



**Operační program
Jan Amos Komenský**

**STUDIE PROVEDITELNOSTI
- VÝZVA OPEN SCIENCE II -**

VÝŇATEK

**KLÍČOVÉ AKTIVITY A HLAVNÍ PLÁNOVANÉ
VÝSTUPY**



**Spolufinancováno
Evropskou unií**



**OPJAK.cz
MSMT.cz**

Obsah

10. 1. KA 2 – TEMATICKÉHO CLUSTERU BIO/ZDRAVÍ/POTRAVINY	4
10.1.1. PODAKTIVITA 2.1 – Národní omický repozitář – Czech Omics Node (OmiCZ)	5
10.1.2. PODAKTIVITA 2.2 – Repozitář strukturní a simulační data (BioSimCZ)	7
10.1.3. PODAKTIVITA 2.3 – Repozitář pro lidská i animální obrazová a fyziologická multimodální data (Imaging repozitář)	10
10.1.4. PODAKTIVITA 2.4 – Repozitář pro data chemické biologie a jeho propojení s vývojem a tréninkem AI modelů MS.....	11
10.1.5. PODAKTIVITA 2.5 - Tvorba nového repozitáře ClinData a FAIRifikace dat	13
10.1.6. PODAKTIVITA 2.6 – Standardizace a nastavení postupů pro ukládání a správu přístupu k DATŮM	15
10.1.7. PODAKTIVITA 2.7 – Vývoj a pilotní zavádění nástrojů pro rozvoj NDI zaměřených na interoperabilitu a uživatelský komfort	16
10.2. KA 3 – TEMATICKÉHO CLUSTERU materiálové vědy a technologie	17
10.2.1. PODAKTIVITA 3.1 – Tvorba nového repozitáře DANTE ^c a FAIRifikace dat.....	19
10.2.2. PODAKTIVITA 3.2 – Vývoj a napojení nástrojů a služeb pro sběr dat od výzkumných pracovníků a z infrastruktur do repozitáře DANTE ^c	23
10.3. KA 4 – TEMATICKÉHO CLUSTERU DATA MANAGEMENT PRO UMĚLOU INTELIGENCI A STROJOVÉ UČENÍ	25
10.3.1. podaktivita 4.1 – TVORBA REPOZITÁŘOVÉ INFRASTRUKTURY DM4AI S PODPOROU FAIR PRINCIPŮ A MEZIOBOROVÉ INTEROPERABILITY	27
10.3.2. podaktivita 4.2 – PROSTŘEDÍ PRO INFERENCI GENERATIVNÍCH AI MODELŮ ULOŽENÝCH V DATOVÉM REPOZITÁŘI	33
10.4. KA 5 – TEMATICKÉHO CLUSTERU sociální vědy	34
10.4.1. Podaktivita 5.1 – INOVACE repozitáře ČSDA a FAIRifikace dat	36
10.4.2. Podaktivita 5.2 – TVORBA nového repozitáře ČSDA pro citlivá data a FAIRifikace	40
10.4.3. Podaktivita 5.3 – inovace repozitáře DataHub a FAIRifikace dat	41
10.4.4. Podaktivita 5.4 – VÝVOJ A PILOTNÍ ZAVEDENÍ SLUŽEB a NÁSTROJŮ PRO Sociálněvědní oblast	42
10.4.5. Podaktivita 5.5 – Vzdělávací platforma a STRATEGIE KOMUNIKACE, ŠÍŘENÍ VÝSTUPŮ PROJEKTU UVNITŘ ODBORNÉ KOMUNITY V OBORECH SOCIÁLNÍCH VĚD	44
10.5. KA 6 – TEMATICKÉHO CLUSTERU FYZIKÁLNÍ VĚDY	46
10.5.1. PODAKTIVITA 6.1 – TVORBA NOVÉHO REPOZITÁŘE FYZIKA A FAIRIFIKACE DAT	47
10.5.2. PODAKTIVITA 6.2 – VÝVOJ A PILOTNÍ ZAVEDENÍ SLUŽEB/ NÁSTROJŮ PRO ROZVOJ NDI..	52
10.5.3. PODAKTIVITA 6.3 – tvorba e-learningových kurzů A MATERIÁLŮ PRO VZDĚLÁVÁNÍ	55
10.6.1. Podaktivita 7.1 – Inovace repozitáře LINDAT/CLARIAH-CZ a FAIRifikace souvisejících dat	57
10.6.2. Podaktivita 7.2 – INOVACE REPOZITÁŘE DIGITALIA MUNI ARTS A FAIRifikace souvisejících dat.....	59

10.6.3. Podaktivita 7.3 – Inovace repozitáře ArchaeoVault a FAIRifikace souvisejících dat	60
10.6.4. Podaktivita 7.4 – Vytvoření Repozitáře pro bibliografická data a FAIRifikace souvisejících dat	62
10.6.5. Podaktivita 7.5 – VÝVOJ A PILOTNÍ ZAVEDENÍ NADSTAVBOVÝCH NÁSTROJŮ PRO REPOZITÁŘE TKA HUMA	65
10.7. KA 8 – TEMATICKÉHO CLUSTERU Enviromentální vědy	66
10.7.1. PODAKTIVITA 8.1 – Inovace a rozšíření repozitáře GENASIS pro prezentaci dat o chemické kontaminaci jednotlivých matric životního prostředí a chemické expozici populace	68
10.7.2. PODAKTIVITA 8.2 – REPOZITÁŘ pro ukládání dat z necílených hmotnostně spektroskopických analýz pro účely hodnocení lidského expozomu	70
10.7.3. PODAKTIVITA 8.3 – Repozitář pro ukládání toxikologických a ekotoxikologických dat	72
10.7.4. PODAKTIVITA 8.4 – Nové Úložiště umožňující propojení geokódovaných dat z různých domén.....	73
10.7.5. PODAKTIVITA 8.5 – Repozitář pro referenční obrazová data flóry a rostlinných společenstev	74
10.7.6. PODAKTIVITA 8.6 – Repozitář pro genetický biomonitoring a genetická data.....	75
10.7.7. PODAKTIVITA 8.7 – Repozitář pro zoologické sbírky	76
10.7.8. PODAKTIVITA 8.8 – Vývoj a pilotní zavedení služeb pro analýzu environmentálních a BIODIVERZITNÍCH dat	77
10. 8. KA 9 – TEMATICKÉHO CLUSTERU CITLIVÁ DATA.....	78
10.8.1. Podaktivita 9.1 – tvorba metodiky a VZDĚLÁVACÍCH MATERIÁLŮ fairifikace citlivých dat	80
10.8.2. Podaktivita 9.2 – Vývoj a pilotní zavedení služeb a nástrojů pro rozvoj NDI	85
10.9. Průřezová TÉMATA PROJEKTU	91
10.9.1. Podaktivita 10.1 – Velké jazykové modely/LLM	Chyba! Záložka není definována.
10.9.2. Podaktivita 10.2 – KYBERBEZPEČNOST	Chyba! Záložka není definována.
10.9.3. Podaktivita 10.3 – COmpliance.....	Chyba! Záložka není definována.
10.9.4. Podaktivita 10.4 – Elektronické laboratorní deníky pro NRP	91
10.9.5. Podaktivita 10.5 – Zpětná dohledatelnost objektů pomocí provenance	96
10.9.6. Podaktivita 10.6 – Komunikační A VZDĚLÁVACÍ AKTIVITY	Chyba! Záložka není definována.
11. HLAVNÍ PLÁNOVANÉ VÝSTUPY	99

10. 1. KA 2 – TEMATICKÉHO CLUSTERU BIO/ZDRAVÍ/POTRAVINY

ZAPOJENÍ PATRNEŘI: ÚOCHB AV ČR (Garant TKA), MU, UPOL

Biomedicínský a biochemický výzkum v ČR v současnosti trpí datovou roztříštěností a absencí společné infrastruktury. Významná data – klinická, obrazová, simulační či omická – jsou často uložena lokálně, nestrukturovaně, nebo nejsou sdílena vůbec. Chybí systémová podpora standardizace, dlouhodobé archivace a FAIR kompatibility. Zároveň však v ČR existují špičková pracoviště (např. BIOCEV, CEITEC) a infrastruktury (např. ELIXIR CZ, Czech-BioImaging, BBMRI CZ), která generují data světové úrovně a jsou aktivně zapojena do evropských projektů.

Cílem KA 2 je vybudování infrastruktury repozitářů pro správu, sdílení a opětovné využití různorodých typů biologických, chemických a klinických dat, která vznikají v základním i aplikovaném výzkumu v oblastech medicíny, biologie, chemie a příbuzných věd. Repozitářové podaktivity sdílejí společný rámec, který je pevně ukotven v principech FAIR¹ – a to jak po stránce technické (standardizovaná metadata, přístupová rozhraní, formáty), tak procesní (kurátorství, verzování, správa oprávnění, citovatelnost).

Jednotlivé repozitáře jsou navrženy jako samostatné, ale interoperabilní komponenty modulárního ekosystému. Vznikne tak infrastruktura, ve které mohou být komplementární data o jednom subjektu (člověku, biomolekule, experimentu) uchovávána napříč více repozitáři a propojena pomocí metadatačních vazeb, persistentních identifikátorů a odkazových integrací (např. přes S3, RDF, Linked Data). Konkrétní zaměření jednotlivých repozitářů je následující:

- 1) ClinData – lidská data včetně klinických studií;
- 2) OmiCZ – multiomická data (např. genomická, transkriptomická, proteomická);
- 3) Imaging – medicínská a biologická obrazová data;
- 4) BioSimCZ – molekulárně dynamická a strukturní simulační data;
- 5) Chemická biologie (ChemBio) – chemicko-biologická data, nástroje strojového učení (MS/AI);

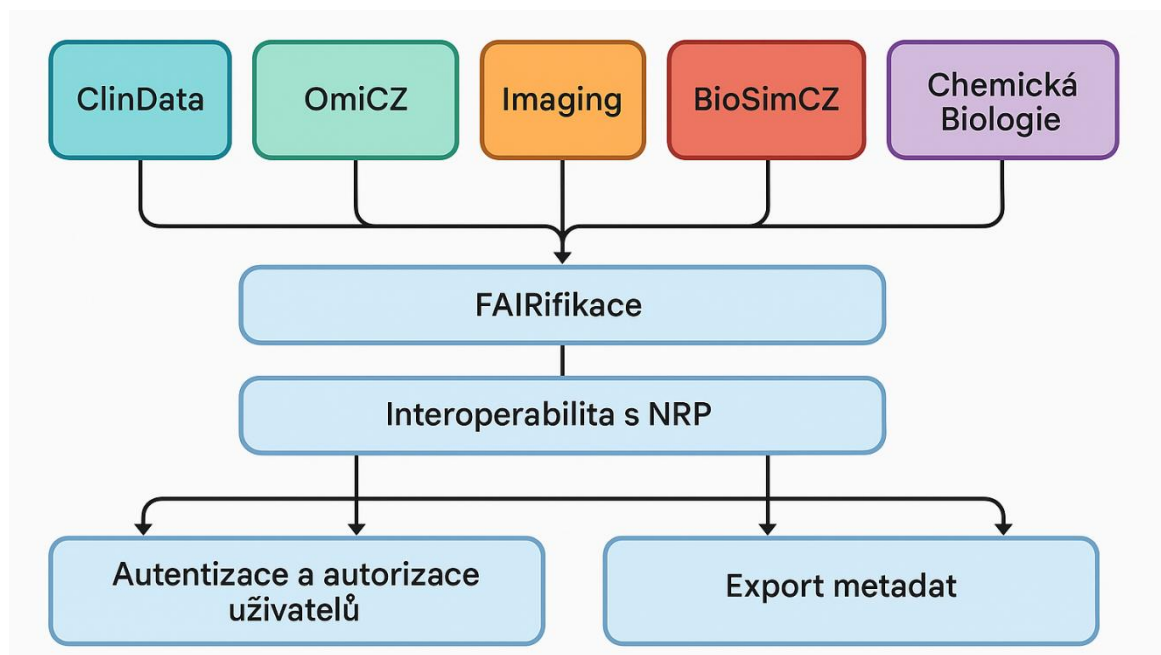
Zásadním rysem všech repozitářových systémů je zajištění interoperability s NRP, využití jednotné autentizace/autorizace prostřednictvím Life Science Login² a možnost exportu metadata do národních a evropských katalogů a infrastruktur (např. BBMRI-ERIC, ECRIN-ERIC, ELIXIR, EOSC, EBRAINS). FAIRifikace není vnímána jako jednorázový výstup, ale jako průběžný a živý rámec, který umožňuje efektivní a udržitelné sdílení dat a jejich opětovné využití napříč obory i institucemi. U vybraných repozitářů je FAIRifikace dále rozšířena o přímou podporu pokročilých nástrojů pro zpracování dat (např. AI modely v chemické biologii, anotace dat pomocí ontologií).

Tato klíčová aktivita tak nepředstavuje izolovanou snahu, ale systémovou změnu v přístupu k datům v oblasti zdraví, biologie a chemie v ČR – s důrazem na mezinárodní kompatibilitu, opakovatelnost výzkumu, otevřenost a udržitelnost. Cílem je vytvořit infrastrukturu, která umožní nejen uchovávání,

¹ <https://faircookbook.elixir-europe.org/content/recipes/introduction/brief-FAIR-principles.html>.

² <https://lifescience-ri.eu/ls-login/>.

ale i plnohodnotné využití dat pro pokročilé výzkumné (meta)analýzy, vzdělávání³, mezioborovou spolupráci a transfer znalostí.



Obr. 3: Grafické schéma aktivit

10.1.1. PODAKTIVITA 2.1 – NÁRODNÍ OMICKÝ REPOZITÁŘ – CZECH OMICS NODE (OMICZ)⁴

Repozitář OmiCZ se specializuje na ukládání, správu a analýzu omických dat (zejména genomických, transkriptomických, proteomických a flowcytometrických), která mají často citlivý charakter a spadají pod regulaci GDPR⁵. Cílem repozitáře je poskytnout bezpečné, interoperabilní a FAIR-kompatibilní řešení pro práci s těmito daty na národní úrovni, s plnou kontrolou nad jejich dostupností a zpracováním, v souladu s právními a etickými požadavky a s napojením na evropské infrastruktury FEGA a Beacon.

Na základě analýzy potřeb vědeckých komunit bude vytvořen metadatový model kompatibilní s evropskými standardy (např. EGA/FEGA), využívající existující ontologie a kontrolované slovníky. Na základě takto definovaného metadatového modelu bude repozitář OmiCZ vytvořen jako instance základního repozitářového systému NRP Invenio, která zajistí efektivní ukládání dat, správu metadat a bezpečný přístup pomocí jednotné autentizační infrastruktury AAI.

