Zpracování dat na CEITECu

Tomáš Raček (CF BioData)

O mně

- Studium FI MU (2008 – 2022)
- Zaměření směrem k výpočetní chemii, strukturní bioinformatice
- Člen Sitoly během doktorského studia
- Vazby na ÚVT
- Příklon k NCBR PřF (výuka, téma doktorského studia,...)
- Formálně na CEITEC MU od 2017
- Zástupce vedoucí centrální laboratoře BioData (2023)

CEITEC

- **Středoevropský technologický institut**
- Založen 2011
- Zaměření
  - Vědy o živé přírodě
  - Pokročilé materiály
  - Nanotechnologie
- Zapojené instituce

MUNI VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ Mendelova univerzita v Brně UNIVERSITAS VETERINARIA BRUNENSIS IPM VÚVeL

MUNI | CEITEC

CEITEC v číslech (2022)

- 28 výzkumných skupin
- **13 centrálních laboratoří (CF – Core Facility)**
  - 720 uživatelů (z 22 zemí)
- 261 výzkumných FTE
- 241 publikací (108 Q1, 24 T5)
- Vlastní Ph.D. program (212 studentů)

C

MUNI MU

Centrální laboratoř Kryo-elektronová mikroskopie a tomografie


Mgr. Jiří Nováček, Ph.D.

MUNI MU

Národní NMR centrum Josefa Dadoka


doc. RNDr. Radovan Fiala, CSc.

MUNI MU

Centrální laboratoř Proteomika


prof. RNDr. Zbyněk Zdráhal, Dr.

MUNI MU

Centrální laboratoř Interakce a krystalografie biomolekul


Mgr. Josef Houser, Ph.D.

MUNI MU

Centrální laboratoř Nanobiotechnologie


Mgr. Jan Přibyl, Ph.D.

MUNI MU

Laboratoř multimodálního a funkčního zobrazování


Ing. Michal Mikl, Ph.D.

MUNI MU

Centrální laboratoř Genomika


MVDr. Boris Tichý, Ph.D.

MUNI MU

Centrální laboratoř Buněčné zobrazování


Mgr. Milan Ešner, Ph.D.

MUNI MU

Laboratoř rostlinného výzkumu


Natallia Madzia Valasevich, PhD

MUNI MU

Centrální laboratoř Bioinformatika


Mgr. Vojtěch Bystrý, Ph.D.

MUNI MU

Centrální laboratoř Správa a analýza biologických dat


doc. RNDr. Radka Svobodová, Ph.D.

MUNI | CEITEC

CF BioData

- **Centrální laboratoř Správa a analýza biologických dat**
- Vedoucí Radka Svobodová
- Založena 2022
- Nabízené služby:
  - **Správa a ukládání dat**
  - Podpora pro přístup k úložným a výpočetním zdrojům
  - Strukturní bioinformatika
  - Propojení s projektem ELIXIR

MUNI | CEITEC

CF BioData – core tým

- **Radka Svobodová (vedoucí CF)**
- Tomáš Raček (strategie)
- Vladimír Horský (rezervační systém)
- Adrián Rošinec (správa metadat, cloud)
- Tomáš Svoboda (správa dat)

