

Pilotní repozitáře NRP

Daniel Mikšík, MU – ÚVT – CERIT-SC – Péče o komunity a uživatele, KA 3 NRP

Open Science Workshop MU

Brno 25. června 2025



Spolufinancováno
Evropskou unií



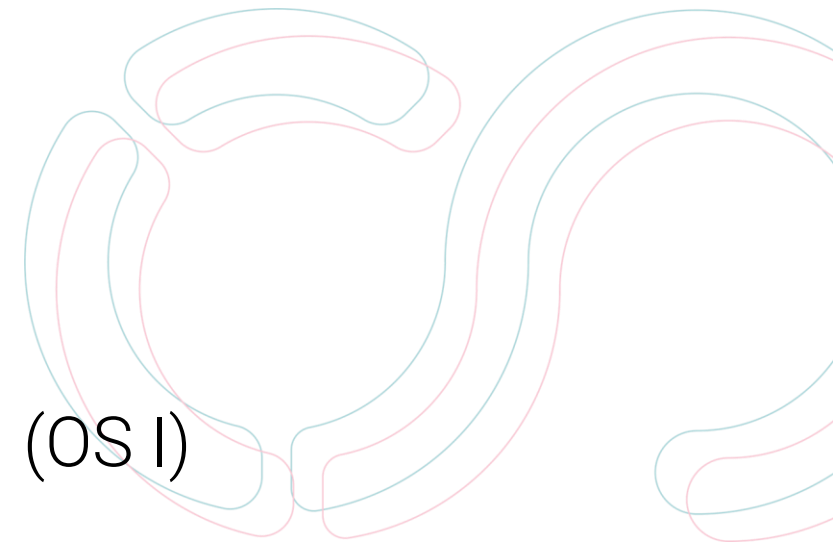
MINISTERSTVO ŠKOLSTVÍ,
MLÁDEŽE A TĚLOVÝCHOVY



Registrační číslo projektu NRP
CZ.02.01.01/00/23_014/0008787

Co je NRP I.

- Národní repozitářová platforma pro výzkumná data
 - web: www.eosc.cz/projekty/nrp
- koordinátor: CESNET, z.s.p.o.
 - hlavní projektový manažer: Luděk Matyska
- partneři (9+1)
 - univerzity: MU, VŠB-TUO, UK, ČVUT, UPOL
 - AV ČR: Heyrovského ústav, BTÚ, KNAV, SSČ
 - NTK (bez finančního příspěvku)
- trvání: 1. 1. 2024 až 31. 12. 2028 (5 let)
- rozpočet: 1,3 mld. Kč, OP JAK, výzva Open Science I (OS I)

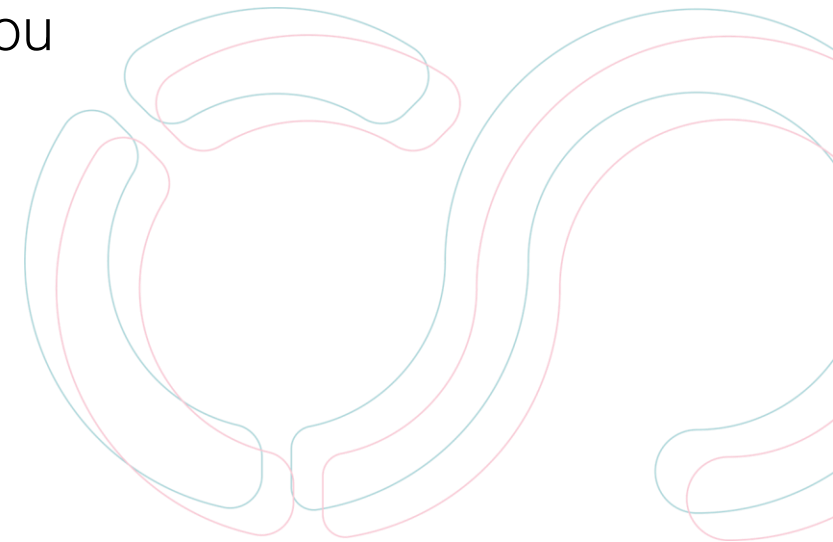


Co je NRP II.