V rámci OmiCZ bude implementován software pro přímé nahrávání omických dat z přístrojů (sekvenátorů, flow cytometrů) a dále uživatelská rozhraní (ProducerUI, ClinUse, CropUI), která umožní oborově specifickou práci s repozitářem s hlavním přidanou funkcionalitou umožnit konfiguraci a spouštění analytických workflow.

³ Návnazné vzdělávací aktivity pro podporu pracovníků datové podpory a dalších odborníků v rámci komunity budou financovány a organizovány Školicím centrem EOSC CZ, které zajistí odpovídající kapacity, technické zázemí.

⁴ V popisu KA jsou plánované aktivity propojeny na realizované výstupy. Výstupy jsou označeny kódem, pro TKA B/H/F v podobě **B_x** (klíčový), **B_x.x** (dílčí, tzn. náležící ke klíčovému).

⁵ <https://mv.gov.cz/gdpr/clanek/co-je-gdpr.aspx>.

Data a metadata uložená v repozitáři budou interoperabilní nejen na národní úrovni (napojení na repozitáře klinických dat a NRP), ale i na evropské úrovni, díky integraci s FEGA a Beacon network infrastrukturami.

Důležitou součástí repozitáře bude podpora bezpečného zpracování citlivých dat prostřednictvím delegovaných výpočetních workflow, kdy se uživatelé neukážou primární data, ale až výsledek workflow. Služba bude zahrnovat vyhodnocení citlivosti výstupů workflow v souladu s etickými a legislativními normami. Také bude zakomponována možnost vytvoření interaktivního prostředí Galaxy v rámci Trusted Research Environment (TRE), které umožní bezpečné zpracování citlivých genomických dat v repozitáři. Návrh a tvorba těchto služeb, které budou využity v OmiCZ repozitáři je plánována a definována v rámci TKA SENSI.

V rámci přípravy systému OmiCZ proběhne také shromáždění testovacích datasetů a návrh reálných scénářů použití, které pokryjí různé typy omických dat (genomická, transkriptomická, proteomická, flowcytometrická) i různé aplikační kontexty (klinická data, rostlinná/agro data apod.). Datasety budou poskytnuty partnery projektu a poslouží k ověření funkčnosti jednotlivých komponent systému, včetně nahrávání, správy metadat, řízení přístupu a analýzy v TRE.

V rámci závěrečné fáze aktivity budou realizována školení pro koncové uživatele systému a pracovníky datové podpory a dalších odborníků v rámci komunity (vědce, datové kurátory, administrátory). Cílem školení bude zajistit dostatečné porozumění správě a využívání citlivých dat v prostředí OmiCZ a napojení na evropskou infrastrukturu.⁶

ČINNOSTI:

- sběr požadavků od komunit (typy produkovaných dat, používané datové formáty, metadatové požadavky);
- vytvoření metadatového modelu vhodného pro omická data, kompatibilního s EGA/FEGA, splňujícího požadavky FAIR a využívajícího existující ontologie B_1.1;
- implementace jádra repozitáře na platformě Invenio včetně autentizace a autorizace přes AAI B_1.2;
- vývoj a implementace nástrojů pro nahrávání a správu dat:
 - rozhraní ProducerUI pro producenty dat,
 - ClinUse pro klinická data,
 - CropUI pro rostlinná/agro data,
 - nasazení funkčního systému umožňujícího nahrávání dat a správu metadat přes ProducerUI a další rozhraní. B_1.3
- zajištění interoperability s evropským a národním datovým prostorem:
 - napojení na FEGA a Beacon,
 - propojení s ClinData přes ClinLink,
 - export metadat do SPARQL engine (kde relevantní).
- funkční interoperabilita s evropskou infrastrukturou (FEGA/Beacon) a ClinData, ověřená exportem metadat a napojením na SPARQL engine B_1.4;
- implementace služeb pro workflow a analytické nástroje:
 - implementace služby Federated Workflow Execution pro dávkové analýzy bez přístupu k datům a implementace služby Galaxy v TRE pro interaktivní analýzy B_1.5.
- příprava testovacích datasetů a scénářů použití:

⁶ Školení budou realizována ve spolupráci se Školícím centrem EOSC CZ, které zajistí odpovídající kapacity, technické zázemí a organizační podporu, a nejsou tak zahrnuta do rozpočtu OS II.

- datasety poskytnuté partnery pokrývající různé typy dat a použití pro validaci systému,
 - testování, příprava dokumentace.
- governance framework (právní, etické a provozní aspekty);
 - uživatelské manuály a validační zprávy B_1.6.

KÓDY VÝSTUPŮ PODAKTIVITY⁷

B_1	Národní omický repozitář – Czech Omics Node (OmiCZ)
B_1.1	Dokumentace návrhu architektury a metadatového modelu repozitáře OmiCZ
B_1.2	Funkční prototyp repozitáře OmiCZ s integrovanou AAI
B_1.3	Funkční rozhraní ProducerUI pro správu dat a metadat
B_1.4	Propojení repozitáře OmiCZ na FEGA, Beacon a ClinData
B_1.5	Systém pro bezpečné spouštění workflow a Galaxy v TRE
B_1.6	Governance a metodická dokumentace k repozitáři OmiCZ

10.1.2. PODAKTIVITA 2.2 – REPOZITÁŘ STRUKTURNÍ A SIMULAČNÍ DATA (BIOSIMCZ)

Hlavním cílem podaktivity je vytvoření českého strukturního simulačního repozitáře, zkráceně označovaného BioSimCZ. Jedná se o nový repozitář, který bude umožňovat ukládání, správu a další zpracování strukturních simulačních dat (tj. výstupů simulací, týkajících se 3D struktur biomakromolekul). Konkrétně půjde o data z atomistické molekulové dynamiky, coarse grain simulací, predikovaných ensemble struktur apod., které v tuto chvíli nemají adekvátní úložiště ani v České republice ani celosvětově. Repozitář bude zahrnovat i další vybrané skupiny strukturních dat, u nichž nyní chybí repozitář (např. anotace ke strukturám nukleových kyselin).

Repozitář bude testován s využitím dat, produkovaných institucemi, zapojenými v této podaktivitě (MU, UPOL a ÚOCHB AV ČR) a také spolupracujících institucí (Univerzita Karlova, VŠCHT). Poté, co bude repozitář vytvořen, bude mít každá laboratoř v ČR možnost nahrát sem svá simulační data (simulace biomolekul provozuje kromě univerzit i několik ústavů Akademie Věd – např. ÚOCHB, IBT, či BTU). Informace o vytvořeném repozitáři budou šířeny prostřednictvím bioinformatické infrastruktury ELIXIR a také na setkáních a konferencích strukturně bioinformatické komunity. Stejným způsobem šíříme informace o nových bioinformatických databázích a nástrojích. Repozitář bude fungovat na stejném principu jako Protein Data Bank – data budou nahrávat přímo uživatelé, takto získají perzistentní datový identifikátor, který mohou využít např. v publikaci.

Pro vyhledávání v rámci BioSimCZ budou využívány identifikátory, běžné pro biomakromolekulární struktury (PDB ID⁸, UniProt ID⁹, apod.) a využívané v rámci standardních bioinformatických databází Protein Data Bank, AlphaFold DB, UniProt, CATH spravovaných např. ESFRI infrastrukturou ELIXIR¹⁰. K vyhledávání budou rovněž sloužit přímo 3D struktury molekul (ve formátu PDB¹¹ a mmCIF¹²).

⁷ Kódy výstupů podaktivity jsou provázány na přílohu č. 3 Harmonogram klíčových aktivit a kap. 11 Hlavní plánované výstupy/produkty Studie proveditelnosti.

⁸ Protein Data Bank identifikační kód – identifikuje záznam v rámci databáze Protein Data Bank (<https://www.rcsb.org/>).

⁹ UniProt identifikační kód – identifikuje záznam v rámci databáze UniProt (<https://www.uniprot.org/>).

¹⁰ ELIXIR (<https://elixir-europe.org/>) je ESFRI projekt, zaměřený na vývoj a údržbu Evropské infrastruktury pro life sciences.

¹¹ PDB je datový formát souborů, využívaný v rámci databáze Protein Data Bank (<https://www.rcsb.org/>) a dalších databází biomakromolekulárních struktur.

¹² mmCIF (Macromolecular Crystallographic Information File) je datový formát souborů, využívaný v rámci databáze Protein Data Bank (<https://www.rcsb.org/>) a dalších databází biomakromolekulárních struktur.

Bude zajištěno propojení s následujícími projekty:

- BioExcel¹³ MDDb¹⁴: Projekt MDDb připravuje síť evropských uzlů, ukládajících v rámci jednotlivých států molekulárně dynamická data, generovaná v dané zemi. Tento distribuovaný projekt vyžaduje, aby v každém zapojeném státě vzniknul národní uzel MDDb, dedikovaný pro ukládání a správu molekulárně dynamických dat v tomto státě. Za architekturu, správu a zapojení národního uzlu do MDDb jsou zodpovědní pracovníci daného uzlu. BioSimCZ je plánován se stát českým národním uzlem MDDb.

BioSimCZ bude začleněn na seznam datových zdrojů ELIXIR CZ a stane se součástí databáze bio.tools¹⁵, která popisuje bioinformatické databáze a nástroje. Dále bude BioSimCZ propojen s klíčovými databázemi biomolekulárních struktur, spravovanými infrastrukturou ELIXIR (např. PDBe-KB¹⁶ či 3D-Beacons Network¹⁷). V rámci projektu bude zajištěna architektura, vývoj a testování. Po vytvoření repozitáře do něho budou importována data ze zdrojů, kde se tato data v současné době nacházejí (ZENODO, Figshare, OSF apod.). Poznámka: Nebudou tedy vytvořeny žádné nové datové sady. Tato data budou importována do repozitáře z následujících důvodů: Poslouží k otestování funkcionality a robustnosti databází, umožní daleko lepší další zpracování dat pomocí výpočetních nástrojů a motivují uživatele nahrávat do repozitáře svá data. Repozitář BioSimCZ bude součástí repozitářové platformy NRP, konkrétně bude implementován pomocí systému Invenio. Metadatové schéma BioSimCZ bude postaveno nad metadatovým schématem NRP, převeze jeho položky a doplní o další položky, které jsou specifické pro BioSimCZ).

Repozitář bude využívat metadatové schéma pro simulační strukturní data (např. molekulovou dynamiku, coarse grain simulace) a rovněž metadatové schéma pro další vybrané skupiny strukturních dat (např. specifické typy reziduí jako jsou nekanonické aminokyseliny, či nukleobáze). Metadatové schéma pro simulační strukturní data bude založeno na datových schématech simulačních nástrojů (např. GROMACS, AMBER apod.), zobecněno a rozšířeno o metadata pocházející od uživatelů (název projektu, kontaktní osoba, popis simulace atd.). Metadatové schéma pro další vybrané skupiny strukturních dat bude vytvářeno na základě strukturní abecedy dostupné v rámci databáze DNATCO¹⁸.

Dále bude repozitář využívat softwarový nástroj pro vyčítání metadat ze simulačních dat pro umožnění ukládání simulací do repozitáře a pro jejich převod do datového formátu dle výše uvedeného metadatového schématu pro simulační strukturní data. Tento nástroj bude také sloužit pro kontrolu vnitřní konzistence simulačních dat a pro převod dalších vybraných skupin strukturních dat do datového formátu dle výše uvedeného metadatového schématu pro tyto skupiny strukturních dat. Uvedený nástroj je nezbytnou součástí repozitáře, protože zaručuje FAIRness dat, importovaných do repozitáře.

Repozitář BioSimCZ bude vyvíjen ve dvou etapách, nejdříve jako prototyp a poté jako kompletní verze.

¹³ BioExcel(<https://bioexcel.eu>) je Evropské centrum excelence pro výzkum v oblasti výpočetní biologie.

¹⁴ MDDb je označení pro projekt MDposit (<https://mmb.mddb.eu/>).

¹⁵ bio.tools (<https://bio.tools/>) je databáze softwarových nástrojů a databází v oblasti life science.

¹⁶ PDBe-KB (Protein Data Bank in Europe – Knowledge Base, <https://www.ebi.ac.uk/pdbe/pdbe-kb/>) je databáze, shrnující různé vlastnosti struktur biomakromolekul.

¹⁷ Varadi, M., Nair, S., Sillitoe, I., Tauriello, G., Anyango, S., Bienert, S., ... & Velankar, S. (2022). 3D-Beacons: decreasing the gap between protein sequences and structures through a federated network of protein structure data resources. *GigaScience*, 11, giac118.

¹⁸ DNATCO (<https://dnatco.datmos.org/>), publikováno zde: Černý J., Božíková P., Malý M., Tykač M., Biedermannová L., Schneider B. (2020). Structural alphabets for conformational analysis of nucleic acids available at dnatco.datmos.org. *Biological Crystallography*, 76 (9), 805–813.

- **Prototyp** bude realizovat ukládání, sdílení a zpracování dat z molekulové dynamiky. Nezbytnou součástí prototypu bude nástroj pro vyčítání metadat z dat molekulové dynamiky a jejich převod do metadatového schématu, který zajišťuje FAIRifikaci dat a jejich import do repozitáře.
- **Kompletní verze** repozitáře bude vzhledem k prototypu rozšířena o ukládání, sdílení a zpracování dat z dalších typů simulací a o práci s dalšími skupinami strukturních dat. Nezbytnou součástí repozitáře bude i nástroj pro vyčítání metadat z uvedených typů simulačních dat a také z dalších skupin strukturních dat a jejich převod do výše uvedených metadatových schémat. Tento nástroj, který zajišťuje FAIRifikaci dat a jejich import do repozitáře, bude rovněž schopen realizovat kontrolu konsistence simulačních dat.

V rámci závěrečné fáze aktivity budou realizována školení pro koncové uživatele systému a pracovníky datové podpory a dalších odborníků v rámci komunity (vědce, datové kurátory, administrátory).¹⁹

ČINNOSTI:

- návrh architektury prototypu repozitáře a nástroje pro automatické vyčítání metadat;
- vytvoření metadatového schématu pro data z molekulové dynamiky **B_2.1**;
- implementace, konfigurace, deployment a testování prototypu repozitáře (včetně nástroje pro vyčítání metadat, s využitím vytvořeného metadatového schématu);
- vytvoření dokumentace k prototypu repozitáře (včetně nástroje pro vyčítání metadat a metadatového schématu) **B_2.2**;
- návrh architektury kompletní verze repozitáře (včetně architektury nástroje pro automatické vyčítání metadat);
- vytvoření metadatového schématu pro uvedená simulační data a také metadatového schématu pro další typy strukturních dat **B_2.3**;
- implementace konfigurace, deployment a testování kompletní verze repozitáře (včetně nástroje pro vyčítání metadat);
- vytvoření dokumentace pro kompletní verzi repozitáře (včetně nástroje pro vyčítání metadat a metadatového schématu) **B_2.4**;
- Návrh a implementace napojení kompletní verze repozitáře na mezinárodní datové zdroje (MDDb a zdroje ELIXIRu) **B_2.5**.

