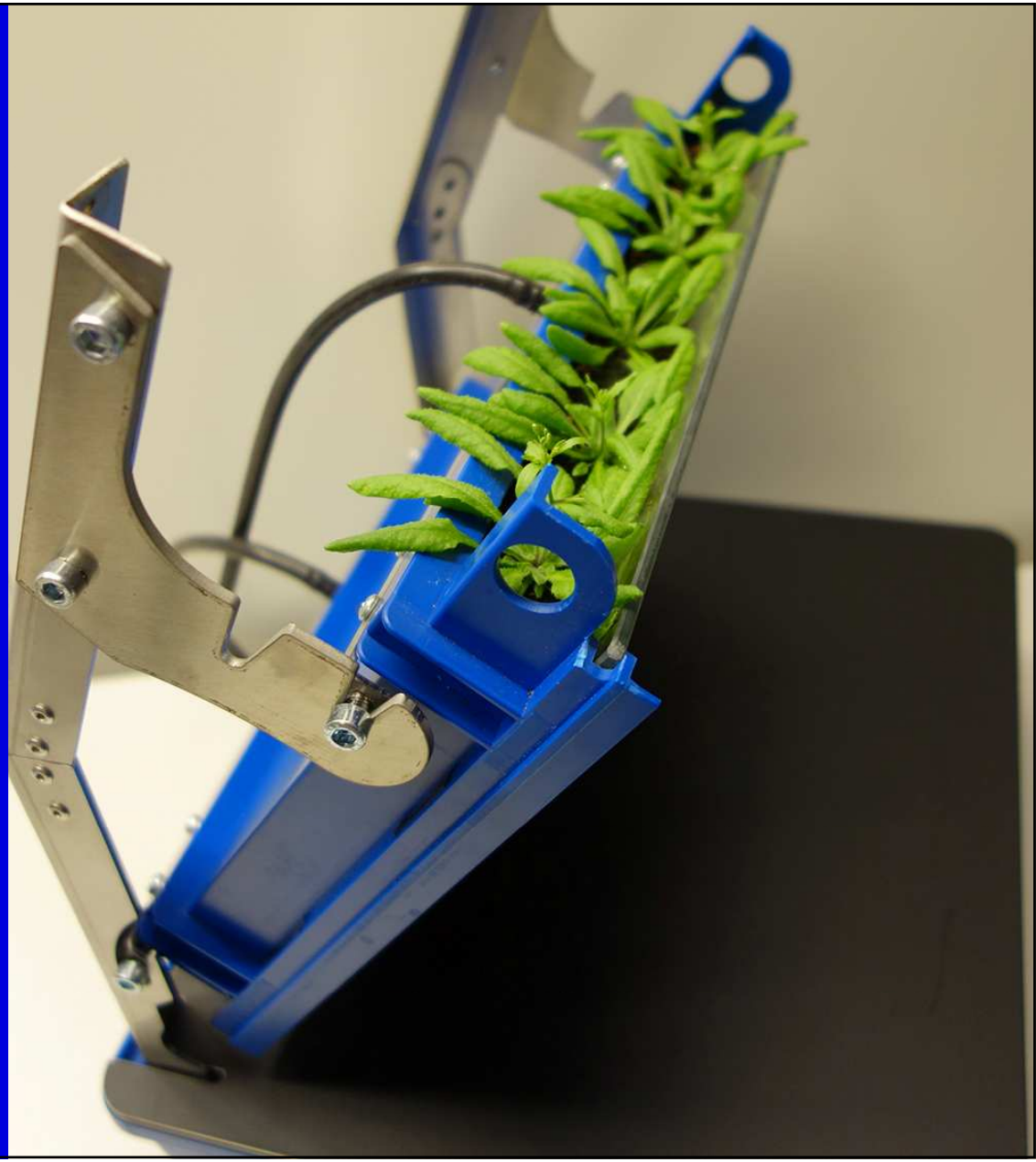
CF CELLIM

CF CELLIM

- Windows Server
- Zpracování a přechodné uložení dat