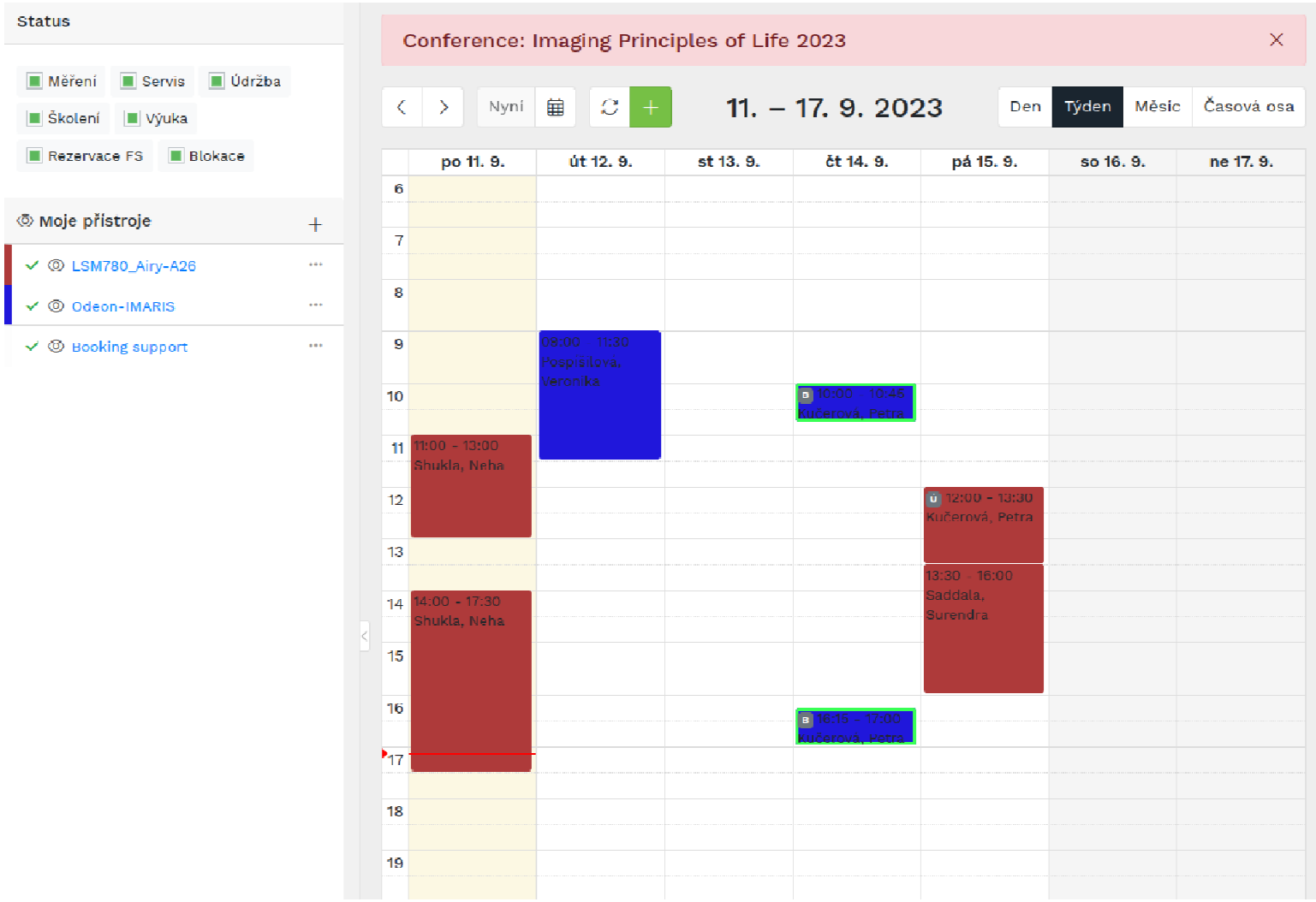
MUNI | CEITEC

Správa dat na CEITEC

- Některé CF si řeší samy
  - Velká nebo citlivá data
  - Vznik CF BioData až od 2022
- U některých reálně neexistuje (flashka, externí disk)
- Vlastní úložiště brno14-ceitec (~ 6 PB)
  - Datová politika (možnosti alokace prostoru pro skupinu?)
- CF by měly poskytovat správu dat svým uživatelům
- Technické detaily → přednáška Tomáše S.

Rezervační systém (booking)

- Rezervační systém primárně pro přístroje CF
- Správa požadavků, cenotvorby, certifikací,...
- Každá CF by jej měla používat
- Vyvíjí externí firma
- Postaveno nad CRM Microsoft Dynamics 365
  - On-site instalace
  - Provozuje VUT
  - Produkční a testovací instance
- Programové úpravy pro jednotlivé CF


Status
Měření
Servis
Údržba
Školení
Výuka
Rezervace FS
Blokace
Moje přístroje
LSM780_Airy-A26
Odeon-IMARIS
Booking support
Conference: Imaging Principles of Life 2023
Nyní
11. – 17. 9. 2023
Den
Týden
Měsíc
Časová osa
po 11. 9.
út 12. 9.
st 13. 9.
čt 14. 9.
pá 15. 9.
so 16. 9.
ne 17. 9.
08:00 - 11:30
Pospíšilova, Veronika
10:00 - 10:45
Kučerová, Petra
11:00 - 13:00
Shukla, Neha
12:00 - 13:30
Kučerová, Petra
13:30 - 16:00
Saddala, Surendra
14:00 - 17:30
Shukla, Neha
16:15 - 17:00
Kučerová, Petra