KÓDY VÝSTUPŮ PODAKTIVITY²⁰

B_2	Repozitář strukturních simulačních dat (BioSimCZ)
B_2.1	Dokumentace návrhu architektury prototypu repozitáře (včetně návrhu architektury nástroje pro vyčítání metadat a metadatového schématu)
B_2.2	Funkční prototyp repozitáře BioSimCZ
B_2.3	Dokumentace návrhu architektury kompletní verze repozitáře (včetně návrhu architektury nástroje pro vyčítání metadat a metadatových schémat)
B_2.4	Kompletní verze repozitáře BioSimCZ
B_2.5	Funkční propojení kompletní verze repozitáře napojena s MDDb a datovými zdroji ELIXIRu

¹⁹ Školení budou realizována ve spolupráci se Školícím centrem EOSC CZ, které zajistí odpovídající kapacity, technické zázemí a organizační podporu, a nejsou tak zahrnuta do rozpočtu OS II.

²⁰ Kódy výstupů podaktivity jsou provázány na přílohu č. 3 Harmonogram klíčových aktivit a kap. 11 Hlavní plánované výstupy/produkty Studie proveditelnosti.

10.1.3. PODAKTIVITA 2.3 – REPOZITÁŘ PRO LIDSKÁ I ANIMÁLNÍ OBRAZOVÁ A FYZIOLOGICKÁ MULTIMODÁLNÍ DATA (IMAGING REPOZITÁŘ)

Repozitář pro lidská i animální obrazová a fyziologická multimodální data (Imaging repozitář) bude sloužit k bezpečnému ukládání, správě a sdílení dat pořízených zejména magnetickou rezonancí (MRI), elektroencefalografií (EEG), pozitronovou emisní tomografií (PET) a počítačovou tomografií (CT). Repozitář bude podporovat ukládání dat z výzkumných studií s lidskými účastníky i z animálních studií.

Projekt se zaměřuje na následující klíčové oblasti:

- **Technická realizace repozitáře** v prostředí NRP, konkrétně s využitím repozitářového systému Invenio. Dostupné šablony uživatelských rozhraní budou přizpůsobeny tematicky zaměřeným metadatovým modelům. Tyto metadatové modely budou navrženy na základě analýzy ustavených ontologií a standardů (například OpenMINDS, BIDS) tak, aby reflektovaly specifika jednotlivých typů dat (např. lidská vs. animální data, mozková vs. kardiologická apod.). Výsledný návrh bude následně implementován do repozitářového systému.
- **Zajištění ochrany citlivých údajů** prostřednictvím řízení přístupu na úrovni jednotlivých datasetů s využitím AAI komponent dostupných v NRP pro repozitářový systém Invenio s cílem maximalizace principů FAIR při práci s citlivými daty. Vzhledem k různé míře citlivosti budou k dispozici různé režimy zpřístupnění dat – od volného přístupu, přes automatizované schvalování, až po ruční posouzení správcem datasetu. Veškerá neanonymizovaná lidská data budou zpracovávána na základě informovaného souhlasu.
- **Napojení na důvěryhodná prostředí (Trusted Research Environments, TRE)**, která umožní bezpečnou interaktivní i dávkovou práci s daty uloženými v repozitáři. Budou připraveny typické sady (kontejnerizovaných) nástrojů rozšiřující dostupná TRE a komponenty zajišťující zpřístupnění vybraných dat repozitáře v TRE na základě přístupových práv konkrétního uživatele. Tato integrace zajistí podporu nejen pro výzkumníky analyzující data, ale také pro datové kurátory při přípravě dat k publikaci a sdílení.
- **Propojení s evropskou výzkumnou infrastrukturou** na úrovni exportu vybraných metadatových záznamů, čímž bude zajištěna interoperabilita a širší využitelnost datasetů v repozitáři. Pro pokrytí neurovědní komunity zde počítáme s infrastrukturou EBRAINS, konkrétně budou připraveny nástroje pro přípravu datasetů pro kurátorský process EBRAINS Knowledge Graph se zaměřením na podporu datasetů spadajících do kategorie citlivých dat.
- **Testování a validace všech komponent repozitáře**, čímž bude zajištěna vhodnost funkčních aspektů (uživatelské rozhraní, nástroje, přístupová práva), stejně jako workflow správy a publikace dat. Zvláštní důraz zde bude kladen na práci s citlivými daty v souladu s etickými a právními požadavky.

ČINNOSTI:

- návrh technické architektury, autorizace a metadatového modelu;
- implementace repozitáře umožňující vývoj navazujících komponent a další rozvoj metadatových modelů;
- integrace systému autentizace a autorizace přístupu k datům na úrovni datasetů;
- návrh a implementace uživatelských rozhraní repozitáře (UI pro vyhledávání a přístup k datům; UI pro správce datasetu; UI pro data stewardy a kurátory);
- vývoj nástrojů pro datového stewarda a kurátora zahrnující nástroje pro anotaci, validaci a přípravu datasetů se zaměřením na interoperabilitu se standardem BIDS;
- propojení s TRE za účelem práce s citlivými daty;
- softwarová konfigurace instance TRE vyhovující hlavním cílovým komunitám;
- zajištění interoperability s evropskou digitální infrastrukturou pro výzkum;
- vývoj nástrojů pro export metadat a podporu kurátorského procesu pro EBRAINS Knowledge Graph;
- testování a validace;
- příprava dokumentace systému a uživatelské návody.

KÓDY VÝSTUPŮ PODAKTIVITY²¹

B_3	Repozitář pro lidská i animální obrazová a fyziologická multimodální data (Imaging repozitář)
B_3.1	Dokumentace technické architektury repozitáře a metadatových modelů
B_3.2	Funkční prototyp jádra repozitáře na platformě Invenio
B_3.3	Funkční uživatelská rozhraní repozitáře pro různé typy uživatelů
B_3.4	Napojení repozitáře na TRE s nástroji pro uživatele
B_3.5	Propojení repozitáře s EBRAINS Knowledge Graph
B_3.6	Funkční a validovaná plná verze repozitáře; dokumentace.

10.1.4. PODAKTIVITA 2.4 – REPOZITÁŘ PRO DATA CHEMICKÉ BIOLOGIE A JEHO PROPOJENÍ S VÝVOJEM A TRÉNINKEM AI MODELŮ MS

Repozitář se specializuje na biologická a chemická data.

Na základě analýzy, jaká data a jaké datové formáty jsou v komunitě používány, bude navržen vhodný datový model. Tento datový model by měl splňovat požadavky FAIR a měl by co nejvíce používat již existující a používané ontologie.

Pro takto navržený datový model bude navržena vhodná architektura a základní infrastruktura, umožňující ukládání dat do repozitáře. Předpokládá se, že vzniknou nástroje umožňující unifikování dat do navrženého datového modelu.

Tento systém bude rozšířen o možnost exportovat metadata, a kde je to možné i data, do vhodného RDF engine, který umožní prohledávání (meta)dat pomocí jazyka SPARQL. Hlavním cílem bude umožnit

²¹ Kódy výstupů podaktivity jsou provázány na přílohu č. 3 Harmonogram klíčových aktivit a kap. 11 Hlavní plánované výstupy/produkty Studie proveditelnosti.

federované dotazy, které umožní propojení uložených (meta)dat s jinými již existujícími zdroji dat postavenými na technologii Linked Data.

ČINNOSTI:

- Sběr požadavků od komunity (jaká data produkují a jaké datové formáty používají).
- Vytvoření datového modelu vhodného pro daná data a splňující požadavky FAIR. **B_4.1**
- Implementace nástrojů pro nahrávání a správu dat. Nástroje musí zajistit převoditelnost dat do vhodné FAIR formy. **B_4.2**
- Propojení a vyhledávání dat. Uložená data nebo jejich metadata musí být exportovatelná do SPARQL engine, který umožní jejich dotazování a propojitelnost pomocí federovaných dotazů s jinými (externími) SPARQL enginy. **B_4.3**
- Zajištění přístupu a workflow. Přístupy budou řízeny napojením na společnou autentizační a autorizační infrastrukturu. **B_4.4**
- Příprava testovacích datasetů. Testovací datasety budou založeny převážně na ukázkových datech poskytnutý jednotlivými partnery.

Propojení repozitáře chemické biologie s vývojem a tréninkem AI modelů MS

V rámci podaktivity bude vedle samotného repozitáře dat chemické biologie vytvořeno propojené prostředí pro vývoj, trénování a validaci modelů strojového učení (AI) zaměřených především na analýzu hmotnostních spekter. Repozitář bude zároveň sloužit jako zdroj anotovaných trénovacích dat i jako cílová platforma pro nasazení výsledných modelů, čímž vznikne uzavřený cyklus mezi ukládáním dat, jejich analýzou a zpětnou validací výsledků.

Bude implementován mechanismus pro průběžnou aktualizaci modelů na základě nově nahraných, kurátorsky ověřených dat. Systém bude podporovat dvě úrovně řízení modelového cyklu:

- Komunitní přístup (crowdsourcing) – výzkumníci budou moci nahrávat své vlastní datasety a přispívat k tréninku či testování modelů.
- Kurátorský režim – odborný tým bude dohlížet na kvalitu trénovacích dat, nastavovat validační scénáře a vydávat „produkční“ verze modelů.

Výsledné modely budou verzovány, dokumentovány a publikovány v rámci stejného infrastrukturního rámce jako repozitář samotný – přístupné přes API či webové rozhraní, a to jak pro výzkumné využití, tak pro integraci do dalších repozitářů a nástrojů.

Součástí vývoje bude rovněž vytvoření nástrojů pro:

- validaci trénovacích a testovacích datasetů;
- monitoring kvality predikcí a následnou adaptaci modelů;
- automatickou anotaci a návrh chemických struktur z hmotnostních spekter pomocí RDF a ontologií (např. ChEBI, PubChem);
- záznam o výpočetní historii a provenance dat i modelů, v souladu s FAIR principy.

Tato podaktivita propojuje repozitář dat chemické biologie s vývojem nástrojů umělé inteligence zaměřených na interpretaci hmotnostních spekter. Cílem je nejen zajistit bezpečné a standardizované ukládání dat, ale také vytvořit prostředí, ve kterém budou tato data využitelná pro automatickou anotaci, predikci chemických struktur a trénink AI modelů. Zvláštní důraz je kladen na otevřenost

a zapojení komunity prostřednictvím etablovaných procesů. Výsledné modely a nástroje budou poskytovány jako otevřené služby a napojeny na sémantické standardy (např. RDF, SPARQL) a znalostní báze, čímž bude zajištěna interoperabilita v evropském i globálním výzkumném prostoru. Tato podaktivita významně přispěje ke zrychlení a zpřesnění analýzy chemických dat, rozšíření otevřené vědy a posílení využití AI v oblasti bioinformatiky a chemické biologie. **B_5**

ČINNOSTI:

- návrh a vývoj nástrojů pro integraci trénovacího prostředí s FAIR repozitářem;
- vytvoření infrastruktury pro správu anotovaných trénovacích datasetů, včetně procesů validace a kurátorství;
- implementace mechanismů pro verzování a průběžnou aktualizaci modelů (např. retrainink, sledování výkonu);
- vývoj nástrojů pro trénink a validaci modelů ve dvou režimech: komunitním (crowdsourcing) a kurátorském;
- zpřístupnění výsledných modelů formou služby (např. API, uživatelské rozhraní, RDF exporty);
- vývoj nástrojů pro automatickou anotaci dat a návrh chemických struktur pomocí ontologií (ChEBI, PubChem aj.);
- příprava podpůrné dokumentace, návodů, příkladových datasetů a doporučení pro přispěvatele.

KÓDY VÝSTUPŮ PODAKTIVITY²²

B_4	Repozitář pro data chemické biologie
B_4.1	Metadatový model
B_4.2	SPARQL endpoint
B_4.3	Testovací dataset
B_4.4	Dokumentace
B_5	Propojení repozitáře chemické biologie s vývojem a tréninkem AI modelů pro MS

10.1.5. PODAKTIVITA 2.5 - TVORBA NOVÉHO REPOZITÁŘE CLINDATA A FAIRIFIKACE DAT

Klinický repozitář (ClinData) nabízí bezpečné a strukturované ukládání, správu a zpracování klinických dat (dat o lidech), především pro klinická data, která jsou součástí základního i aplikovaného klinického výzkumu. Dalšími typy ukládaných dat mohou být např. s klinickými daty pro konkrétní subjekt související data multiomická, obrazová, případně jiné typy dat. Repozitář zajišťuje možnost archivace uzavřených studií a integraci s multiomickými a obrazovými repozitáři. Dále umožňuje napojení dat na další relevantní repozitáře s využitím softwarově vytvořených konvertorů.

ClinData se stanou sama o sobě repozitářovým systémem NRP.

Umožní archivaci uzavřených studií (doplnění funkcionality repozitáře) a napojení na repozitáře multiomických a obrazových dat a také do metadatových adresářů a katalogů, například LifeData²³,

²² Kódy výstupů podaktivity jsou provázány na přílohu č. 3 Harmonogram klíčových aktivit a kap. 11 Hlavní plánované výstupy/produkty Studie proveditelnosti.

²³ LifeData portál: <https://lifedata.cz/>.

cBioPortal²⁴, MOLGENIS²⁵, ECRIN-ERIC²⁶ a BBMRI-ERIC²⁷. Vznikne tak unikátní platforma, která bude na jednom místě bezpečně uchovávat a integrovat data o lidech, která bude široce využitelná pro oblast výzkumu v medicíně, biologii i společenských a humanitních vědách.

Realizaci bude zaměřena mj. na ochranu citlivých dat a napojení na TRE²⁸.

Stávající portál ClinData je navržený pro průběžný sběr a management klinických dat. Tento systém bude doplněný o možnost trvalé archivace ukončených studií včetně dynamického metadatového modelu a harmonizace datového modelu v OMOP²⁹.