- klíčový prvek Národní datové infrastruktury („NDI“), základního prostředí v ČR pro správu výzkumných dat
 - **ukládání, sdílení a práci** s výzkumnými daty v souladu s principy FAIR: vyhledatelná, dostupná, interoperabilní, znovu využitelná
 - kromě technického řešení také **portfolio klíčových služeb, vzdělávání** uživatelů, datových odborníků a dalších pracovníků ve výzkumných organizacích
- komplementární projekty
 - předcházející: IPs CARDS (metadata, identifikátory), IPs EOSC CZ (administrativní zázemí, pracovní skupiny, národní metadatový adresář)
 - následné: Open Science II (OS II)

Repozitáře v NRP I.

- projekt NRP buduje, provozuje a poskytuje prostředí NRP pro tvorbu a provoz repozitářů
- repozitáře budují a provozují výzkumné uživatelské skupiny
 - typicky zaštitěné institucí/institucemi s právní subjektivitou
 - očekáváme oborové/tematické repozitáře
- součástí projektu NRP je 5 pilotních repozitářů
- právě podaný projekt Open Science II obsahuje
 - 22 nových repozitářů
 - 13 inovovaných repozitářů



Repozitáře v NRP II.

250 PB fyzicky
50 PB uživatelsky

25. června 2025

Uživatelé

interakce s repozitáři: vkládání, vyhledávání, stahování dat

Instance repozitářových systémů

jednotlivé tematické datové repozitáře

povinné role: „Správce repozitáře“, „Datový kurátor“

Repozitář A

Repozitář B

Repozitář C

Implementace repozitářových systémů

softwarové balíky pro instanci repozitářů

CESNET Invenio

spravuje: CESNET

CLARIN DSpace

spravuje: Matfyz UK

ASEP/ARL

spravuje: SSČ AV ČR

Technický hw/sw základ

hw úložné a výpočetní kapacity → objektové úložiště (Ceph+S3) a prostředí pro běh kontejnerů (Kubernetes)

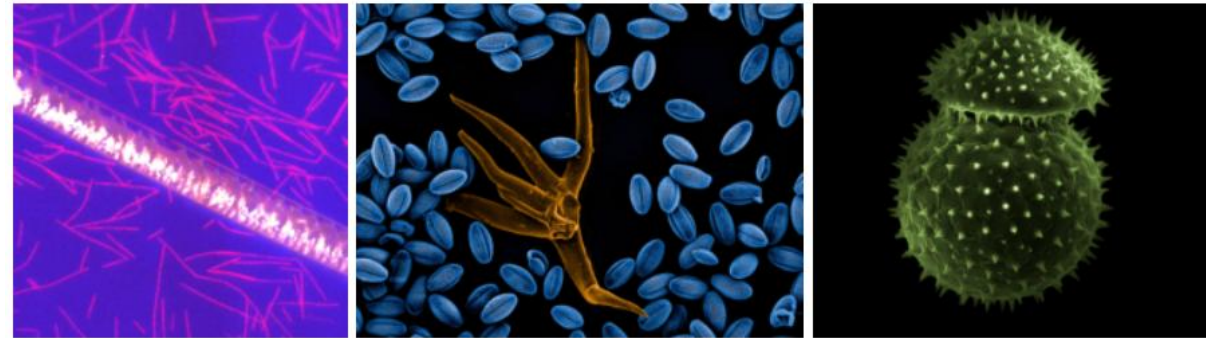
spravuje: E-INFRA.CZ, SSČ AV ČR

Pilotní repozitáře NRP

- repozitáře budují výzkumné uživatelské skupiny
 - používají prostředí NRP v procesu jeho výstavby
 - poskytují zpětnou vazbu správcům infrastruktury
 - ověřují různé scénáře využití NRP
- skupiny mají vlastní odborné zázemí, financování, plány udržitelnosti
 - projekt NRP poskytuje prostředky na vývoj repozitářů v prostředí NRP nebo na integraci vlastního řešení do NRP
 - projekt NRP nepokrývá správu dat v repozitáři (data curation)