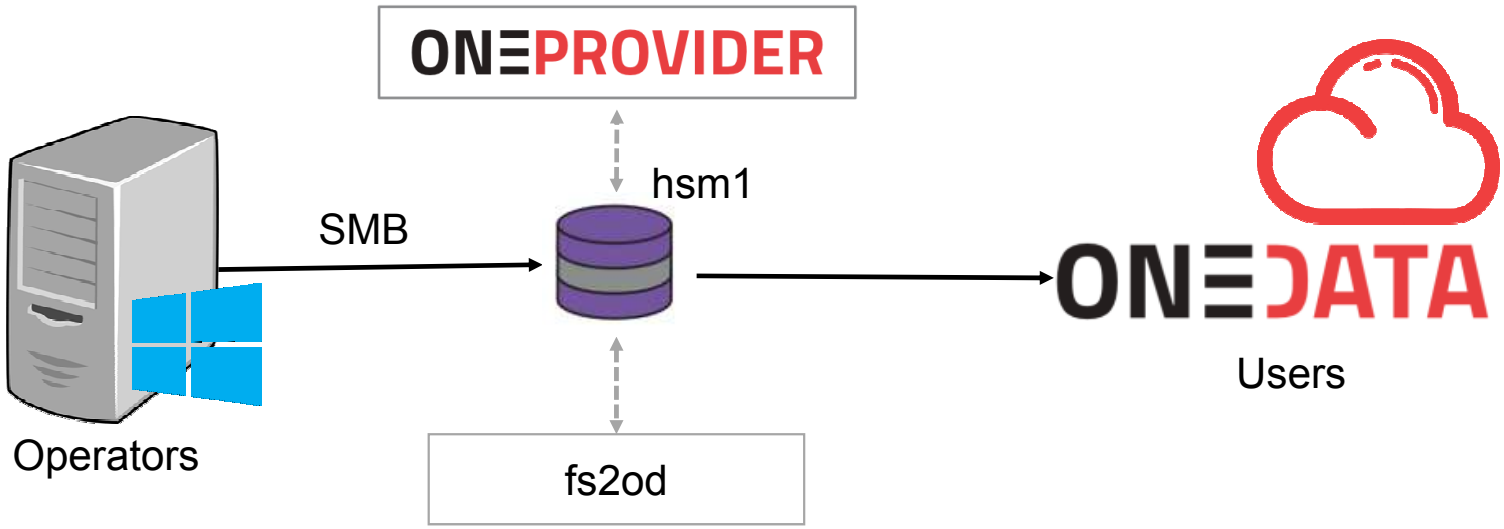


CF PLANTS

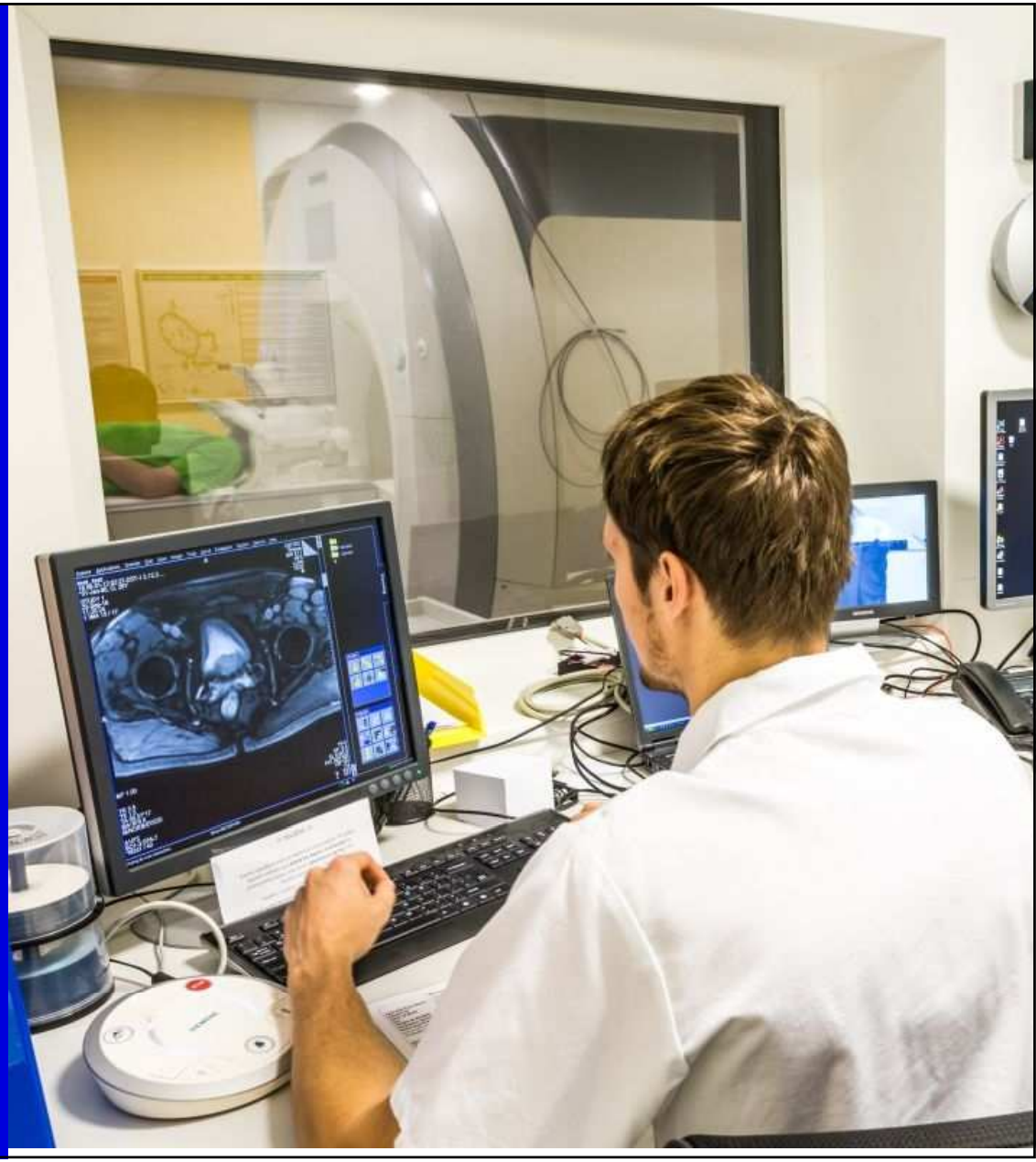