Booking systém – využití (2022)

	BIC	CELLIM	CEMCOF	Genomics	NMR	MAFIL	Nanobio	Plants	Proteomics
Počet přístrojů	88	23	27	142	16	27	13	254	43
Počet rezervací	676	5 141	1 686	3 030	730	2 592	463	305	-
Celkový čas rezervací [h]	9 909	11 070	19 992	11 536	39 200	4 400	4 396	882 870	20 390

MUNI | CEITEC

Otázky

- Jaké jsou objemy dat?
- Jsou data citlivá?
- Kdo a kde data produkuje?
- Je potřeba data replikovat?
- Je potřeba ukládat raw data?
- Jak dlouho data ukládat?
- Kdo má práva k datům přistupovat?
- Je potřeba přístup pro uživatele mimo MU?

CEMCOF

- **Centrální laboratoř kryoelektronové mikroskopie a tomografie**
- Aktuálně asi 2.7 PB dat
- > 1 000 datových sad
- Vlastní webové rozhraní pro správu experimentů


```
AccessRoute: Ciisb
Autoname: 2022323990_Pinkas
Center: CEMCOF,EM-Instruct-CZ,Brno
ClassType: lims.Cemcof.Microscopes.SPAExperiment
ExperimentData:
  Archive: true
  AutopickingModel: null
  Binning: 1
  ClassType: lims.Cemcof.Microscopes.SpaTomoData
  Clean: true
  Cs: 2.7
```


MUNI | CEITEC

CF Plants

- **Centrální laboratoř rostlinného výzkumu**
- Fenotypovací stanice
  - Automatické měření vlastností v čase
  - Fotografie
  - CSV s naměřenými hodnotami
  - ~ desítky MB / experiment

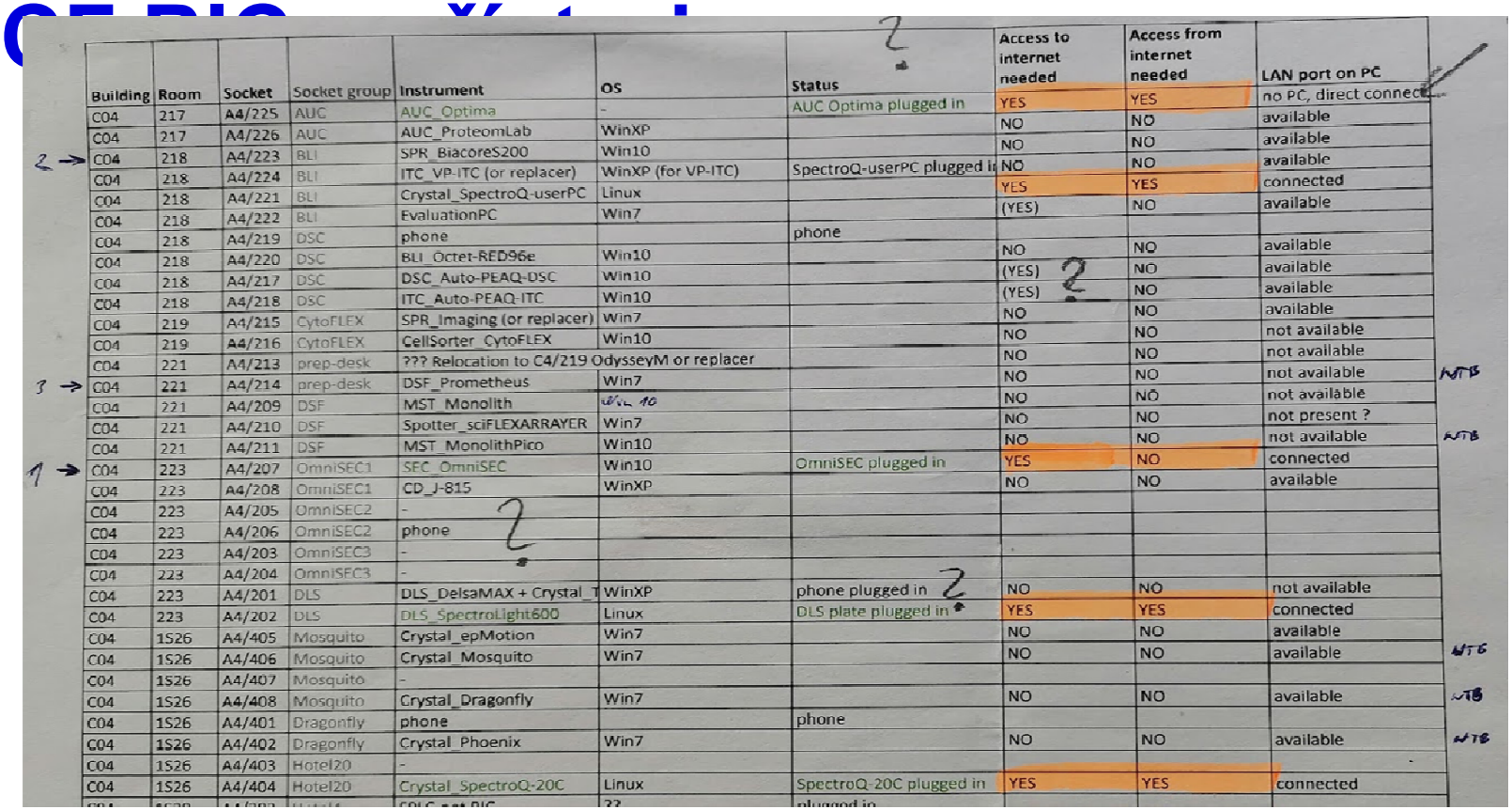
- Experimenty provádí zaměstnanci laboratoře
  - Potřeba dostat výsledky experimentů k uživatelům
  - Ručně generované přehledy