Repozitář pro biologická zobrazovací data kdo, co

- VVI Czech-BioImaging
 - primárně zajišťuje přístup k [technologiím](#), poskytuje odborné [vzdělávání](#)
 - data ve správě výzkumníků (a jejich institucí)
 - repozitář má povýšit ukládání a zpřístupňování snímků a souvisejících dat (FAIR atd.)
 - „V naší skupině byl v datech takový nepořádek, že jsme se jejich správě museli začít věnovat, protože už by se v tom nikdo nevyznal.“ [M. Cebecauer, Univesitas, 5/2025](#)
- Ústav fyzikální chemie J. Heyrovského AV ČR
- Fakulta informatiky MU



Repozitář pro biologická zobrazovací data proč

- ukládání velkých souborů
- ukládání velkého množství malých souborů
- adekvátní metadatový popis
- extrakce metadat z přístrojových dat
 - návaznost na další aktivity projektu NRP a navazujících projektů (OS II)
- využití elektronických laboratorních deníků pro (před-)zpracování přístrojových dat a jejich odeslání do repozitáře
 - návaznost na další aktivity projektu NRP a navazujících projektů (OS II)

Repozitář pro biologická zobrazovací data kde jsme I.

– milníky

- začátek 1/2025 – CESNET Invenio
- Q2/2025: Pilotní provoz repozitáře pro beta testování uživateli
- Q4/2025: Repozitář napojen na služby NRP – základní verze umožňující AAI, přidělování licencí a práci s dvěma metadatovými modely

– stav

- probíhá tvorba metadatového modelu
 - v globální komunitě existuje doporučení (REMBI), již nedostatečné, aktivity s cílem zvýšit standardizaci (např. [foundingGIDE](#))
 - založen na základním metadatovém modelu NRP ([Czech Core Metadata Model](#), CCMM), zohledňuje doporučení komunity

Repozitář pro biologická zobrazovací data kde jsme II.

- probíhá tvorba metadatového modelu
 - komunikace s oborovými partnery v oboru (konference, přímé osobní zapojení do obdobných iniciativ)
 - spolupráce s experty na metadatové modely v NRP – úvodní zaškolení a průběžné konzultace
 - využití nástroje [Dataspecer](#), probíhá intenzivní iterativní vývoj (částečně podpořen z projektu NRP)
- probíhá seznámení s repozitářovým systémem CESNET Invenio
 - repozitář využije základ poskytovaný v prostředí NRP
 - zajistí vývoj vlastních komponent nad rámec standardní podpory

Repozitář pro biologická zobrazovací data kde jsme III.

- tvorba metadatového modelu vyžaduje expertní znalost
 - lidí s touto znalostí je málo
 - křivka učení u „výzkumníků s přesahem“ do (meta-)datového specialisty není tak strmá, jak skupina očekávala
- je potřeba standardizovat komunikaci mezi skupinou a správci infrastruktury
 - aneb cesta od přímé komunikace (e-mail + chat) k helpdesku
 - vytvoření helpdesku (standardizované podpory správců repozitářů) je v běhu

Molecular Biophysics Database (MBDB)