CF PLANTS

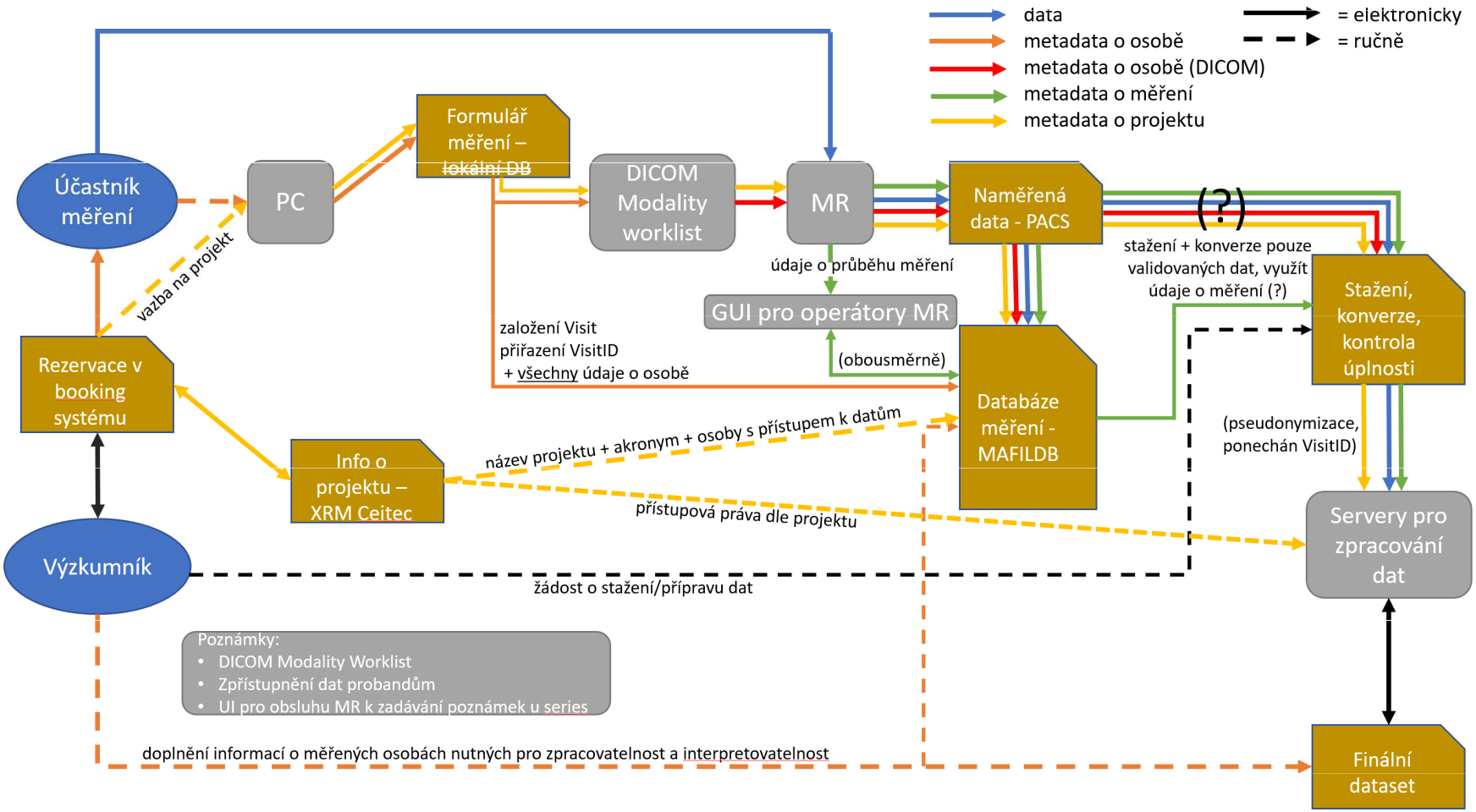
- Předání výsledků uživatelům
- Úvahy o repozitáři