Stávající systém pro autentizaci a autorizaci uživatelů bude doplněný o možnost autentizace uživatelů například prostřednictvím LifeSci AAI. **B_6.1**

Sensitivní klinická data budou v rozsahu stanoveném uživatelem automaticky exportovatelná do relevantních oborových metadatových katalogů, kde budou analyzovatelná budoucími uživateli. Data pseudonymní nebo neanonymizovatelná bude možné zpřístupnit uživatelům na základě DTA³⁰ uzavřeném mezi vlastníkem dat a uživatelem přímo v zabezpečeném systému ClinData. **B_6.2**

Součástí dat o člověku jsou často genetická a jiná omická data i výsledky obrazových vyšetření (MRI, CT, PET/CT, PET/MRI, SPECT, ultrazvuk, atd., ale také fotografie, zvukové nebo audionahrávky. Abychom tato data mohli efektivně analyzovat na úrovni jednotlivce ve vazbě na jeho klinické informace (například laboratorní vyšetření, klinický obraz nemoci, odpověď na léčbu, přežití atd.), propojíme v repozitáři ClinData klinické, multiomické a obrazové informace vztahující se ke stejnému jedinci v podobě odkazů do interního nebo externího úložiště. **B_6.3**

Průběžné testování všech funkcionalit repozitáře. Příprava podpůrné dokumentace, service desku, školení uživatelů a pilotní nasazení systému pro klíčové vědecké komunity.³¹ **B_6.4**

ČINNOSTI

- návrh architektury pro vznik trvalého repozitáře pro klinická data;
- interoperabilita autorizace a autentizace uživatelů ClinData s LifeSci AAI **B_6.1**;
- napojení do oborových metadatových adresářů a zpřístupňování dat **B_6.2**;
- napojení do oborových repozitářů **B_6.3**;
- testování **B_6.4**.

²⁴ <https://www.cbioportal.org/>.

²⁵ <https://molgenis.org/>.

²⁶ <https://ecrin.org/>, European Clinical Research Infrastructure Network.

²⁷ <https://www.bbMRI-eric.eu/>.

²⁸ Trusted research environment.

²⁹ Observational medical outcomes partnership.

³⁰ Data transfer agreement.

³¹ Školení budou realizována ve spolupráci se Školicím centrem EOSC CZ, které zajistí odpovídající kapacity, technické zázemí a organizační podporu a nejsou tak zahrnuta do rozpočtu OS II.

KÓDY VÝSTUPŮ PODAKTIVITY³²

B_6	Repozitář ClinData
B_6.1	AAI
B_6.2	Metadatový model
B_6.3	Interoperabilita
B_6.4	Testovací dataset

10.1.6. PODAKTIVITA 2.6 – STANDARDIZACE A NASTAVENÍ POSTUPŮ PRO UKLÁDÁNÍ A SPRÁVU PŘÍSTUPU K DATŮM

Současný stav ukládání a správy dat z regulovaných klinických studií je charakterizován fragmentárním a nesystematickým přístupem, kdy nedostatečná standardizace brání efektivní interoperabilitě a identifikaci studií vhodných např. pro metaanalýzy.

Očekávaný cílový stav po naplnění aktivity zahrnuje vytvoření jednotných metodických postupů, které umožní nejen dodržení standardů ECRIN³³ a zvýšení kvality a dostupnosti dat, ale také integraci dat mezi oborově specifickými repozitáři, například mezi omickým repozitářem a repozitářem klinických dat. Implementace jednotné struktury metadat a dat také významně sníží náklady nutné na jejich harmonizaci a tím usnadní jejich sdílení v rámci evropské infrastruktury (např. GDI³⁴, FEGA, EOSC) (3.3). Adaptace existujících metadatových a datových modelů ECRIN zajistí plnou právní a etickou shodu s legislativními normami (GDPR, IPR), čímž se výrazně sníží bariéry pro mezinárodní a interdisciplinární výzkum.

Aktivita pomůže k odstranění překážek, které brání FAIRifikaci dat z regulovaných klinických studií, a tím zvýší dohledatelnost a dostupnost těchto dat širší odborné veřejnosti. Klinikům, výzkumníkům a dalším odborníkům přinese aktivita jednotný a interoperabilní systém pro správu dat z regulovaných klinických studií, což usnadní přístup k datům pro metaanalýzy a evidence-based medicínu.

Dvě hlavní tematické části metodiky – tvorba standardizovaného popisu zdrojových dat (metadata) a standardizace samotných dat (struktura, formát) – napomohou vytvoření interoperabilního systému, který odpovídá předpokladům ECRIN MetaData Repository (MDR) a Data Sharing Repository (crDSR). Umožní tak jednotnou a interoperabilní datovou infrastrukturu, která podporuje otevřenou vědu a efektivní sdílení dat mezi evropskými partnery.

Metodika FAIRifikace dat z ukončených regulovaných klinických studií bude vytvořena pro medicínská data. Metodika navazuje na repozitář klinických dat a také na omický repozitář, kde rozšiřuje jejich použití, propojení a interoperabilitu, zejména v kontextu výzkumné infrastruktury ECRIN.

ČINNOSTI:

- Rešerše struktury a formátu obecných metodik FAIRifikace dat a metodik FAIRifikace dat z klinického výzkumu.
- Stanovení metodických postupů pro standardizaci, ukládání a správu přístupů k datům z ukončených regulovaných klinických studií.

³² Kódy výstupů podaktivity jsou provázány na přílohu č. 3 Harmonogram klíčových aktivit a kap. 11 Hlavní plánované výstupy/produkty Studie proveditelnosti.

³³ ECRIN, the European Clinical Research Infrastructure.

³⁴ The Genomic Data Infrastructure (GDI).

- Adaptace ECRIN metadatového modelu pro metodiku FAIRifikace dat z ukončených regulovaných klinických studií. **B_7**
- Sběr zpětné vazby k metodice FAIRifikace dat z ukončených regulovaných klinických studií, její evaluace a implementace změn.
- Vytvoření finální verze metodiky FAIRifikace dat z ukončených regulovaných klinických studií. **B_8**

KÓDY VÝSTUPŮ PODAKTIVITY³⁵

B_7	Metodika FAIRifikace dat z ukončených regulovaných klinických studií
B_8	Adaptace ECRIN metadatového modelu pro metodiku FAIRifikace dat z ukončených regulovaných klinických studií

10.1.7. PODAKTIVITA 2.7 – VÝVOJ A PILOTNÍ ZAVÁDĚNÍ NÁSTROJŮ PRO ROZVOJ NDI ZAMĚŘENÝCH NA INTEROPERABILITU A UŽIVATELSKÝ KOMFORT

Aktuálně čelíme několika problémům, které projekt cíleně řeší. Současný stav je charakterizován řadou problematických bodů: Fragmentovaná vědecká data s nejednotnými formáty. Omezená interoperabilita mezi národními a mezinárodními repozitáři. Manuální a časově náročná analýza hmotnostních spekter. Nedostatečné využití AI pro automatizaci analytických procesů. Nízká úroveň školení pro práci s FAIR daty a umělou inteligencí.

Navrženy a vyvinuty budou nástroje pro „Repozitář pro biologická zobrazovací data“ (**B_03**). Jedná se o oborově specifické nástroje, které nejsou vyvíjeny v rámci NRP z důvodu tematicky omezené použitelnosti. Zároveň jsou tyto nástroje nezbytné pro zajištění širokého portfolia služeb NDI.

Vytvoříme standardizované datasety a AI benchmarky pro validaci a analýzu vědeckých dat uložených ve vytvářených repozitářích. Vývojem automatizovaných nástrojů pro správu a anotaci dat, které minimalizují manuální práci. Propojením národní a evropské vědecké infrastruktury pro snazší sdílení dat.

Budou navrženy, vyvíjeny a testovány nástroje řešící řadu témat:

- predikci chemických vzorců využívající vysoce přesná hmotnostní spektra a integraci standardizovaného datasetu a benchmarku pro umělou inteligenci (Chemical Formula Prediction Tool);
- tvorba standardizovaných workflow pro validaci dat z hmotnostních spektrometrů, zaměřených na reprodukovatelnost a interoperabilitu v oblasti chemické biologie (Validace a benchmarking datových sad);
- vývoj metadatových standardů a modelů zajišťujících interoperabilitu mezi stávajícími datovými úložišti, např. ELIXIR CZ, EGA (FAIRifikace datových sad);
- automatizace anotace dat hmotnostní spektrometrie pomocí RDF a napojení na otevřené znalostní báze, např. ChEBI, PubChem (Nástroje pro sémantickou anotaci dat).

Nástroje pro zvýšení uživatelského komfortu při práci s daty z biologického zobrazování **B_9**, zejména:

- anotační nástroje pro manuální, semiautomatické a automatické anotování dat včetně multimodálních, multidimenzionálních a časosběrných (vývoj nástrojů pro anotaci obrazových dat)

³⁵ Kódy výstupů podaktivity jsou provázány na přílohu č. 3 Harmonogram klíčových aktivit a kap. 11 Hlavní plánované výstupy/produkty Studie proveditelnosti.

ze světelné a elektronové mikroskopie potřebných pro tvorbu trénovacích a testovacích dat pro algoritmy strojového učení, konverze mezi různými anotačními formáty, kontrola konzistence anotací, provázání vyvinutých nástrojů s platformou pro správu dat a dalšími nástroji NDI);

- implementace a integrace platformy Onedata pro data management mikroskopických obrazů ve světelné mikroskopii (zajišťování sběru a přesunu dat z menších přístrojových PC na větší serverová úložiště, sdílení dat mezi interními a externími pracovníky, extrakce a prohlížení přístrojových metadat z nativních souborů hlavních výrobců světelných mikroskopů, přidělování perzistentních identifikátorů pro veřejné datasety, vytvoření přehledné dokumentace pro uživatele);
- implementace a integrace systému Omero s platformou pro správu dat a ostatními nástroji NDI (nasazení nástroje OMERO sloužícího k snadnému prohlížení, analýze a anotaci obrazových dat, vytvoření nástrojů pro automatizované importy dat a metadat, vývoj nástrojů zajišťující funkční provázanost systému OMERO s data management platformou a dalšími nástroji vytvářenými v rámci projektu, automatizované propojení obrazových a experimentálních dat, SSO autentizace a jednotná autorizace během komunikace mezi nástroji Onedata a databází Omero);
- rozvoj implementace iRODS pro management životního cyklu cryo-EM dat (vývoj a implementace nástrojů pro zpracování, správu a publikaci dat generovaných elektronovými mikroskopy, které se výrazně liší od světelných zmíněných výše, s využitím federativního cloudového řešení iRODS se zaměřením na tři klíčové oblasti: (1) rozšíření stávajícího systému přenosu dat na další typy vědeckých dat a jeho implementaci v jiných sdílených laboratořích, (2) vývoj řešení pro automatizovanou depozici cryo-EM dat do oborově specifických repozitářů EMDB a EMPIAR včetně integrace „single-click“ workflow pro snadnou publikaci, (3) rozšíření real-time analýzy dat a automatizace nastavení přístrojů s využitím metod strojového učení; cílem je vytvořit univerzální platformu, která usnadní efektivní správu dat v celém jejich životním cyklu, od akvizice přes analýzu až po dlouhodobé uchování a sdílení v souladu se zásadami European Open Science Cloud).

Služby budou dostupné a udržované díky podpoře ze strany infrastruktury Czech-Biolmaging.

ČINNOSTI:

- návrh a vývoj nástrojů, příprava datových modelů;
- zavedení prvních standardizovaných datasetů a AI modelů;
- integrace s EOSC a ELIXIR, testování s uživateli;
- pilotní implementace a testování;
- plná implementace, otevřený přístup k datovým modelům a workflow;
- dokončení dokumentace, spuštění služeb.

KÓDY VÝSTUPŮ PODAKTIVITY³⁶

B_9	Klíčový	Nástroje pro Repozitář pro biologická zobrazovací data
-----	---------	--

10.2. KA 3 – TEMATICKÉHO CLUSTERU MATERIÁLOVÉ VĚDY A TECHNOLOGIE

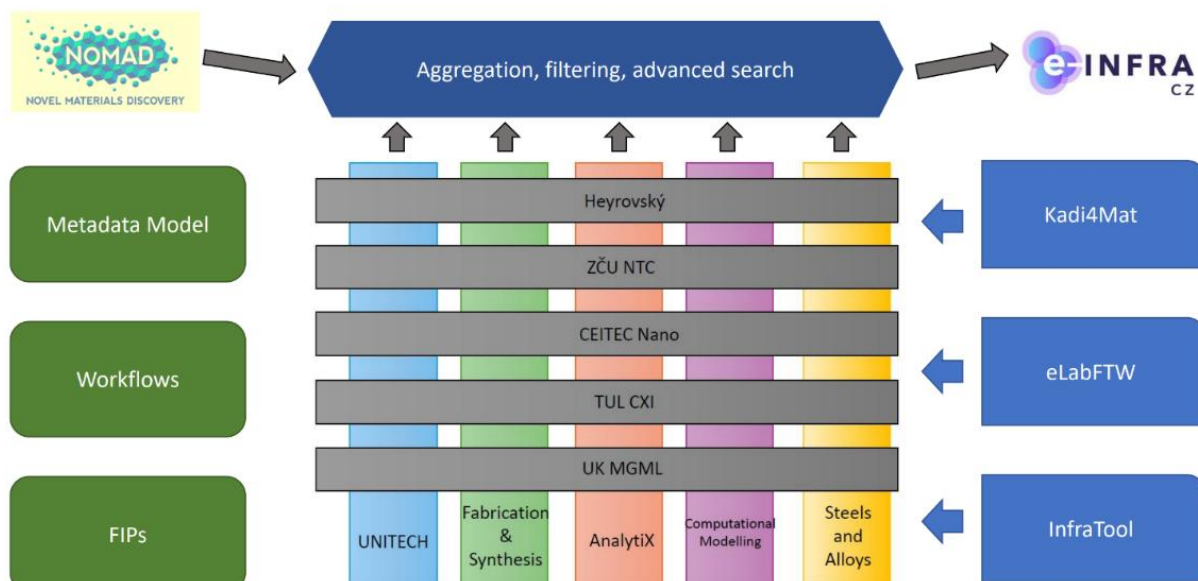
³⁶ Kódy výstupů podaktivity jsou provázány na přílohu č. 3 Harmonogram klíčových aktivit a kap. 11 Hlavní plánované výstupy/produkty Studie proveditelnosti.

ZAPOJENÍ PATRNEŘI: ÚFCH JH AV ČR (Garant TKA), VUT, ZČU, MU, UK

V návaznosti na projekt NRP a výstupy PKA vytvoří klíčová aktivita 3 prostředí pro kvalitní správu FAIR dat výzkumu v oblasti materiálových věd a technologií v České republice. Tyto obory v době počátku projektu nemají své oborové repozitáře, sémantické významy a modely jsou neustálené a mezinárodně neukotvené. V oboru chybí nástroje pro přímé ukládání dat do repozitářů, které by zkrátily dobu potřebnou pro kvalitní správu, ukládání a znovu využívání dat.