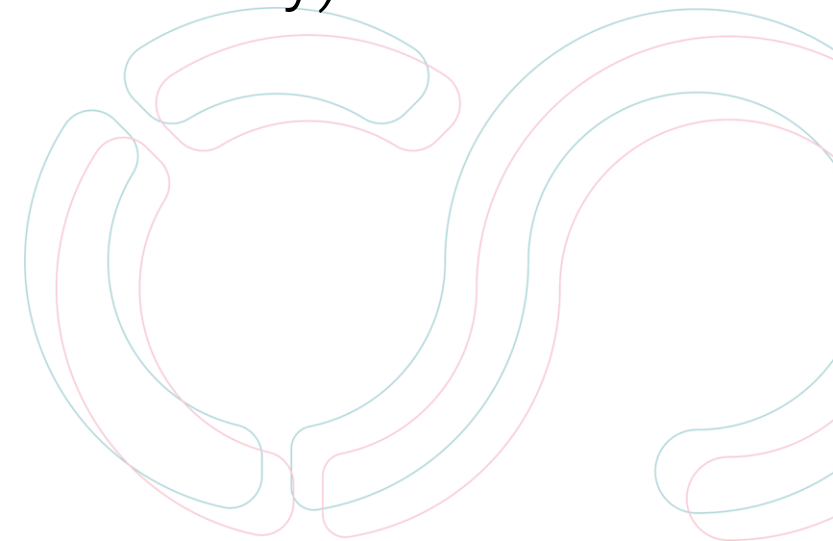
kdo, co

- Biotechnologický ústav AV ČR
- MOlecular-Scale Biophysics Research Infrastructure – MOSBRI
 - přístup k [technologiím a postupům](#), [vzdělávání](#)
 - work packages, Joint Research Activities [WP 4](#): Standards for data archiving & exploitation
 - Provision of new data standards for biophysical data
 - Support for the development of improved data processing protocols
 - **Provision of a pilot database for biophysical data**
- různé biofyzikální metody/techniky
 - složitě strukturovaný metadatový popis + soubor(y) v (proprietárních) formátech

Molecular Biophysics Database (MBDB)

proč

- více technik = více metadatových modelů v jednom repozitáři
- velké a složité metadatové modely
- synchronizace lokálně udržovaných údajů (řízené slovníky)
 - s externími zdroji
 - taxonomie NCBI pro anotaci biologického organismu
 - projekty OpenAire pro informace o financování
 - ORCID pro autentizaci
 - Pubchem pro chemické druhy



Molecular Biophysics Database (MBDB)

kde jsme I.

- milníky
 - vývoj zahájen 2021/2022, CESNET Invenio
 - 2025: Přenesení repozitáře do produkčního režimu
- průběh
 - průzkum potřeb komunity (členové MOSBRI + svět): metody – které, četnost
 - přehled existujících repozitářů a datových formátů
 - výběr biofyzikálních metod pro další kroky
 - specifikace vlastností repozitáře, výběr technického řešení (CESNET Invenio)
 - tvorba metadatových schémat
 - přidělování unikátních identifikátorů, definice řízených slovníků
 - sw vývoj
- první produkční verze repozitáře 3/2025: mbdb-data.org, vč. [dokumentace](#)

Molecular Biophysics Database (MBDB)

kde jsme II.

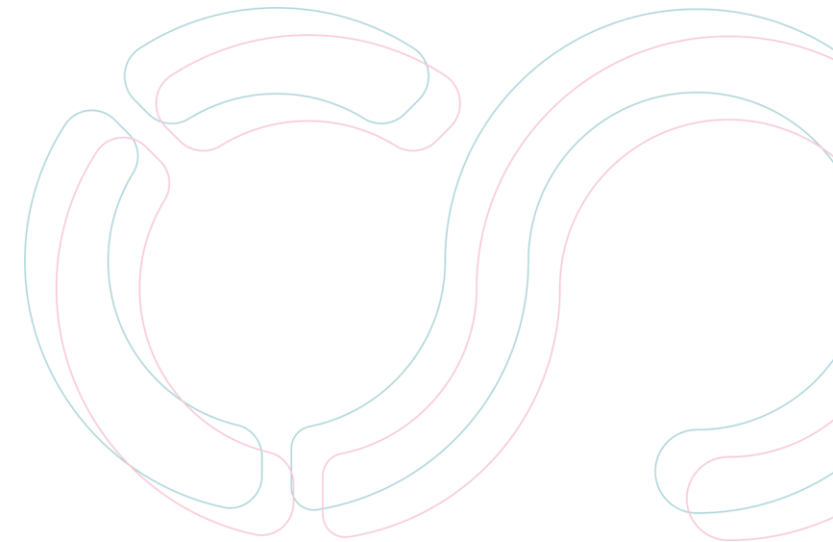
– aktuální témata

- standardizovat komunikaci mezi skupinou a správci infrastruktury
- zjednodušit pro vývojáře MBDB přístup k logům
- standardizovat zatím organický proces testování nových verzí, staging, nasazení do produkce
- doladit vkládání dat přes API
- upgrade základního kódu CESNET Invenio
 - zajišťuje NRP
- napojení na autentizační a autorizační infrastrukturu (AAI) e-INFRA.cz
 - zprovoznění komunit v repozitáři
 - zprovoznění nástrojů pro komunikaci s koncovými uživateli (zpětná vazba, hlášení chyb)