MUNI | CEITEC

CF BIC

- **Centrální laboratoř Interakce a krystalografie biomolekul**
- Desítky různých přístrojů (= zdrojů dat)
  - Velká heterogenita
  - Malá textová data (často i <1 MB / experiment)
- Obslužné počítače často zastaralé
  - Windows Vista, XP
  - Problematické rozšíření (notebook)
  - Někdy už nepodporovaný SW
- Uživatelé měří na přístrojích sami
  - Sdílené účty
  - Data odnáší na flashce

MUNI | CEITEC



Building	Room	Socket	Socket group	Instrument	OS	Status	Access to internet needed	Access from internet needed	LAN port on PC
C04	217	A4/225	AUC	AUC_Optima	-	AUC Optima plugged in	YES	YES	no PC, direct connect
C04	217	A4/226	AUC	AUC_ProteomLab	WinXP		NO	NO	available
C04	218	A4/223	BLI	SPR_BiacoreS200	Win10		NO	NO	available
C04	218	A4/224	BLI	ITC_VP-ITC (or replacer)	WinXP (for VP-ITC)	SpectroQ-userPC plugged in	NO	NO	available
C04	218	A4/221	BLI	Crystal_SpectroQ-userPC	Linux		YES	YES	connected
C04	218	A4/222	BLI	EvaluationPC	Win7		(YES)	NO	available
C04	218	A4/219	DSC	phone		phone			
C04	218	A4/220	DSC	BLI_Octet-RED96e	Win10		NO	NO	available
C04	218	A4/217	DSC	DSC_Auto-PEAQ-DSC	Win10		(YES)	NO	available
C04	218	A4/218	DSC	ITC_Auto-PEAQ-ITC	Win10		(YES)	NO	available
C04	219	A4/215	CytoFLEX	SPR_Imaging (or replacer)	Win7		NO	NO	available
C04	219	A4/216	CytoFLEX	CellSorter_CytoFLEX	Win10		NO	NO	not available
C04	221	A4/213	prep-desk	??? Relocation to C4/219	OdysseyM or replacer		NO	NO	not available
C04	221	A4/214	prep-desk	DSF_Prometheus	Win7		NO	NO	not available
C04	221	A4/209	DSF	MST_Monolith			NO	NO	not available
C04	221	A4/210	DSF	Spotter_sciFLEXARRAYER	Win7		NO	NO	not present ?
C04	221	A4/211	DSF	MST_MonolithPico	Win10		NO	NO	not available
C04	223	A4/207	OmniSEC1	SEC_OmniSEC	Win10	OmniSEC plugged in	YES	NO	connected
C04	223	A4/208	OmniSEC1	CD_J-815	WinXP		NO	NO	available
C04	223	A4/205	OmniSEC2	-					
C04	223	A4/206	OmniSEC2	phone					
C04	223	A4/203	OmniSEC3	-					
C04	223	A4/204	OmniSEC3	-					
C04	223	A4/201	DLS	DLS_DelsaMAX + Crystal_T	WinXP	phone plugged in	NO	NO	not available
C04	223	A4/202	DLS	DLS_SpectroLight600	Linux	DLS plate plugged in	YES	YES	connected
C04	1S26	A4/405	Mosquito	Crystal_epMotion	Win7		NO	NO	available
C04	1S26	A4/406	Mosquito	Crystal_Mosquito	Win7		NO	NO	available
C04	1S26	A4/407	Mosquito	-					
C04	1S26	A4/408	Mosquito	Crystal_Dragonfly	Win7		NO	NO	available
C04	1S26	A4/401	Dragonfly	phone		phone			
C04	1S26	A4/402	Dragonfly	Crystal_Phoenix	Win7		NO	NO	available
C04	1S26	A4/403	Hotel20	-					
C04	1S26	A4/404	Hotel20	Crystal_SpectroQ-20C	Linux	SpectroQ-20C plugged in	YES	YES	connected