Ústřední aktivitou je zde implementace nového oborového repozitáře DANTE^c s adekvátním metadatovým profilem, výběrem relevantních licencí a uživatelským prostředím, které pomůže zlepšit ukládání, ale i vyhledávání a znovuvyužití výzkumných dat. Vznik nového repozitáře jednoznačně povede ke konsolidaci uživatelů z komunity oborového clusteru. Repozitář a tato KA se kromě obecného oborově-specifického úložiště dat zaměří také na rozšíření možností pro automatizaci jednotlivých kroků správy a FAIRifikace dat, včetně jejich vytěžování pro rozvoj ML-AI přístupů. Pro tento účel budou v repozitáři vytvořeny vysoce specializované kolekce s rozšířenými metadatovými profily pro detailní popis nejen ukládaných dat, ale i procesů spojených s tvorbou těchto dat a informacemi o provenance. Sémantické a strukturní nastavení repozitáře a jednotlivých kolekcí bude aktivně harmonizováno s podobnými aktivitami ve světě pro zvýšení jejich interoperability. Očekáváme, že tím dojde ke zviditelnění repozitáře nejen v komunitě samotné, ale i mimo ni, a ke zkvalitnění ukládaných FAIR dat.

V komunitě oborového clusteru byla identifikována potřeba pro ukládání výstupů z komplexních výzkumných a technologických, převážně automatizovaných, procesů. Pro tento účel bude repozitář DANTE^c adaptován pro práci s hierarchickými záznamy. Tím bude umožněno manuální, ale hlavně automatizované, ukládání dat z komplexních procesů (workflows) přímo do repozitáře, propojení souvisejících dat v repozitáři a záznam informací o provenance dle nejnovějších poznatků z oboru. Repozitář bude napojen na tři nástroje pro správu dat uživateli a v rámci výzkumných infrastruktur a jeden nástroj pro následnou práci s daty v repozitáři. Toto propojení přispěje ke zkvalitnění ukládaných oborově-specifických dat, ale i umožní široké mezioborové využití repozitáře pomocí moderních nástrojů na bázi ML a AI.



Obr. 4: Grafické schéma aktivit

10.2.1. PODAKTIVITA 3.1 – TVORBA NOVÉHO REPOZITÁŘE DANTE^c A FAIRIFIKACE DAT³⁷

Vznikne jeden nový repozitář DANTE^c s pěti kolekcemi (vlastní metadata, vlastní UI) pro specifická data. V rámci aktivity bude také konsolidováno pět komunit s vlastním kurátorem a pravidly pro proces schvalování záznamů (Heyrovský, NanoENVI, CEITEC Nano, NTC-ZČU, CXI TUL). Mimo komunitu Heyrovský se jedná o komunity z výzkumných infrastruktur. Vzniknou pravidla pro přístup k datům a online dokumentace ke všem částem repozitáře DANTE^c (včetně interaktivní nápovědy v rámci UI a uživatelsky-orientovaných pomůcek na komunitních stránkách datamaterialized.eu).

Pro jednotlivé kolekce repozitáře DANTE^c budou vytvořeny metadatové profily (5 kolekcí = 5 metadatových profilů), komunitní standardy (FAIR Implementation Profiles, FIPs³⁸) a licenční profily pro sdílení otevřených dat, ale i dat, u kterých bude potřebné kontrolovat přístup individuálně (citlivá data). Návrh pravidel a postupů pro práci s citlivými daty v repozitáři DANTE^c proběhne v rámci TKA SENSI. Před implementací do repozitáře DANTE^c proběhne analýza a sladění navržených standardů a procesů s mezinárodní komunitou. Podobně bude sledována možnost sdílení dat s jinými obory a řešení sledující původ a změny záznamů – provenance (viz plánovaná metodika). Zajištění kvality ukládaných dat bude probíhat pomocí školení a hackathonů.³⁹ Zde budou také řešeny různé přístupy k FAIRifikaci dat a jejich ukládání do repozitáře (např. s využitím nadstavby repozitáře pro teoretická data nebo z PKA oblast ELN).

V rámci projektu vytvoříme výstupy, které zkvalitní správu a sdílení výzkumných dat v novém repozitáři. Přínosy zahrnují: 1. pět nových metadatových modelů a FIPs zlepší FAIRifikovatelnost ukládaných dat⁴⁰; 2. vznikne jeden nový nástroj pro vyhledávání, filtrování a agregaci příbuzných dat pro další využití, např. pro trénování ML a AI. Bez této aplikace by byla příprava dat časově náročná a bez propojení s úložištěm dat NOMAD⁴¹ (rozvíjen konsorciem FAIRmat, NFDI, Německo). Nástroj zvýší znovuvyužívání (re-use) dat⁴²; 3. NeXuS⁴³ parser a metodika pro AI/ML zkrátí dobu zpracování dat do mezinárodních strojově zpracovatelných standardů a zvýší možnost znovuvyužití dat pomocí AI/ML; 4. výstupy zkvalitní práci v řadě institucí.⁴⁴

Vzhledem k celkovému rozsahu činností dělíme níže popis činností na jednotlivé logické celky.

1/ Nový oborový repozitář DANTE^c

Repozitář využije softwarový systém Invenio CESNET (z projektu NRP). Tím se zjednoduší napojení repozitáře DANTE^c na AAI eINFRA, vytěžování metadat do NMA a registrace v NKR. Vznik repozitáře dále zahrnuje:

- i. **Vytvoření politik** pro vkládání záznamů do kolekcí v repozitáři DANTE^c a následných schvalovacích a editačních/ kurátorských procesů.

³⁷ V popisu KA jsou plánované aktivity propojeny na realizované výstupy. Výstupy jsou označeny kódem, pro TKA MATECH v podobě **M_x** (klíčový), **M_x.x** (dílčí, tzn. náležící ke klíčovému).

³⁸ <https://www.go-fair.org/how-to-go-fair/fair-implementation-profile/>.

³⁹ Školení budou realizována Školicím centrem EOSC CZ, které zajistí odpovídající kapacity, technické zázemí a organizační podporu.

⁴⁰ Předpokládáme 40–50 záznamů se skóre FAIR v aplikaci F-UJI nad 59 %.

⁴¹ <https://nomad-lab.eu/nomad-lab/>; <https://www.fairmat-nfdi.eu/fairmat>.

⁴² Očekáváme 20 znovuvyužitých záznamů do konce roku 2028.

⁴³ <https://www.nexusformat.org/>.

⁴⁴ Očekáváme minimálně 20 kurátorovaných záznamů do konce roku 2026 s nárůstem minimálně 100 % v každém následujícím roce.

- ii. **Vytvoření uživatelských rozhraní** pro vkládání záznamů do kolekcí v repozitáři DANTE^c s využitím systému REACT/SB (včetně sladění designu s NRP/NDI).
- iii. **Nastavení licenčních profilů** pro záznamy v repozitáři DANTE^c s využitím výstupů KA4.2 projektu NRP.
- iv. **Administrace a správcovství** repozitáře DANTE^c, včetně všech pěti kolekcí.
- v. **Editace a kurátorství dat** vkládaných do kolekcí v repozitáři DANTE^c.
- vi. **Vytvoření online dokumentace**, nápovědy pro UI repozitáře a uživatelsky-orientovaných pomůcek v komunitním webu (datamaterialized.eu).

Repozitář DANTE^c bude na konci projektu provozovat 5 kolekcí s vlastním metadatovým profilem. Kolekce, tj. sekce pro ukládání různých typů dat budou zahrnovat: *UNITECH* – obecná část repozitáře bez přísné specifikace technických metadat, *Fabrication&Synthesis* – příprava materiálů a syntéza chemických látek, *AnalytiX* – charakterizace materiálů a analytické postupy, *Computational Modelling* – *in silico* modelování materiálů a jejich vlastností, a *Steels and Alloys* – databáze ocelí a slitin. Tyto kolekce byly identifikovány v komunitě na fyzických setkání PS MATECH EOSC CZ v roce 2024. Jednotlivé kolekce budou mít odlišné metadatové profily, a tak i odlišné uživatelské rozhraní pro vkládání záznamů člověkem, resp. odlišné schéma v API pro automatizované záznamy z přístrojů a ELN. Kolekce UNITECH bude využívat základní oborově-specifický metadatový model DANTE^{core}, který bude sdílený i s ostatními kolekcemi repozitáře DANTE^c pro vyšší interoperabilitu. V komunitě jsme se rozhodli využít faktu, že mezinárodní repozitáře a datové infrastruktury pro obory MATECH zatím neexistují. V tom se tato komunita liší od jiných TKA (např. BIO nebo HUMAN), které se v rámci NRP napojují na existující mezinárodní systém datových infrastruktur a repozitářů. To nám umožnilo sjednotit potřeby komunity do jedné instance repozitáře CESNET Invenio. Kolekce pak budou fungovat jako virtuálně oddělené repozitáře pro specifické potřeby podoborů. Zároveň bude jednodušší správa instance Invenio, napojených SW nástrojů a služeb a, předpokládáme, i interoperabilita dat⁴⁵.

Repozitář DANTE^c bude v průběhu projektu využívat minimálně 5 komunit s vlastním datovým kurátorem a workflows pro vkládání záznamů do repozitáře⁴⁶. Očekáváme, že se v první fázi budou datoví kurátoři formovat kolem kolekcí a budou sloužit zároveň celé komunitě oborového clusteru. Po zavedení dobrých praktik (předpoklad koncem projektu) budou datoví kurátoři sloužit primárně svým komunitám. Z důvodu procesu postupného napojování repozitáře na automatizační nástroje budou v průběhu projektu komunity používat převážně podobné workflow/politiky pro vkládání záznamů a jejich schvalování až po publikování. Očekáváme, že toto se může ke konci projektu změnit z důvodu specifických potřeb zapojených komunit. Repozitář bude postupně napojen na relevantní služby v NRP dle jejich dostupnosti v prostředí NRP.

Základní verze repozitáře DANTE^c napojeného na služby NRP bude využívat eINFRA AAI, umožňovat přidělování základní selekce licencí a práci s jedním metadatovým modelem (DANTE^{core}). V průběhu projektu pak budou na repozitář napojeny další služby, které vznikly nebo vznikají v rámci projektu NRP. Jde zejména o FAIRifikační nástroj⁴⁷, automatizované vkládání záznamů pomocí zjednodušených nástrojů⁴⁸ nebo ELN vytvářených v rámci PKA tohoto projektu a selekční nástroj pro výběr licence⁴⁹.

⁴⁵ Interoperabilita je zajištěna využitím totožného metadatového modelu, díky čemu budou mít shodný základ a tím budou interně interoperabilní.

⁴⁶ Komunity jsou obdobné jako u evropského repozitáře Zenodo. Očekáváme, že si komunity budou zakládat instituce, nebo jejich podjednotky. Komunity mají své „submission workflow“ a své kurátory.

⁴⁷ výstup KA 5.1 projektu NRP.

⁴⁸ výstup KA 5.3 projektu NRP.

⁴⁹ výstup KA 4.2 projektu NRP.

Výsledkem bude repozitář napojený na NRP, včetně vybraných pokročilých služeb a funkcemi pro práci se širším spektrem metadatových modelů včetně hierarchických záznamů z komplexních přístrojových systémů a výzkumně-vývojových procesů. Součástí nového repozitáře DANTE^c bude podrobná dokumentace v podobě nápovědy v uživatelském rozhraní a formou online dokumentu v dokumentovém úložišti NRP. Komunitní web datamaterialized.eu bude doplněn o zjednodušené návody a pomůcky pro rychlé zvládnutí práce v repozitáři DANTE^c. **M_1**

ČINNOSTI:

- repozitář DANTE^c napojený na služby NRP – základní verze s napojením na NMA, eINFRA AAI a základním seznamem licencí;
- implementace komunit a zapojení komunitních správců a kurátorů v repozitáři DANTE^c;
- repozitář napojený na služby NRP – produkční verze;
- plná funkcionality repozitáře, včetně automatického sběru dat a metadat. **M_1**

2/ Metadatové profily a standardy pro repozitář DANTE^c

Pro repozitář DANTE^c bude primárně rozšířen základní metadatový profil NMA (Czech Core Metadata Model; CCMM) pomocí ontologie Metadata4Ing⁵⁰ a korejského datového modelu pro popis materiálů⁵¹. Tak vznikne tzv. DANTE^{core} metadatový profil s ontologií pro oborově-specifická data s obecným využitím. Tento model bude využíván v kolekci UNITECH a bude základem pro modely ostatních kolekcí v repozitáři DANTE^c. Bude také snaha tento model harmonizovat s jinými oborově-specifickými repozitáři v NRP. Pro tento účel provedeme analýzu metadatových profilů využívaných v jiných TKA (např. TKA FYZIKA nebo TKA DM4AI). Ve specializovaných kolekcích (viz výše) se zaměříme na technické a procesní parametry spojené s danými podoblastmi výzkumu a s nimi spojenými datovými toky. Vyžijeme primárně kurátorované ontologie EMMO⁵² a PMDco⁵³ a obecnou terminologii definovanou v Schema.org⁵⁴. Specifické třídy a atributy pak umožní detailní záznam z výzkumných procesů ve výše uvedených specializovaných kolekcích. Jenom detailní záznamy s bohatými informacemi nejenom o samotných datech, ale i o navázaných procesech, jsou využitelné pro tvorbu kvalitních modelů ML a AI. Pro ověření funkcionality a interoperability repozitáře DANTE^c provedeme analýzu FAIR dat, která budou do repozitáře průběžně ukládána. Případné nedostatky využijeme pro úpravu repozitáře, nebo na repozitář napojených nástrojů. **M_1.1**

Metadatové profily budou doplněny o datové standardy definované ve FAIR Implementation Profiles (FIPs), které pro obory materiálůvých věd a technologií vytvoříme pomocí nástroje FIP Wizard⁵⁵ pro různé typy využití repozitáře DANTE^c. FIPs nadefinují standardy podporované a v některých případech i vyžadované danou kolekcí a budou využívány i pro FAIRifikaci dat. Očekáváme vytvoření více verzí FIP, a to i v případě, že pro danou specifikaci zatím nebude vytvořena kolekce v repozitáři DANTE^c. **M_2**

V průběhu projektu provedeme analýzu (meta)datových standardů užívaných nebo vyvíjených v rámci podobných aktivit v zahraničí. Typickým příkladem jsou aktivity v Německu (NFDI4Ing, NFDI-

⁵⁰ Ontologie vyvíjená konsorciem NFDI4Ing v Německu, <https://nfdi4ing.pages.rwth-aachen.de/metadata4ing/metadata4ing/>; <https://nfdi4ing.de/>.