Datový catch-all repozitář

kdo, co

- CESNET
 - správce repozitáře: CESNET
 - datový kurátor: NTK + CESNET
- repozitář poslední volby
 - obecné využití pro libovolnou datovou sadu



Datový catch-all repozitář proč

- ukládání dat v prostředí národní e-infrastruktury, v jurisdikci České republiky, s dostatečnou kapacitou
- pokrytí přechodného období, kdy se jednotlivé odborné komunity teprve formují, vyhovující oborový repozitář ještě nemají k dispozici, přesto potřebují ukládat datové sady do repozitářových systémů

Datový catch-all repozitář kde jsme I.

– milníky

- 2024: zveřejněn pre-produkční repozitář (CESNET Invenio)
- 2025: převod repozitáře do produkčního prostředí NRP
- 2025: příprava materiálů pro certifikaci a zahájení procesu certifikace

– stav

- upgrade kódu CESNET Invenio
 - současně změna základního URL data.narodni-repozitar.cz -> datarepo.eosc.cz
 - návrh grafického uživatelského rozhraní a vizuální identity
- **přechod vnitřního metadatového modelu na CCMM, vč. slovníků**

Datový catch-all repozitář kde jsme II.

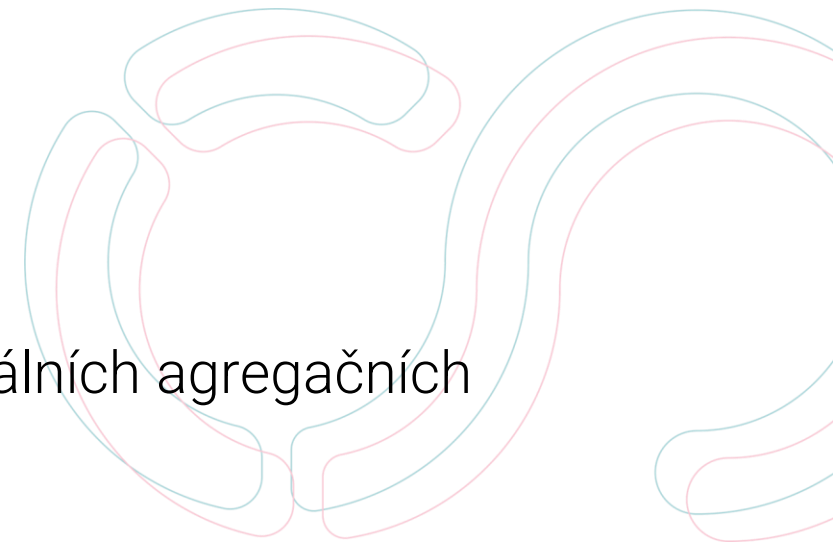
– stav

- nastavení exportu metadat pro sklizení přes OAI-PMH do Národního metadatového adresáře (NMA)
- formulace politiky repozitáře, finalizace workflow pro ukládání dat a přístup k nim
- popis životního cyklu záznamu (doba uložení, důvody výmazu)
- definice uživatelských skupin a pravidel pro vytváření komunit
- rozsáhlá aktualizace dokumentace (obsahová, grafická, technická)
- aktualizovaná produkční verze: Q3/Q4 2025

Repozitář pro herbářové sbírky

kdo, co

- Přírodovědecká fakulta UK
- další instituce spravující herbářové sbírky
 - MU
 - Národní muzeum
 - Moravské zemské muzeum
 - Botanický ústav AV ČR
 - a další
- fotografie/skeny herbářových položek
 - pro využití v záznamech o herbářových položkách v globálních agregačních službách databází [JACQ](#) -> [GBIF](#)