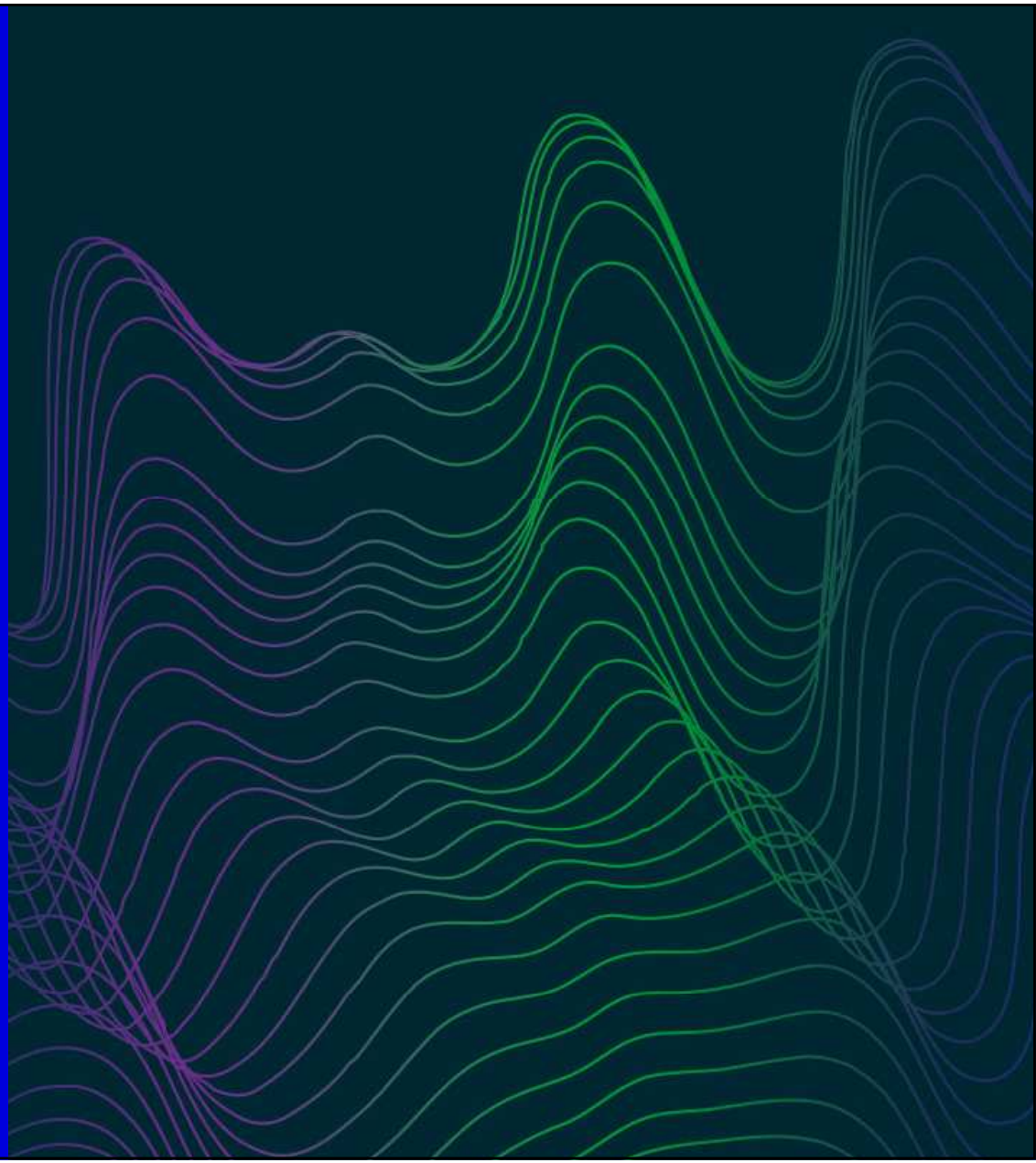
CF MAFIL



CF MAFIL


data
metadata o osobě
metadata o osobě (DICOM)
metadata o měření
metadata o projektu
= elektronicky
= ručně
Účastník měření
PC
Formulář měření – ~~lokální DB~~
DICOM Modality worklist
MR
Naměřená data - PACS
(?)
stažení + konverze pouze validovaných dat, využít údaje o měření (?)
Stažení, konverze, kontrola úplnosti
údaje o průběhu měření
GUI pro operátory MR
(obousměrně)
založení Visit
přiřazení VisitID
+ všechny údaje o osobě
vazba na projekt
Rezervace v booking systému
Databáze měření - MAFILDB
(pseudonymizace, ponechán VisitID)
Info o projektu – XRM Ceitec
název projektu + akronym + osoby s přístupem k datům