CF CELLIM

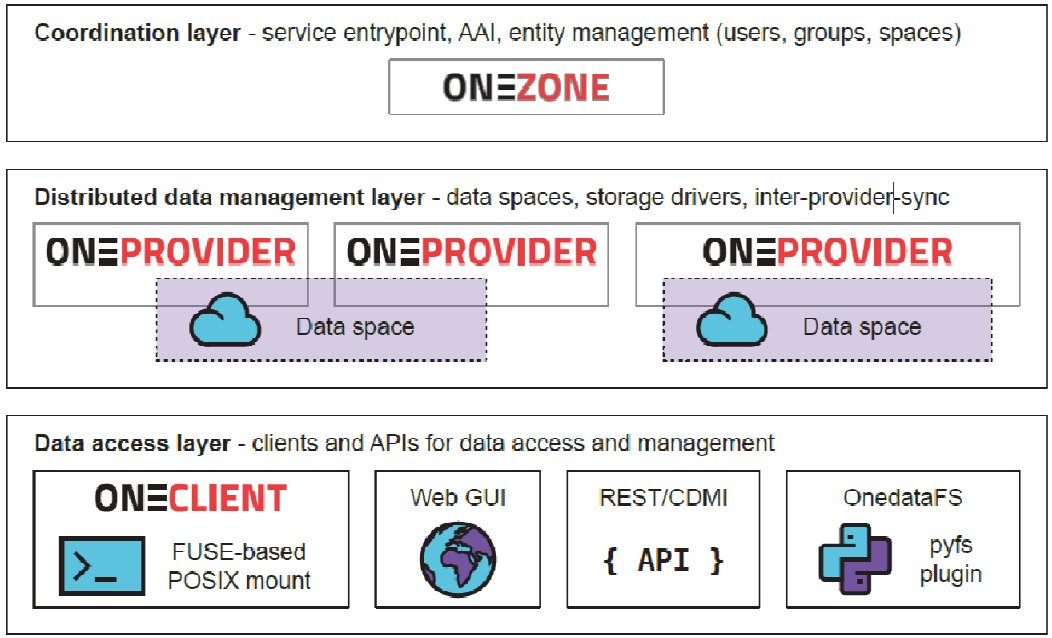
- **Sdílená laboratoř Buněčné zobrazování**
- ~ 10 mikroskopů (většinou Carl Zeiss)
  - Uživatelé měří sami
  - Nutnost explicitního povolení přístupu po proškolení ke konkrétnímu typu mikroskopu
  - Stovky GB na dataset, někdy i více než 1 TB (light-sheet)
- Vlastní sdílený HW
  - Windows Server (postprocessing, licencovaný SW)
  - 250 TB diskové pole
  - Datová politika
  - Účty pro jednotlivé uživatele (manuálně spravované)