⁵¹ autor Kwang-Ryeol Lee, National Centre for Materials Research and Development (NCMRD), Seoul, Jižní Korea; <https://github.com/krlee227/MatResData-Standard-Committee>.

⁵² <https://emmo-repo.github.io/>.

⁵³ <https://materialdigital.github.io/core-ontology/>.

⁵⁴ <https://schema.org/docs/schemas.html>.

⁵⁵ Výstup KA 5.1 projektu NRP.

Matwerk⁵⁶, NFDI4Chem⁵⁷, Platform Material Digital – PMD⁵⁸), Francii (DIADEM⁵⁹), Dánsku (CAPeX⁶⁰), Norsku (SINTEF⁶¹), Japonsku (DICE⁶²), Korei (NCMRD) a USA (MGI⁶³). Součástí této analýzy bude i monitorování a katalogizace technických rozhraní (API) využívaných SW systémy vyvíjenými v těchto aktivitách. Cílem je maximální interoperabilita řešení navržených pro repozitář DANTE^c. Tyto analýzy a práce na FIP pak budou úzce konzultovány na mezinárodní úrovni, konkrétně prostřednictvím pracovních skupin RDA (v současnosti existuje skupina „Harmonised terminologies and schemas for FAIR data in materials science and related domains“⁶⁴) nebo pomocí přímých jednání s koordinátory výše uvedených aktivit. S touto aktivitou souvisí zahraniční pracovní cesty z důvodu harmonizace standardů (metadatový profil, ontologie, kontrolované slovníky, formáty souborů) pomocí konzultací v online prostoru, ale i aktivní účast na různých fyzických konferencích a workshopech zaměřených na vyspělé materiály pro průmysl a jejich digitalizace, FAIR data a otevřená data. **M_1.1, M_1.2, M_2**

Zanalyzujeme potřeby integrace repozitáře DANTE^c se současnými standardy pro uchovávání informací o provenance. Výstupem bude definice potřeb souvisejících se vznikem dat na různých úrovních a v různých institucích pomocí provenance metadat, a to zejména pro účely dohledatelnosti původu/předchůdců datových sad, a to jak uživatelem žádajícím o přístup k datové sadě, tak i správcem repozitáře, resp. datovým kurátorem. Tato aktivita využije podklady z PKA podaktivity Provenance, aby informace o provenance byly v souladu s mezinárodním standardem ISO 23494, podkladovým datovým modelem CPM a přispěje ke znovuvyužití (R ve FAIR) dat uložených v repozitáři DANTE^c. **M_1.3**

ČINNOSTI:

- definice (meta)datových standardů – FIP; **M_2**
- vznik metadatového modelu DANTE^{core};
- rozšíření metadatového profilu pro jednotlivé kolekce, 5 kolekcí = 5 profilů; **M_1.1**
- analýza kompatibility sémantických standardů a SW rozhraní v mezinárodním prostředí; **M_1.2**
- analýza potřeb integrace repozitáře DANTE^c se současnými standardy pro provenance; **M_1.3**
- analýza datové výzvy a hackathonu.

3/ SW nástroj pro synchronní prohledávání a filtrování dat, včetně jejich extrakce, v repozitářích DANTE^c a NOMAD

Pro lepší využití dat ukládaných v repozitáři DANTE^c vytvoříme SW nástroj pro synchronní prohledávání a filtrování dat, včetně jejich extrakce, v repozitářích DANTE^c a NOMAD⁶⁵. NOMAD je v této době nejrozvinutější datový repozitář pro materiálové vědy v Evropě. Propojení repozitářů DANTE^c a NOMAD proběhne v několika fázích. Nejdřív provedeme mapování datových modelů DANTE^{core} a NOMAD. V případě potřeby bude navržen rozšířený metadatový model DANTE^{core+} pro lepší součinnost (interoperation) obou služeb. SW nástroj vyvineme tak, aby sloužil ke komunikaci s API obou repozitářů, nabízel unifikovaný způsob filtrování dat v repozitářích a jejich následné získávání

⁵⁶ <https://nfdi-matwerk.de/>.

⁵⁷ <https://www.nfdi4chem.de/>.

⁵⁸ Platform Material Digital, <https://www.materialdigital.de/>.

⁵⁹ <https://www.pepr-diaDEM.fr/le-pepr/>.

⁶⁰ <https://capex.dtu.dk/>.

⁶¹ <https://www.sintef.no/en/main-research-areas/materials/>.

⁶² <https://dice.nims.go.jp/>; National Institute of Materials Sciences, Tsukuba, Japan – <https://www.nims.go.jp/eng/>.

⁶³ Materials Genome Initiative, <https://www.mgi.gov/>;

⁶⁴ <https://www.rd-alliance.org/groups/harmonised-terminologies-and-schemas-fair-data-materials-science-and-related-domains-wg/members/all-members/>.

⁶⁵ Tj. samostatná služba s vlastním GUI doplňující současné způsoby vyhledávání a filtrace a agregace.

pro následné zpracování (primárně pro rozvoj AI nástrojů v materiálových vědách). Funkcionality filtrovacího a data procesujícího nástroje nastavíme i pro práci s hierarchickými záznamy z automatizovaných procesů. Nakonec budou pro tento SW nástroj připraveny detailní návody a metodika práce (v součinnosti s KA7 NRP). **M_1.4**

ČINNOSTI:

- mapování metadatových profilů DANTE^{core} a NOMAD,
- SW nástroj pro zajištění synchronizace práce s daty v repozitářích DANTE^c a NOMAD. **M_1.4**

4/ SW nástroj Invenio NeXuS „parser“ a metodika pro AI/ML

Vyvineme nástroj Invenio NeXuS „parser“ pro teoretická data a experimentální data. Jedná se o Plugin do Invenia, který zpracuje datasety nahrávané do repozitáře DANTE^c a automaticky opatří technickými metadaty. Vzniklý „parser“ přečte výstupní formáty teoretických výpočetních balíků nebo formát NeXuS, vstupně výstupní data teoretických modelů a další datové formáty a převede data do rozšířeného metadatového formátu repozitáře DANTE^c. Zde bude kontrolován také soulad s portálem NOMAD. Bude vytvořena metodologie pro FAIRifikaci dat určených pro komplexní analýzu experimentálních a teoretických materiálových dat (primárně pro AI/ML technologie) v repozitáři DANTE^c. **M_1.5, M_1.6**

ČINNOSTI:

- SW nástroj: Invenio NeXuS „parser“ pro teoretická a experimentální data; **M_1.5**
- metodika FAIRifikace dat pro AI/ML v repozitáři DANTE^c. **M_1.6**

KÓDY VÝSTUPŮ PODAKTIVITY⁶⁶

M_1	Repozitář DANTE ^c
M_1.1	Metadatové profily pro kolekce repozitáře DANTE ^c
M_1.2	Analýza kompatibility semantických standardů a SW v mezinárodním prostředí
M_1.3	Analýza potřeb provenance pro repozitář DANTE ^c se současnými standardy pro provenance
M_1.4	SW nástroj umožňující synchronní práci s daty v repozitářích DANTE ^c a NOMAD
M_1.5	SW nástroj Invenio NeXuS „parser“ pro teoretická data a experimentální data
M_1.6	Metodika FAIRifikace dat pro AI/ML v repozitáři DANTE ^c
M_2	FAIR Implementation profiles pro komunity v repozitáři DANTE ^c

10.2.2. PODAKTIVITA 3.2 – VÝVOJ A NAPOJENÍ NÁSTROJŮ A SLUŽEB PRO SBĚR DAT OD VÝZKUMNÝCH PRACOVNÍKŮ A Z INFRASTRUKTUR DO REPOZITÁŘE DANTE^c

Pro zlepšení prostředí, v kterém se spravují vědecká data, jsou potřebné jak nástroje přímé automatizace pro sběr a zpracování (meta)dat (např. z vědeckých přístrojů nebo v kombinaci s výpočetními kapacitami)⁶⁷, tak nástroje pro každodenní práci vědeckých pracovníků se svými daty.

⁶⁶ Kódy výstupů podaktivity jsou provázány na přílohu č. 3 Harmonogram klíčových aktivit a kap. 11 Hlavní plánované výstupy/produkty Studie proveditelnosti.

⁶⁷ Náplň projektu NRP (KA 5.3 a KA 5.4).

Mezi tyto nástroje patří elektronické laboratorní (terénní) deníky, tzv. ELN. PKA ELN se zabývá integrací dvou mezinárodně etablovaných ELN systémů (Kadi4Mat⁶⁸ a eLabFTW⁶⁹) v prostředí NDI.

Cílem této aktivity je asistovaný a poloautomatizovaný sběr záznamů do repozitáře DANTE^c. Repozitář napojíme na dva výše popsané existující SW nástroje pro uživatelskou správu dat (ELN). Nástroje zjednoduší každodenní práci s FAIR daty, a tím přispějí ke zkvalitnění sbíraných dat bez zbytečného zvyšování zátěže vědeckých a podpůrných pracovníků VO. Pomocí těchto nástrojů se FAIRifikace dat posouvá blíže k momentu jejich vzniku a umožňuje průběžný záznam informací o provenance. Napojení nástrojů na repozitář DANTE^c dále rozšíří možnosti práce s jedním ze základních repozitářových systémů NRP – CESNET Invenio.

V rámci této aktivity také vyvineme nový open-source SW balíček, který naplní požadavky správců infrastruktur, tj. řešení umožňující integrovanou správu uživatelů, přístrojů a produkovaných vědeckých dat výzkumné infrastruktury v jednom softwarovém systému. Tento SW systém napojíme na repozitář DANTE^c pro přímý sběr záznamů z výzkumných infrastruktur. Následně bude toto řešení implementováno ve dvou výzkumných infrastrukturách zaměřených na materiálový výzkum a vývoj technologií: MGML na MFF UK v Praze⁷⁰ a laboratoře infrastruktury CzechNanoLab⁷¹ (ve Fyzikálním ústavu AV ČR v Praze a CEITEC Nano na VUT v Brně⁷²).

Tato aktivita využije výsledky PKA oblast ELN, která integruje SW nástroje ELN Kadi4Mat a eLabFTW do prostředí NRP. V rámci této aktivity dojde k přímému napojení těchto ELN na repozitář DANTE^c a sladění metadatových profilů.

V rámci aktivity implementujeme ELN Kadi4Mat minimálně v pěti výzkumných skupinách a připravíme uživatelskou dokumentaci pro přímou depozici záznamů do repozitáře DANTE^c. Služba integrující ELN Kadi4Mat s repozitářem umožní přímý výběr dat v ELN a jejich odeslání do vybraných kolekcí repozitáře DANTE^c. Služba také umožní kontrolu povinných a doporučených metadat pro danou kolekci. V případě nesplnění podmínek pro záznam bude odesílatel informován o nesplnění podmínek nebo bude vystaven formulář pro doplnění chybějících metadat. Služba bude upravena i pro práci s hierarchickými záznamy z automatizovaných procesů. Vzniknou templáty (vzorové záznamy) pro ELN, které budou v souladu s profily kolekcí repozitáře DANTE^c. Dojde k úpravě repozitáře pro příjem dat z instancí ELN Kadi4Mat. V České republice zatím neexistuje obecný postup propojení ELN s vybraným repozitářem pro obory materiálových věd a technologií. **M_3**

Implementujeme také ELN eLabFTW minimálně v pěti výzkumných skupinách a připravíme uživatelskou dokumentaci pro přímou depozici záznamů do repozitáře. Dojde k vývoji služby a templátů a úpravě repozitáře DANTE^c podobně, jako je to popsané pro ELN Kadi4Mat. **M_4**

Vyvineme open-source SW nástroj (balíček) pro správu uživatelů, přístrojů a jejich datových výstupů pro velké výzkumné infrastruktury. Nástroj umožní registraci uživatelů do infrastruktury, kontrolu jejich účtů, nastavení přístupových práv k přístrojům a k datovým úložištím, rezervaci vědeckých přístrojů (booking), automatický sběr (meta)dat a kontrolu povinných metadat pro předdefinované úložiště. Tento nástroj také zabezpečí neměnnost vznikajících dat (transparentnost). Součástí balíčku je registrace přístrojů a jejich technických parametrů. Tyto parametry budou automaticky přikládány

⁶⁸ <https://kadi.iam.kit.edu/>.

⁶⁹ <https://www.elabftw.net/>.

⁷⁰ Materials Growth and Measurement Laboratory, <https://mgml.eu/>.

⁷¹ <https://www.czechnanolab.cz/>.