přístupová práva dle projektu
Servery pro zpracování dat
Výzkumník
žádost o stažení/přípravu dat
Poznámky:
• DICOM Modality Worklist
• Zpřístupnění dat probandům
• UI pro obsluhu MR k zadávání poznámek u series
doplnění informací o měřených osobách nutných pro zpracovatelnost a interpretovatelnost
Finální dataset

Projekt RELIEVE



Projekt RELIEVE

- Výzkum neurologických nemocí
- Návrh datového workflow (EEG data)

MUNI
ICS

Maj data bez metadat smysl?

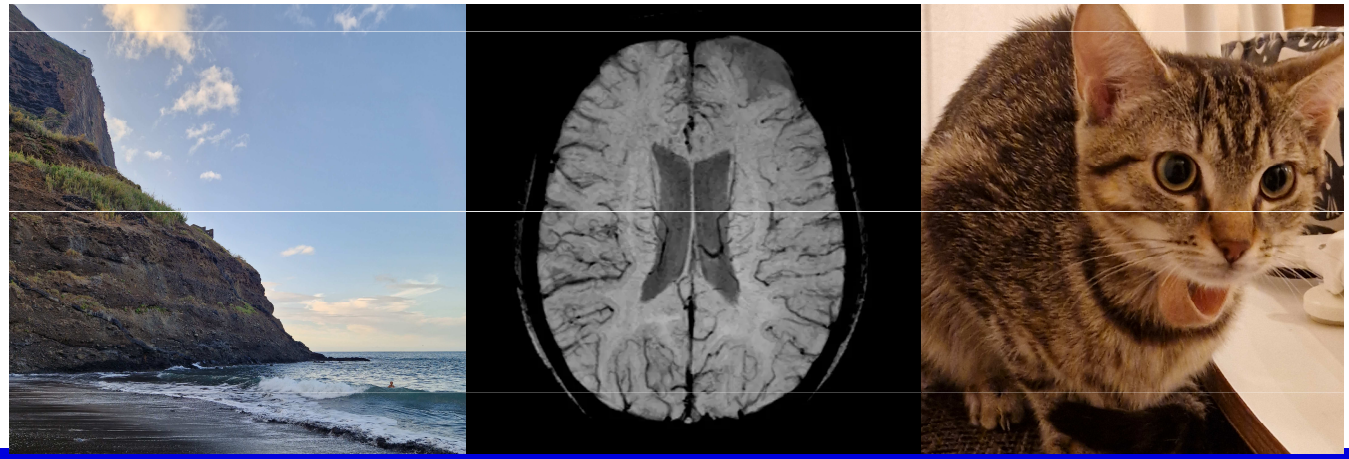
Adrián Rošinec

O mně

- Studium FI MU (2017 – 2022)
- ÚVT MU (od 2018)
  - WinAdmin, identity, cloud, e-infra, coffee, ...
- Doktorské studium PřF MU (od 2023)
- ☺