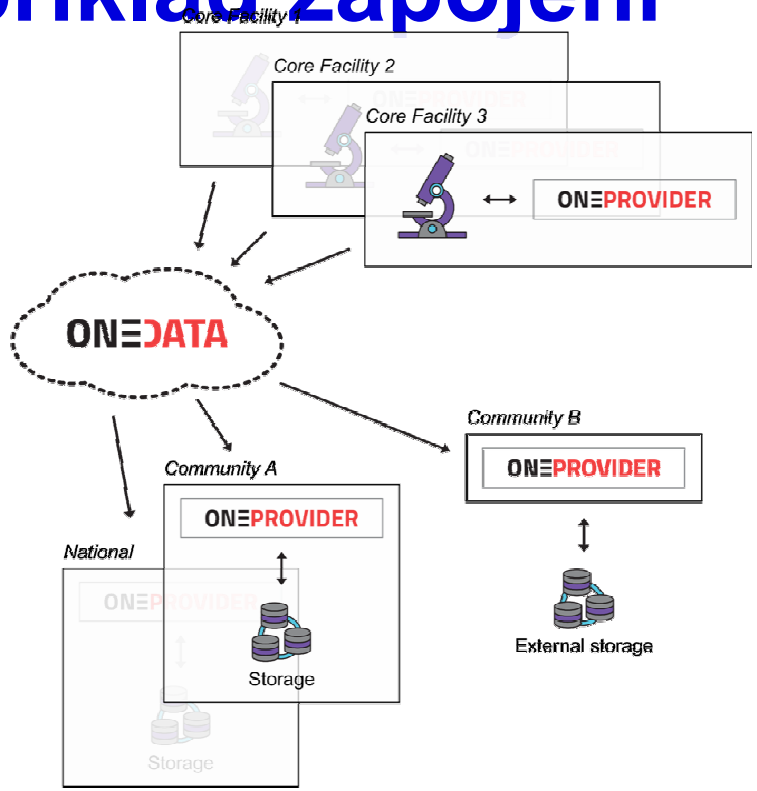
MUNI | CEITEC

Onedata

- Systém distribuovaného úložiště
- Komplexní webové rozhraní
- Podpora pro replikaci datových sad
- Perzistentní identifikátory
- Podpora pro uložení metadat
- ERLANG (?!)
- Komunikace s autory

Onedata – komponenty a příklad zapojení





MUNI | CEITEC

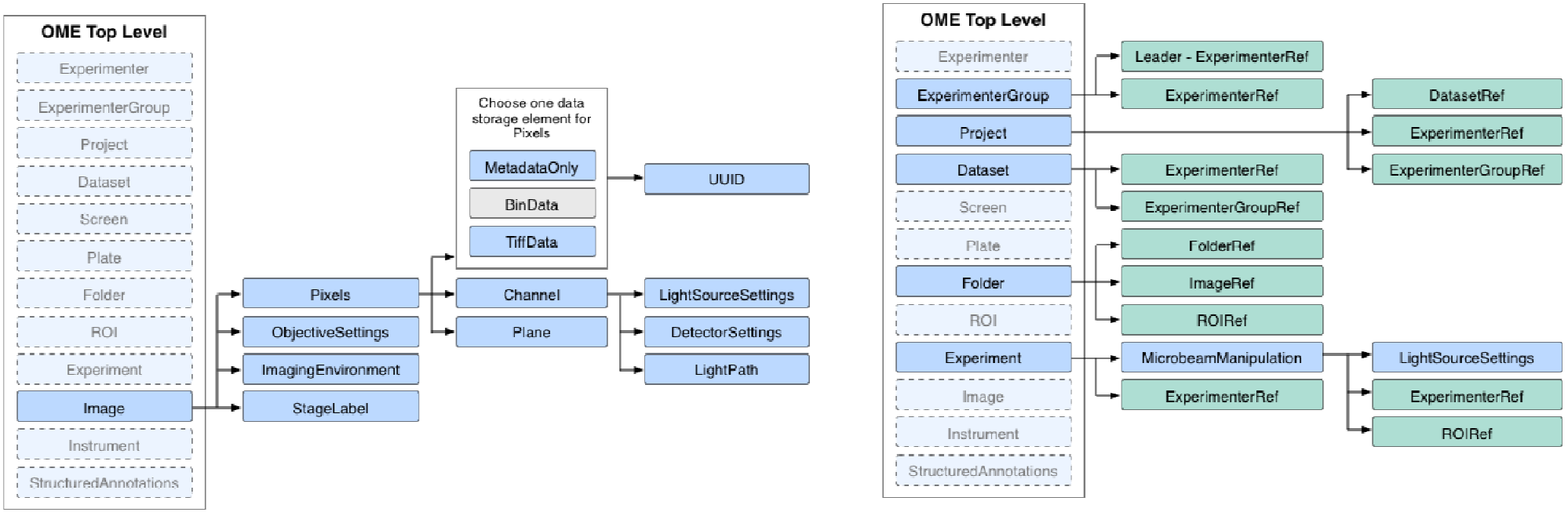
Nadstavby nad Onedata

- fs2od – “Filesystem to Onedata”
  - Automatická tvorba datových sad
  - Monitoring adresářové struktury
  - Možnosti nastavení replikace
  - Podpora pro metadata (YAML)
- Skript pro download celého datasetu
- Původně pro CEMCOF