Repozitář pro herbářové sbírky proč

- v ČR cca 10 mil. herbářových položek (rostliny, lišejníky, mechy, semena)
- 75 % v 5 největších herbářových sbírkách PRC, PR, BRNM, PRM, BRNU
 - online zpřístupněno cca 250 tis.
 - fotografie/skeny obsahuje cca 50 tis. položek
- cíl
 - vytvořit prostředí a nástroje pro efektivní zpřístupňování dalších herbářových položek vč. fotografií/skenů
 - ověřit napojení skupinou vyvíjeného sw řešení na základní vrstvu NRP a navazující služby NDI (Národní metadatový adrsář, Národní katalog repozitářů, AAI e-INFRA.CZ ad.)

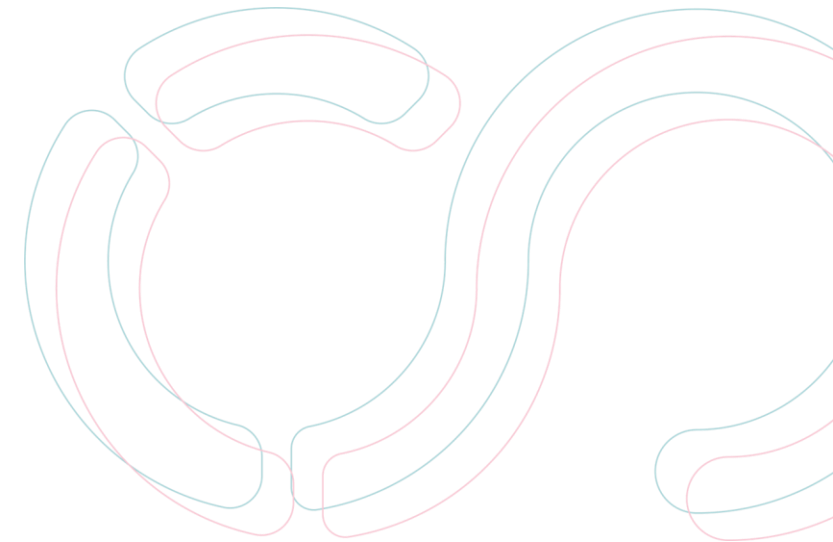
Repozitář pro herbářové sbírky kde jsme I.

- milníky
 - začátek: 5/2024 (vlastní sw řešení)
 - 2024: pilotní provoz na zdrojích e-INFRA.cz: herbarium.biodiversity.cz
 - 2025: integrace pilotního repozitáře pro herbářové sbírky na definitivním úložišti NRP
- průběh a stav
 - další části depozičního workflow
 - přímý zápis do JACQ po zveřejnění snímků v repozitáři – změna příznaku *hasImage*
 - integrace do NDI: AAI, PID, CCMM
 - licence, smluvní zajištění provozu repozitáře

ArchaeoVault

český repozitář pro neterénní archeologická data – kdo

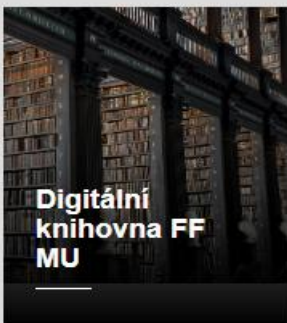
- ÚAM FF MU (projekt RES-HUM), AÚ AV ČR Praha a Brno
 - výzkumná uživatelská skupina
- KISK FF MU + CIT FF MU
 - analýza, architektura, vývoj
- infrastruktura [Digitalia MUNI ARTS](#)
 - vývoj a správa
 - součást VVI [LINDAT/CLARIAH-CZ](#)
 - repozitářový systém Islandora
 - integrace do NDI (S3, K8s, AAI, NMA atd.)