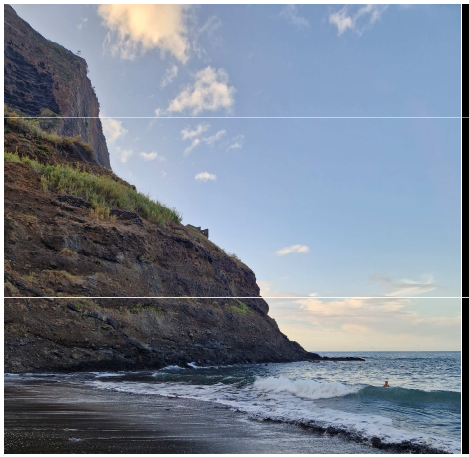
Co jsou data?

- Súbor informácií, faktov
- Reprezentované v štrukturovanej alebo neštrukturovanej podobe
- Napr. obrázok, zvuková stopa, tabuľky, ...



A metadata?

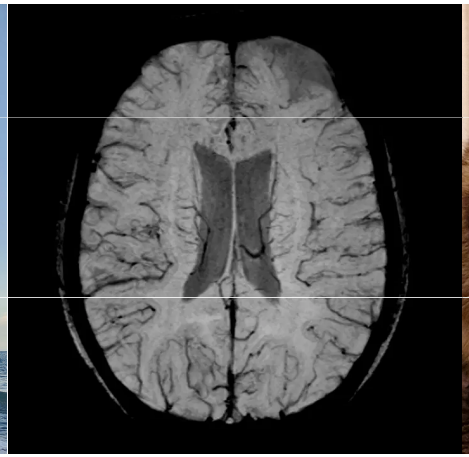
- Doprovodné informácie
- Pomáhajú pochopiť kontext a detaily datasetu
  - Dátum vytvorenia
  - Veľkosť
  - Lokalita
  - Typ/formát



Creation: 07.09.2023

Device: Samsung SM-S906B
f/1.8 1/320 ISO 50 No-Flash

Location: Praia do Faial, Faial, PT



Creation: 17.04.2019

Gender: M

Handedness: Left

SliceLocation: 40.115494018811

SliceThickness: 1.5

EchoTime: 3.13

NumberOfAverages: 4

Creation: 19.02.2023

Device: Samsung SM-S906B
f/1.8 1/25sec ISO 640 No-Flash

Location: Kavárna Pelíšek, tř. Kpt. Jaroše, Brno, CZ

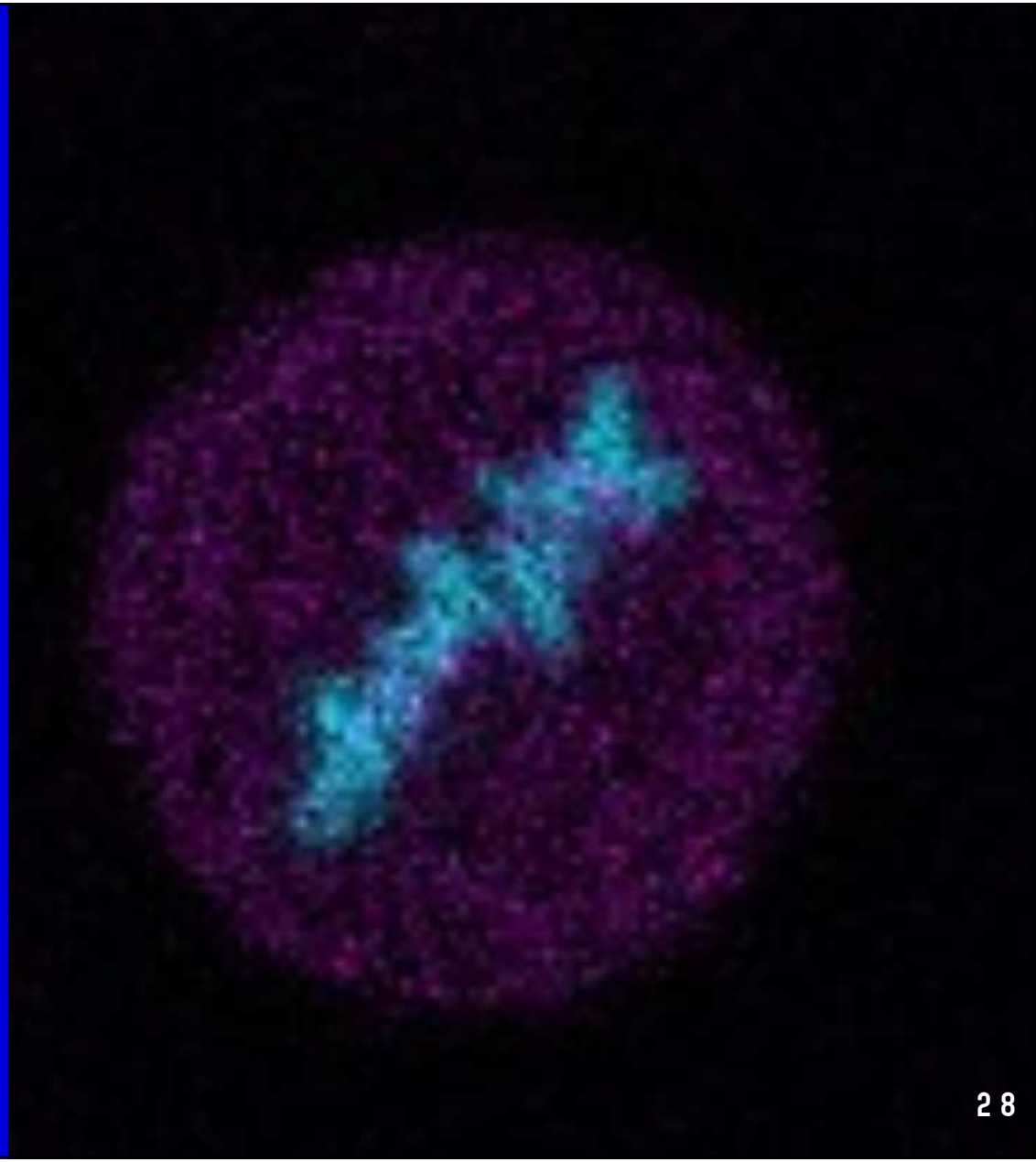
Prečo chceme anotované datasety?