Správa metadat

- **Mají data bez anotací smysl?** (přednáška Adriána)
- Metadata souborů
  - Často lze automaticky extrahovat (např. nastavení mikroskopu)
- Metadata experimentu / datové sady
- Metadata společná pro všechny CF
- Metadatová schémata
  - Automatické vyplňování hodnot
  - Validace
  - Ontologie?
- Exporty do oborových katalogů

Příklad – Open Microscopy Schema



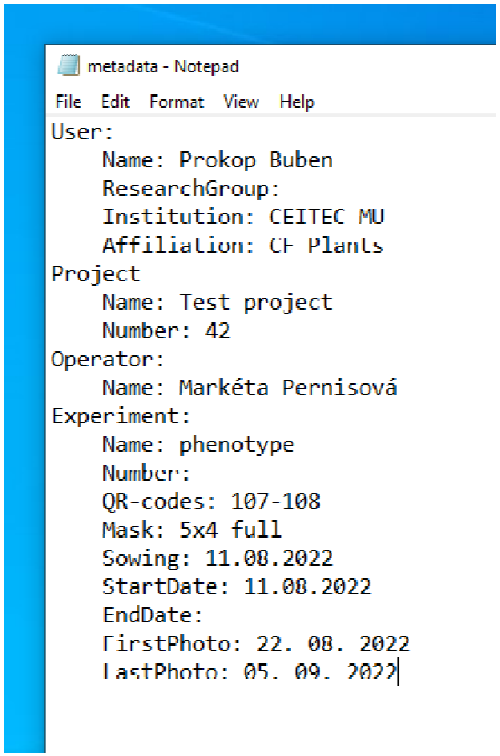
MUNI | CEITEC

FR CESNET 2022 (Tomáš Svoboda)

- Cíl 1: Rozšíření fs2od pro jiné zdroje dat (CF)
  - CF PLANTS
  - CF CELLIM
- Cíl 2: Podpora pro spouštění aplikací nad datovými sadami
  - K8s (driver)
  - Scipion
- Konec projektu: červen 2023
- Aktuálně: draft článku Onedata4Sci

Aktuální výzvy

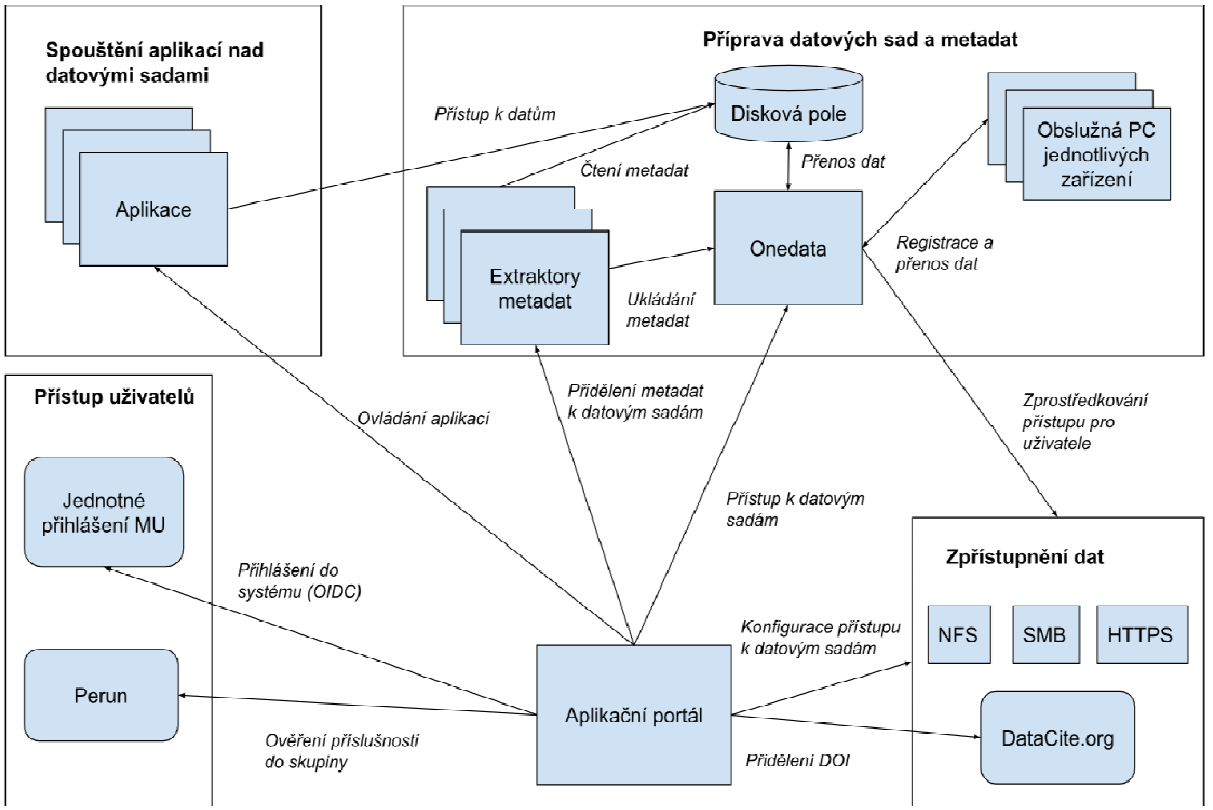
- Chybějící uživatelská přívětivost:
  - Přehledy datových sad
  - Vyhledávání podle metadat
  - Manuální založení datové sady
  - Uzamknutí datové sady
  - Přidělení DOI
  - Stažení celé datové sady
- Automatická extrakce metadat
- Windows client
- Spouštění uživatelských aplikací nad datovými sadami
- Změny v Onedata