ArchaeoVault

co, proč

- **neterénní** archeologická data = vznikla z objektu některou **analytickou** metodou
 - chemická, materiálová, fyzikální, antropologická (DNA), traseologie
- pro všechny **archeologické datasety z ČR** (české výzkumy, zahraniční výzkumy nad daty z ČR) nespadaající pod Archeologický informační systém ([AIS](#))
 - provazba s AIS (identifikátory), stejné řízené slovníky, kde to bude možné
- ověřit scénář technologický základ NRP + vlastní repozitářový systém (Islandora) + napojení na NDI (NMA, NKR, AAI, PIDs ad.)



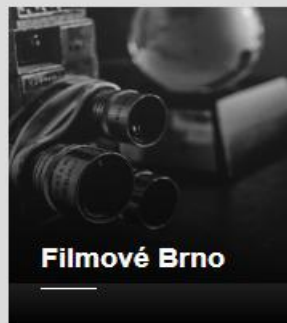
Digitální knihovna FF MU

Digitální knihovna obsahuje volně dostupné plné texty časopisů, monografií a sborníků vydaných na Filozofické fakultě Masarykovy univerzity.



Digitální knihovna Arna Nováka

Cílem projektu je zpřístupnit rozsáhlé dílo Arna Nováka. Kolekce zahrnuje digitalizované materiály z fondu Ústřední knihovny FF MU.



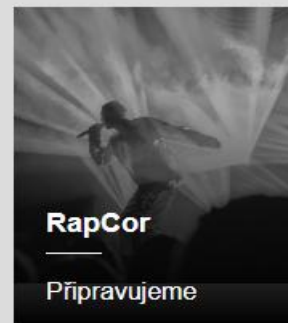
Filmové Brno

Databáze obsahuje informace o jednotlivých kinech, jejich programu a o promítaných filmech v Brně v letech 1918–1945.



Vzory krásy

Obrazová databáze nabízí několik stovek vzorů z liturgických oděvů a dalších bohoslužebných tkanin převážně z 18. století.



RapCor

Připravujeme

Jazykový korpus zaměřený na mluvenou francouzštinu v rapových písních, určený nejen pro socio-lexikální výzkum.



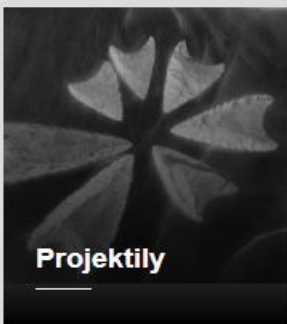
Herbaria manuscripta

Evidence středověkých rukopisných herbářů latinského i českého původu vzniklých do roku 1500 a uchovávaných ve fondech českých institucí.



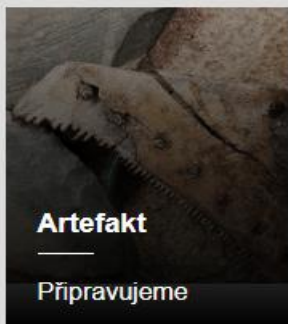
Archeo3Data

Databáze souhrnně prezentující výsledky badatelských výzkumů archeologických artefaktů různého typu a původu.



Projektily

Databáze zpřístupňuje výzkumná data, 3D modely a traseologické snímky více než 400 kamenných lukostřeleckých



Artefakt

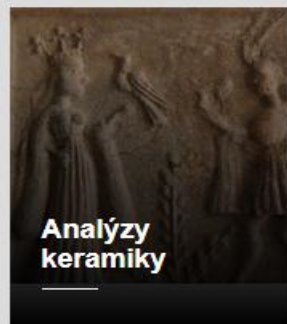
Připravujeme

Databáze zpřístupňuje výzkumná data archeologických nálezů – mikrozbytky, traseologie, struktura, technologie a provenience



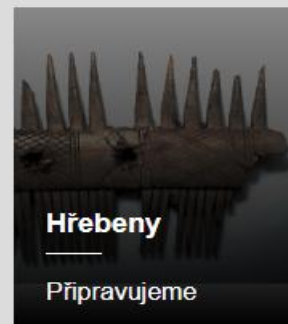
NEOFIG

Faktografická databáze zpřístupňuje výzkumná data téměř 400 neolitických figurálních plastik z moravských a rakouských nálezů. Zahrnují petrografii