⁷² <https://nano.ceitec.cz/>.

k naměřeným datům pro zvýšení reproducibility měření. Nástroj umožní administrativní správu vědeckých přístrojů, včetně jejich údržby. **M_5**

Tento SW nástroj napojíme na repozitář DANTE^c pro asistovaný sběr záznamů z výzkumných infrastruktur. Tato služba umožní uživatelům infrastruktury přímo ukládat získané datové záznamy a jejich metadata (včetně podrobných technických parametrů) do vybraných kolekcí repozitáře DANTE^c. Nástroj bude upraven i pro práci s hierarchickými záznamy z automatizovaných procesů. Dojde k úpravě repozitáře pro příjem dat ze SW nástroje pro správu dat ve výzkumných infrastrukturách. Vytvoříme uživatelskou dokumentaci k vytvořenému SW balíčku pro implementaci ve výzkumných infrastrukturách. Na základě provedených uživatelských testů (včetně hackathonů) následně dotvoříme metodologii pro používání SW nástrojů a ukázkové příklady použití. Tento SW nástroj implementujeme ve dvou českých infrastrukturách (CzechNanoLab a MGML). **M_6**

ČINNOSTI:

- napojení ELN Kadi4Mat na repozitář DANTE^c; **M_3**
- implementace ELN Kadi4Mat ve výzkumném prostředí materiálových věd a technologií;
- napojení ELN eLabFTW na repozitář DANTE^c; **M_4**
- implementace ELN eLabFTW ve výzkumném prostředí materiálových věd a technologií;
- komplexní SW nástroj pro správu uživatelů, přístrojů a jejich datových výstupů ve výzkumných infrastrukturách, včetně metodiky; **M_5**
- integrace SW nástroje pro správu uživatelů, přístrojů a jejich datových výstupů s repozitářem DANTE^c;
- implementace SW nástroje pro správu uživatelů, přístrojů a jejich datových výstupů ve výzkumných infrastrukturách MGML a CzechNanoLab. **M_6**

KÓDY VÝSTUPŮ PODAKTIVITY⁷³

M_3	Integrace ELN Kadi4Mat s repozitářem DANTE ^c
M_4	Integrace ELN eLabFTW s repozitářem DANTE ^c
M_5	SW nástroj pro správu uživatelů a přístrojů a jejich datových výstupů ve výzkumných infrastrukturách
M_6	Integrace SW nástroje pro správu uživatelů a přístrojů a jejich datových výstupů ve výzkumných infrastrukturách s repozitářem DANTE ^c

10.3. KA 4 – TEMATICKÉHO CLUSTERU DATA MANAGEMENT PRO UMĚLOU INTELIGENCI A STROJOVÉ UČENÍ

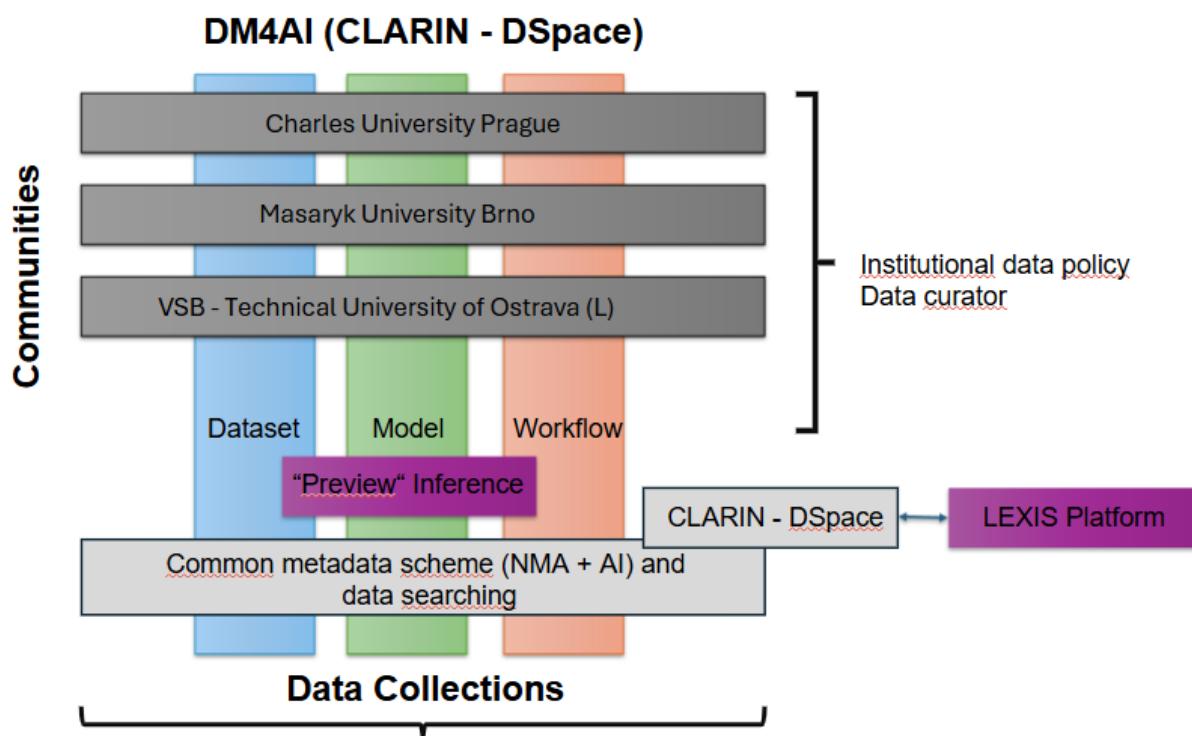
ZAPOJENÍ PARTNEŘI: VŠB – TUO (Garant TKA), MU, UK

V České republice doposud neexistuje centralizovaná platforma pro sdílení a ukládání AI/ML dat/modelů/workflow, existují jenom na sobě často nezávislé iniciativy a projekty, které často využívají platformy jako GitHub nebo Hugging Face pro sdílení svých modelů a dat a zvyšují tak svou závislost na zahraničních platformách. Cílem této aktivity je tedy zřídit podobně inspirovanou platformu/repozitář, zasazený do českého prostředí, a to včetně českého a evropského právního prostředí v případě užití AI nad citlivými daty. Ten výrazně zvýší kontrolu nad daty, výstupy a vývojem

⁷³ Kódy výstupů podaktivity jsou provázány na přílohu č. 3 Harmonogram klíčových aktivit a kap. 11 Hlavní plánované výstupy/produkty Studie proveditelnosti.

technologií v oblasti AI/ML a dojde ke konsolidaci uživatelské komunity a posílení nezávislosti na zahraničních infrastrukturách.

Tvorba nového, oborově zaměřeného AI/ML (Artificial Intelligence/Machine Learning) repozitáře Data management pro umělou inteligenci a strojové učení (DM4AI) založeného na repozitářovém systému CLARIN DSpace má za cíl nabídnout jednotnou platformu inspirovanou globální aplikací Hugging Face, která umožní efektivní sdílení a správu AI/ML modelů, datasetů a workflow, a to včetně poskytnutí pokročilých nástrojů pro práci s daty a možnostmi napojení na výpočetní infrastruktury prostřednictvím LEXIS Platformy.



Obr. 5: Schéma aktivit

Nová repozitářová platforma (repozitář + všechny její nástroje) bude odrážet AI/ML doménová specifika a bude jim plně přizpůsoben jak po stránce struktury, tak i funkcionality, aby co nejlépe podporoval práci s daty, modely a AI/ML workflow. Z hlediska AI/ML specifík přinese tento nový repozitář několik klíčových vlastností pro výzkumníky a komunitu DM4AI:

- Robustní, komplexní a strukturovaná metadata, která zohledňují specifické atributy AI/ML dat, modelů případně licencí.
- Oddělení klíčových částí AI/ML workflow (data, modely), zvyšuje přehlednost a znovu použitelnost dat.
- Data v repozitáři budou spravována v souladu s principy FAIR, čímž se zajistí jejich maximální kvalita a využitelnost. V oblasti umělé inteligence je kvalita dat naprosto zásadní – model je jen tak dobrý, jako data, na kterých se učil (např. problém se snížením datové zaujatosti).
- Umělá inteligence je silně závislá na výpočetních zdrojích, protože složité modely a algoritmy vyžadují obrovské množství výpočetní kapacity pro efektivní trénování a inferenci, a proto AI/ML repozitář bude poskytovat možnost napojení na výpočetní infrastruktury a zároveň poskytovat funkci zjednodušeného náhledu na AI/ML modely přímo v prostředí repozitáře („náhled“ inference).
- AI/ML repozitář umožní lepší kontrolu nad daty a jejich původem/transformacemi, licencemi

a dodržováním AI regulací lokální/globální (např. AI Act⁷⁴) a využití AI pro inovace.

Členové odborného týmu zapojeni do řešení projektu budou absolvovat zahraniční služební cesty za účelem získání nových znalostí formou účasti na zahraničních konferencích/seminářích/workshopech, navázání kontaktů se zahraničními odborníky a šíření výsledků projektu. Konkrétně se jedná zejména o cesty spojené se zapojením do evropského EOSC – participace na pracovních skupinách (EOSC task forces), EOSC Symposium, zimních/letních školách EOSC a souvisejícími outreach aktivitami, dále např. konference Open Repositories, COLING nebo ACL.

10.3.1. PODAKTIVITA 4.1 – TVORBA REPOZITÁŘOVÉ INFRASTRUKTURY DM4AI S PODPOROU FAIR PRINCIPŮ A MEZIOBOROVÉ INTEROPERABILITY⁷⁵

Cílem aktivity je vytvořit oborový repozitář specifický pro oborový cluster Data management pro umělou inteligenci a strojové učení, poskytující přístup k datasetům, jejich metadatům, AI/ML modelům a jejich workflow, a to včetně možnosti napojení na výpočetní infrastruktury. Bude nasazena instance softwarového repozitářového systému CLARIN-DSpace.

Repozitář bude určen pro datové sady (existující či nové) vytvořené výzkumníky s vazbou na AI/ML a bude také zahrnovat datové sady (existující či nové) vytvořené jinými subjekty, s nimiž čeští výzkumníci pracují a které nejsou dostupné v jiných repozitářích. Uživatelé poskytneme možnost ukládat informace o celé AI/ML workflow, které popisují použitá vstupní data (dostupná v tomto repozitáři, případně formou odkazu na externí datový zdroj), aplikací pro AI/ML trénování (použitý algoritmus a cluster pro jeho spuštění, včetně konfigurace prostředí) a výsledný model (např. váhy neuronové sítě), což dohromady tvoří FAIR Digital Object. Repozitář bude umožňovat efektivní správu dat a možnosti vyhledávání.

Vzhledem k celkovému rozsahu činností dělíme níže popis činností na jednotlivé logické celky.

1/ Nový repozitář DM4AI a rozvoj kompetencí a šíření potřebného know-how do oborově orientovaných výzkumných komunit

Při tvorbě repozitáře bude využita infrastruktura NRP, platforma LEXIS, softwarová platforma CLARIN-DSpace a výkonné výpočetní prostředky e-Infra (např. výpočetní cluster Karolina). Repozitář bude primárně využívat EOSC/e-INFRA CZ společnou autentizační a autorizační infrastrukturu (AAI) a bude řádně registrován v NKR.

Plně využijeme CLARIN-DSpace vyvíjený v rámci NRP s vytvořením specifického metadatového modelu pro AI/ML modely. Grafické uživatelské rozhraní repozitáře bude navrženo tak, aby vyhovovalo výše uvedeným požadavkům a bude inspirováno mj. uživatelským rozhraním, jež poskytuje populární služba Hugging Face⁷⁶ (konkrétně: Model Hub⁷⁷, Datasets Library⁷⁸, Train⁷⁹ na externích výpočetních zdrojích beroucí v potaz specifika AI a ML komunity). Rozhraní nasazené instance CLARIN-DSpace bude přizpůsobeno tak, aby odpovídalo potřebám oborové skupiny, a to včetně vytvoření potřebných

⁷⁴ AI Act | Shaping Europe's digital future.

⁷⁵ V popisu KA jsou plánované aktivity propojeny na realizované výstupy. Výstupy jsou označeny kódem, pro TKA DM4AI v podobě **AI_x** (klíčový), **AI_x.x** (dílčí, tzn. náležící ke klíčovému).

⁷⁶ <https://huggingface.co>. Platformy jako např. Kaggle nebo Hugging Face poskytující podobné funkcionality, připojené na komerční poskytovatele výpočetních zdrojů. Tyto platformy dovoluují uživateli zjednodušenou přímou interakci s vybraným modelem pro první krátké posouzení.

⁷⁷ <https://huggingface.co/docs/hub/en/models-the-hub>.

⁷⁸ <https://huggingface.co/docs/datasets/en/index>.

⁷⁹ <https://huggingface.co/docs/transformers/en/training>.

metadatových schémat a integrace s platformou LEXIS⁸⁰ pro distribuované řízení dat a spouštění komplexních výpočetních úloh na výkonných clusterech. AI_1

Nedílnou součástí klíčové aktivity je tvorba tutoriálů a specializovaných podkladů určených pro nové uživatele infrastruktury z komunity AI/ML budou vytvářené napříč klíčovými činnostmi v oborovém clusteru, zejména budou provázané s aktivitami spojených s tvorbou a užíváním oborově specifického repozitáře a software (SW). Mezi hlavní cíle výstupu patří: (1) Příprava analytických materiálů souvisejících s potřebami správy výzkumných dat ve vazbě na TKA Data management pro umělou inteligenci a strojové učení. (2) Příprava instruktážních výukových materiálů na principu „how to do“ (repozitář, software) / návodů/ vzdělávacích materiálů pro uživatele/ přednášky/ metodiky (3) Školení uživatelů.⁸¹ AI_1.11

ČINNOSTI:

- založení AI/ML specializovaného oborového repozitáře (instance);
- přizpůsobení grafického uživatelského rozhraní systému CLARIN-DSpace;
- testování a pilotní provoz repozitáře;
- poskytnutí přístupu výzkumníkům ze skupiny DM4AI a vylepšení systému na základě jejich zpětné vazby (pilotní fáze);
- DM4AI začne využívat repozitář pro svá data; AI_1
- přenos know-how, tvorba tutoriálů. AI_1.11

2/ Zajištění efektivního vyhledávání dat/meta-dat

Vytvoříme metadatový model, který bude AI/ML⁸² oborově specifický a bude navazovat na NMA (integrace s Národním metadatovým adresářem je nutná z důvodu poskytování povinných metadat do NMA). Model bude pokrývat všechny aspekty interoperability, které jsou nezbytné pro další bezproblémové používání, resp. sdílení. Metadatový model bude ještě v rámci projektu udržován a pravidelně aktualizován, aby odrážel všechny klíčové vlastnosti uložených dat. Tvorbu metadatového schéma bude doprovázet řada analýz a studií, které budou pokrývat: (i) informace o stávajících schématech, (ii) identifikaci rozsahu potřeb a (iii) určení struktury schématu. AI_1.1

Naprogramujeme SW pro správu dat a možností vyhledávání a integrujeme ho do repozitáře jako rozšíření funkcionalit poskytovaných CLARIN-DSpace. Repozitář tak bude umožňovat efektivní vyhledávání dat/meta-dat, tak aby bylo možné snadno identifikovat relevantní data a připravit je pro rozsáhlé učení nebo inferenci. Specializované datové struktury, které odrážejí typ požadavků na dotazy a strukturu dat/meta-dat, jsou předpokladem pro efektivní možnosti vyhledávání dat. Příkladem mohou být kategoriální data, časové řady, grafy, obrazová a video data nebo 3D mračna bodů a sítě. Repozitář bude podporovat jako jeden z typů uložených objektů (pro replikovatelnost experimentů) popis celé AI/ML workflow včetně přijatých dat/meta-dat, aplikací pro AI/ML trénování (např. zdrojový kód, binární soubory, kontejnery – verzování), natrénovaných modelů a výsledků. AI_1.2

ČINNOSTI:

- vytvoření a údržba oborově specifického metadatového modelu pro AI/ML;
- zajištění interoperability pro sdílení a používání metadatového modelu;

⁸⁰ <https://www.lexis.tech>. LEXIS Platforma – pokročilá platforma, která poskytuje snadný a bezpečný přístup k vysokovýkonným výpočetním zdrojům – včetně webového rozhraní pro distribuovanou správu dat, spouštění HPC aplikací a orchestraci složitých workflows napříč více lokalitami.

⁸¹ Školení budou realizována ve spolupráci se Školícím centrem EOSC CZ, které zajistí odpovídající kapacity, technické zázemí a organizační podporu a nejsou tak zahrnuta do rozpočtu OS II.