- Zlepšenie prístupu k dátam
  - vyhľadávanie, filtrácia, kategorizácia, identifikácia datasetov
- Pomáha pochopiť kontext vzniku datasetu
  - Ako a prečo dataset vznikol
  - Aká metóda bola využitá pre získanie dat
    - Prístroj (spektrometer/mikroskop/MRI) a jeho parametre
  - Zdroj dát
    - Pacient a jeho diagnóza
- Podporuje interoperabilitu
  - Umožnuje výmena datasetov a využitie inými vyskumníkmi
- Reprodukovatelnosť
- Provenance
- Licensing

Prečo chceme anotované datasety?

- Umožňuje nám kontrolovať kvalitu datasetov, prípadne vedeckých výstupov v čase
  - Vďaka agregáciám a štatistickým prehľadom

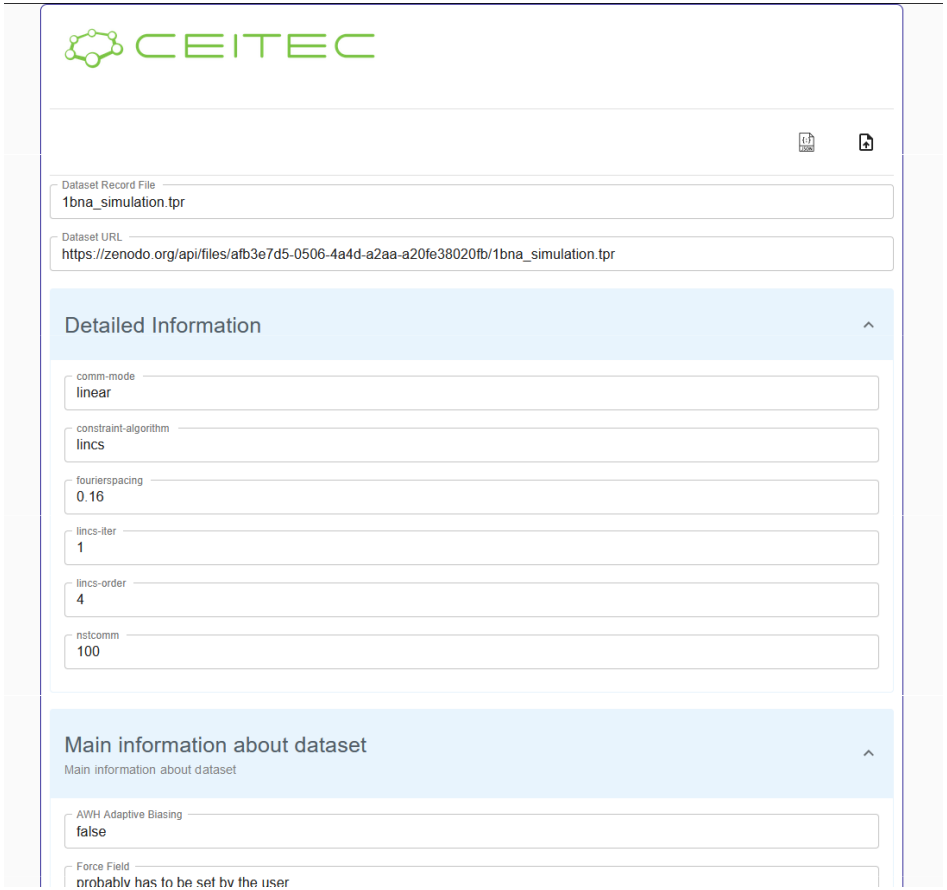
Príklad z praxe



Ako získavame metadáta

- Manuálne anotácie
  - Veľmi pracne s vysokou mierou chybovosti, populárne
  - Cieľ „spríjemňovať“ život anotátorom našeptávačmi, validátormi, rozpoznaním kontextu
- Automatické anotácie
  - Veľmi častý základ sú metadáta z nastavenia prístroja, administratívne metadáta (kto dataset vyrobil), ...
  - Adaptácia analytických nástrojov pre výstup do metadatového súboru
  - Zo štrukturovaných/neštrukturovaných dát potreba vyberať dôležité údaje
    - parametre simulácie, rozpoznávanie v obrázkoch, ...