Analýzy keramiky

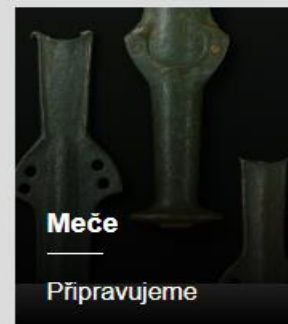
Archeologická databáze převážně středověkých keramických výrobků včetně obrazové dokumentace a podrobných výsledků přírodovědných



Hřebeny

Připravujeme

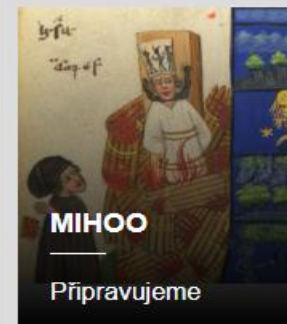
Dokumentace hřebenů z kosti a parohu z území Česka a střední Evropy. Obsahuje fotografie artefaktů a traseologických stop, 3D modely, CT



Meče

Připravujeme

Výzkumná databáze mečů z doby bronzové s fotodokumentací artefaktů, 3D modely, údaji o opotřebení a výsledky analýz.



MIHOO

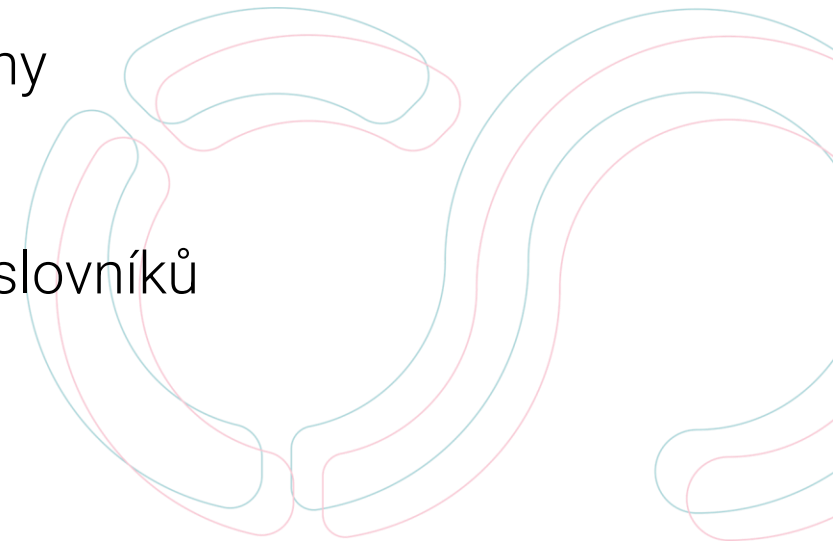
Připravujeme

Literární dílo Jana Husa v kontextu středověkých rukopisů, edic, překladů a dalších souvisejících informací.

ArchaeoVault

kde jsme

- milníky
 - zahájení 10/2024
 - 2025: přenesení repozitáře do pilotního režimu v rámci NRP
- průběh a stav
 - změna věcného obsahu, žádost o změnu, schválení změny
 - ověřen provoz Islandory v Kubernetes a nad S3
 - vývoj CLI nástroje pro správu obsahu v repozitáři
 - výběr a implementace metadatových schémat, řízených slovníků
 - dataset: CCMM
 - objekty: CARARE
 - měření: CDWA
 - vzorky: IGSN
 - media: Dublin Core
 - přístroj: PIDINST



Partneři



UNIVERZITA
KARLOVA



Univerzita Palackého
v Olomouci



Středisko společných činností
Akademie věd ČR



Děkuju za váš čas.

Komentáře a dotazy směrujte prosím na daniel.miksik@muni.cz.



Spolufinancováno
Evropskou unií



MINISTERSTVO ŠKOLSTVÍ,
MLÁDEŽE A TĚLOVÝCHOVY

Registrační číslo projektu NRP

CZ.02.01.01/00/23_014/0008787