MUNI | CEITEC

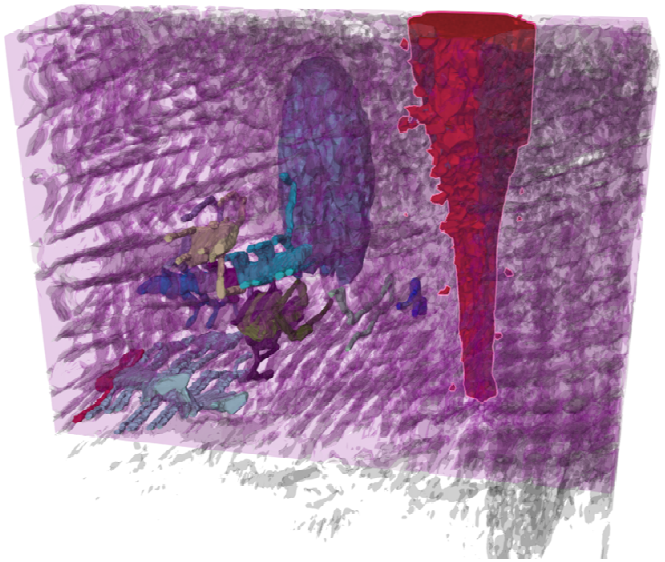
FR CESNET 2023 (Tomáš Pašek)

- Cíl: Rozhraní pro správu datových sad a jejich metadat
  - Konsolidace jednotlivých use cases
  - Definice metadatových schémat pro jednotlivé CF
  - Vyhledávání
  - Přehledy/statistiky?



Další projekty – CELLIM + Mol*

- CF CELLIM (světelná mikroskopie)
  - Snímání ve více z-rovinách
  - Rekonstrukce 3D obrazu
  - Segmentace
- Mol*
  - Nástroj pro vizualizaci molekulárních struktur
  - Rozšíření pro segmentační data (Junior Star GAČR)
- Cíl: Pipeline
  - Uložení v Onedata
  - Downsampling pro coordinate server
  - Zobrazení uživateli ve webovém prohlížeči



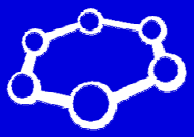
MUNI | CEITEC

Další menší projekty

- **Projekt A-C-G-T** (Adrián)
  - Databáze genetické informace české populace
  - Jednotky TB dat
  - Potřeba autentizace (odborná veřejnost)
  - Preprocessing dat a nasazení v K8s (Adrián)
- **MAFILDB** (Tomáš S.)
  - Databáze MRI měření
  - Informace o pacientech, experimentech
- **GOLEM** (Adrián)
  - Hosting webové aplikace

CF I

MUNI | CEITEC

MUNI |  CEITEC