Editor metadát



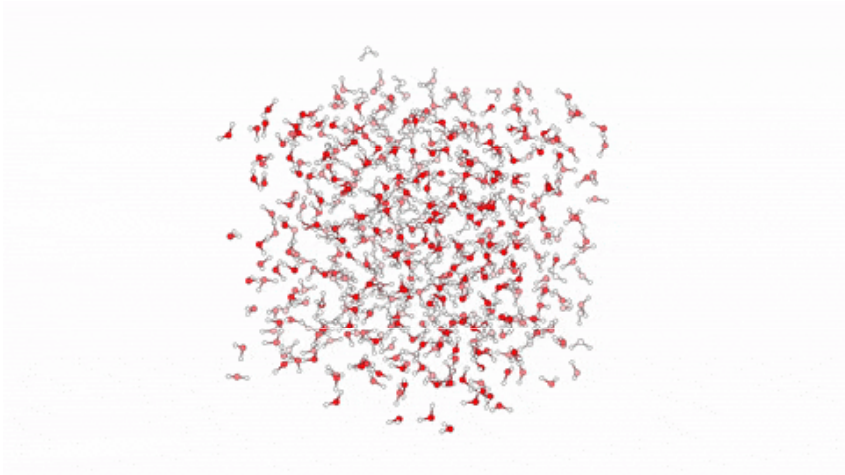
- „Klikátko“
- Skvelé pre manuálne anotovanie datasetu
- Validácie oproti JSON schema
- Potreba rozšírenie o referencie/dependency

Definície metadata schémat

- Snaha definovať/adopovať metadatové schémy pre publikované datasety
  - CEITEC CFs + spolupracujúca komunita
- Pre simulácie molekulovej dynamiky
  - Róbert Vácha a „Molecular Simulations and Design research group“ z Max Planck Insititute
- Formálna schéma ale aj sémantika anotácií

Modifikácia analytických nástrojov

- Gromacs
- Analytické/simulačné nástroje modifikovať aby výstup experimentov obsahoval aj metadatové popisy
  - Parametre spustenia, referencia na vstupné dáta



Metadatový katalog

- Cieľom je budovať metadatové katalogy
  - Pre odborné komunity / vyskumné tímy
- Možnosť jednoducho prehľadávať datasety
  - Zjednodušiť orientáciu v publikovaných ale aj vlastných datasetoch
- Poskytnúť nástroj pre prípravu datasetu pre publikáciu
  - Anotovanie datasetu
  - Validácia schémy
  - FAIR checking

Majú data bez metadat zmysel?

- Jo.
  - V pripade, že ich je veľa, zdieľajú sa, ťažko sa v nich orientuje


MUNI