⁸² Artificial intelligence (AI) / machine learning (ML).

- správa dat a podpora efektivního vyhledávání dat a metadat; AI_1.1
- definování kritérií pro efektivní identifikaci relevantních dat;
- výběr vhodných struktur pro různé typy dat (kategoriální data, časové řady, grafy, 3D mračna bodů...);
- optimalizace pro efektivní vyhledávání a dotazování;
- implementace funkcí pro ukládání a verzování AI/ML workflow. AI_1.2

3/ Podpora činností nezbytných pro FAIRifikaci dat v repozitářích, zajištění provenance a kurátorství

Využijeme nástroje pro generování a správu standardizované provenance, která budou vytvořeny v rámci PKA podaktivity Provenance, a to konkrétně: 1/ SW pro ukládání a správu provenance; 2/ SW knihovna pro práci s provenance dle CPM. Nástroje budou integrovány s repozitářem a SW pro přenosy dat dle provenance metodiky a stanou se jejich součástí tak, aby bylo umožněno poskytnutí strojově zpracovatelné dokumentace původu modelů a dat v standardizované reprezentaci (vazba na standardy W3C PROV⁸³, Common Provenance Model:ISO 23494⁸⁴, Workflow Run RO-Crate⁸⁵).

V současnosti neexistuje SW řešení, které by umožňovalo ukládání, správu a použití AI modelů a dat v souladu s uvedenými provenance standardy. Z toho důvodu se nedá předpokládat, že kompletní dokumentace AI modelů (natrénovaný model -> datová sada vstupující do trénování a testování modelu -> „zdrojová“ datová sada -> vzorky, z kterých byly sady generovány -> původní zdroj vzorky) bude interoperabilní. Umožnění existence a vyhledatelnosti interoperabilní a důvěryhodné provenance, která by dokumentovala celý životní cyklus modelů a souvisejících datových sad s důrazem na zajištění jejich důvěrnosti a soukromí souvisejících osob.

Bude vytvořena metodika obsahující postupy pro naplňování FAIR principů u výzkumných dat, se zohledněním specifik domény umělé inteligence (etické otázky, bias, komplexita), tj. Metodika přípravy, anotaci a kurátorství dat.

Metodika bude popisovat přípravu, anotaci a kurátorství přijatých dat z pohledu FAIR principů s ohledem na potřeby AI/ML workflow a souvisejících dat. To bude zahrnovat například metodické postupy, jak používat nástroje, které kombinují strojové učení s lidskou verifikační smyčkou nebo automatickou schopností odstraňování duplicit.

V současné době vznikly různé nástroje, iniciativy a metodiky sloužící jako návod pro tvůrce dat, na jejichž základě lze implementovat FAIR principy. Naprostá většina z nich však pracuje na obecném, a nikoli doménově specifickém základě. Nástroje jako FAIR Wizard⁸⁶ nebo FORCE 11⁸⁷ pracují pouze s obecnými prvky FAIR principů a nezohledňují tak specifika jednotlivých oborů. Dostupné jsou však specifické FAIR principy pro sociální (Data Management Expert Guide⁸⁸) a přírodní vědy (RDMkit⁸⁹). Komplexní metodika z pohledu FAIR principů s ohledem na potřeby AI/ML workflow a souvisejících dat zatím v oboru není dostupná.

ČINNOSTI:

- definice požadavků týkajících se ověřování znovupoužitelnosti AI modelů v souvislosti s provenance;

⁸³ <https://www.w3.org/TR/prov-overview/>.

⁸⁴ <https://www.bbmri-eric.eu/news-events/common-provenance-model-bbmri-led-iso-ts-23494-12023-released/>.

⁸⁵ <https://www.researchobject.org/workflow-run-crate/>.

⁸⁶ <https://fair-wizard.com/>.

⁸⁷ <https://force11.org>.

⁸⁸ <https://dmeg.essda.eu/Data-Management-Expert-Guide>.

⁸⁹ <https://rdmkit.elixir-europe.org/>.

- analýza možností adresování požadavků z předchozího bodu v repozitáři pomocí výstupů PKA podaktivity Provenance;
- návrh architektury pro technickou integraci výstupů PKA podaktivity Provenance;
- implementace architektury z předchozího bodu, pilotní provoz, demonstrace využití; **AI_1.3**
- definice klíčových principů a standardů pro přípravu, anotaci a kurátorství dat v oblasti AI/ML;
- vývoj a formulace metodických kroků a doporučení;
- popisování postupů přípravy dat v souladu s FAIR principy;
- dokumentace metodiky a její distribuce mezi cílové skupiny. **AI_1.10**

4/ Návrh mechanismů licencování AI modelů

V oblasti licencování bude aktivita fázována takto: (i) právně-technologická analýza, (ii) vypracování sady vzorových licenčních smluv, (iii) vytvoření a pilotní implementace metadatového modelu, (iv) začlenění a ověření specifik pro AI nad citlivými daty, (v) validace a úprava navržených řešení.

Zanalyzujeme povahu AI modelů z právně-technologického pohledu. Výsledný dokument bude shrnovat právní a technologické aspekty AI modelů v kontextu klasifikace jako software, databáze či speciální typ duševního vlastnictví. Obsahuje přehled relevantních právních předpisů (evropských i národních), licenčních mechanismů a doporučení pro praxi. **AI_1.4**

Vytvoříme Licenční modely a vzorové smlouvy pro AI modely. Jednotná sada licenčních modelů a vzorových smluv (např. pro vícepartnerové projekty, práci se studenty, řešení citlivých dat) umožňuje jasné vymezení práv a povinností při tvorbě, sdílení a využití AI modelů, reflektuje zároveň požadavky GDPR⁹⁰ a obchodního tajemství.

V rámci projektu se zaměřujeme na vytvoření konceptu licencování AI modelů, který zohledňuje evropský právní kontext a současné trendy v ochraně počítačových programů, datových sad i výsledků modelů. Projekt identifikuje možnosti licencování zdrojových kódů AI modelů (např. GPL⁹¹, Apache, MIT⁹², BSD⁹³), ochranu trénovacích parametrů formou databází (podle směrnice 96/9/ES a licencí typu ODbL⁹⁴ či PDDL⁹⁵), a také specifické licence přizpůsobené AI modelům, jako je BigScience OPEN RAIL⁹⁶ nebo koncepty typu Model Openness Framework⁹⁷ (MOF).

Vyvíjený koncept je orientován na dva hlavní scénáře: situace, kdy do vývoje AI modelů přispívá více partnerů uvnitř jedné instituce nebo napříč institucemi, a situace sdílení hotových AI modelů směrem ke koncovým uživatelům i dalším vývojářům. Zvláštní pozornost je věnována doménám pracujícím s citlivými daty, kde AI modely mohou tato data přímo obsahovat. **AI_1.5**

Licenční modely a vzorové smlouvy integrujeme s metadatovým modelem (Metadatový model pro popis licenčních aspektů AI modelů), kde aspekty související s licencemi musí být součástí metadat. Metadatový model pro popis licenčních aspektů AI modelů, bude strojově čitelný metadatový model zachycující licenční a právní informace, včetně aspektů ochrany citlivých dat či autorských práv. Model bude navržen tak, aby byl kompatibilní s FAIR principy i s národními a evropskými standardy (např. návaznost na Národní metadatový adresář). **AI_1.6**

⁹⁰ General Data Protection Regulation (GDPR).

⁹¹ General Public License (GPL).

⁹² Massachusetts Institute of Technology.

⁹³ Berkeley Software Distribution (BSD).

⁹⁴ Open Data Commons Open Database License (ODbL).

⁹⁵ Planning Domain Definition Language (PDDL).

⁹⁶ <https://bigscience.huggingface.co/blog/the-bigscience-rail-license>.

⁹⁷ <https://isitopen.ai>.

ČINNOSTI:

- právně-technologická analýza povahy AI modelů; AI_1.4
- vzorové smlouvy AI modelů dalším uživatelům;
- vypracování sady vzorových licenčních smluv pro AI modely; AI_1.5
- vytvoření a pilotní implementace metadatového modelu;
- začlenění a ověření specifik pro AI nad citlivými daty;
- validace a úprava navržených řešení;
- metadatový model popisující licencování AI modelů strojově čitelným způsobem, v návaznosti na analýzu a modelové licenční smlouvy;
- metadatový model popisující využití datové zdroje pro trénink modelů. AI_1.6

5/ Validace vytvořených metadatových modelů a depozičních metodik pomocí AI modelů zpracovávajících citlivá medicínská data

Další náplní aktivity bude validace vytvořených metadatových modelů a depozičních metodik vyvinutých iniciálně pro AI systémy běžně trénované na otevřených datech. Pilotní nasazení těchto modelů na citlivých datech – nebo na kombinaci citlivých a otevřených zdrojů – umožní ověřit jejich použitelnost, spolehlivost a výkonnost i při zpracování chráněných medicínských či biopotravinařských dat. Součástí řešení bude úzká spolupráce s expertními skupinami TKA SENSI a pracovním týmem TKA B/H/F, díky čemuž zajistíme, že validační mechanismy budou plně reflektovat specifika obou domén. V rámci pilotního ověřování nebude probíhat výzkum, ale pouze validace architektury a metadatových modelů propojujících citlivá data a citlivé nebo potenciálně citlivé AI modely, a bude využívat existujících vyřešených případně aktuálně řešených problémů. Využijeme existující pipeline výzkumné skupiny RationAI na MU, která úspěšně vyvíjí modely s nasazením v klinické praxi, spolu s citlivými daty z několika různých medicínských institucí.

Aktivita bude primárně validovat metadatové modely a depoziční metodiky a bude poskytovat vstup pro jejich vývoj, a nevyžaduje tudíž samostatnou udržitelnost nad rámec udržitelnosti modelů a metodik. Sekundárně bude výstupem rozšířená RationAI pipeline pro zpracování klinických dat pro trénink a inference AI modelů a tato bude i nadále k dispozici jako open-source řešení.

Provedeme analýzu otevřených datových zdrojů a repozitářů na národní a evropské úrovni se zaměřením jak na technické možnosti strojového přístupu k nim, tak podmínkami přístupu přes externí službu. Taktéž bude provedena analýza a pilotování napojení AI modelů na zdroje citlivých dat a jejich zpracování v adekvátně zabezpečených prostředích (zejm. SensitiveCloud).

ČINNOSTI:

- stanovení hlavních cílů aktivity, požadavků na validaci metadatových modelů a depozičních metodik, a identifikace klíčových zainteresovaných stran;
- analýza vstupních dat: Identifikace a kategorizace citlivých medicínských dat, která budou použita pro validaci, včetně obrazových, časových a tabulkových dat;
- integrace RationAI⁹⁸ pipeline: implementace pipeline RationAI pro zpracování a analýzu klinických dat ve spojení s AI modely;
- nasazení SensitiveCloud⁹⁹: využití infrastruktury SensitiveCloud pro bezpečné ukládání, přenos a zpracování citlivých medicínských dat;

⁹⁸ <https://rationai.fi.muni.cz>.

⁹⁹ <https://www.cerit-sc.cz/infrastructure-services/sensitivecloud>.

- vývoj validačních scénářů: definování testovacích případů a metodiky validace metadatových modelů a depozičních metodik pomocí AI modelů;
- testování a validace: provádění testovacích běhů, ladění AI modelů a iterativní validace metadatových modelů;
- dokumentace a reportování: Příprava výstupních zpráv a dokumentace v souladu s požadavky na doložení výstupu;
- zveřejnění výsledků: publikace výsledků, včetně validovaných metadatových modelů a depozičních metodik v otevřeném repozitáři (SW validace vytvořených metadatových modelů a depozičních metodik pomocí AI modelů zpracovávajících citlivá medicínská data). **AI_1.7**

6/ Analýza možností integrace repozitáře s externími datovými zdroji

Aktuálně není možné jednoduše využívat a vyhledávat data z různých Evropských zdrojů pro potřeby české AI/ML komunity. Je nutná osobní prohledávání různých repozitářů a následné získání dat pomocí metody daných repozitářů. V rámci aktivity provedeme širší analýzu různých Evropských repozitářů, Data Spaces a dalších zdrojů na evropské úrovni zaměřených na možnosti jejich integrace s repozitářem DM4AI. Cílem je zjednodušit výzkumné aktivity české AI/ML komunity, jež vyžadují externí datové zdroje.¹⁰⁰

Druhá analýza bude provedena na možnostech integrace s datovými zdroji na národní úrovni. Analýza se zaměří na možnosti integrace s datovými zdroji poskytující otevřená data státní správy, samospráv a ostatních institucí na národní úrovni (např. ČHMÚ) za účelem zpřístupnění těchto dat pro obohacení aktivit české AI/ML komunity o práci s daty nepocházející z výzkumné činnosti.

Ted' není možné jednoduše získávat data z různých otevřených národních datových zdrojů. Je nutné vyhledávat a procházet tyto datové zdroje jednotlivě a využívat metody pro stahování dat k nim náležící. Bude provedena analýza vybraných otevřených národních datových zdrojů a možnost jejich integrace s repozitáři z Aktivity 2.2.

ČINNOSTI:

- shromáždění a hodnocení existujících datových rozhraní na evropské úrovni;
- identifikace klíčových požadavků a standardů pro integraci;
- analýza dostupných technologií a nástrojů pro integraci datových rozhraní;
- vypracování závěrů a doporučení pro integraci datových rozhraní na evropské úrovni; **AI_1.8**
- identifikace relevantních datových poskytovatelů poskytující otevřená data státní správy, samospráv a ostatních institucí na národní úrovni (např. ČHMÚ¹⁰¹);
- identifikace klíčových požadavků a standardů pro integraci s těmito zdroji;
- analýza existujících nástrojů a technologií pro integraci datových zdrojů;
- vypracování závěrů a doporučení pro efektivní integraci datových zdrojů na národní úrovni. **AI_1.9**

KÓDY VÝSTUPŮ PODAKTIVITY¹⁰²

AI_1	Repozitář DM4AI
AI_1.1	Metadatový model AI/ML

¹⁰⁰ Příklady iniciativ, se kterými skupina předpokládá úzkou spolupráci: BDVA/DAIRO (<https://bdva.eu/dairo/>), CLAIRE (<https://cairne.eu/>), ELLIS (<https://ellis.eu/>), EUDAT (<https://eudat.eu/>), BBMRI (<https://www.bbMRI.cz/>) a BigScience (<https://bigscience.huggingface.co/>).

¹⁰¹ <https://www.chmi.cz/>.

¹⁰² Kódy výstupů podaktivity jsou provázány na přílohu č. 3 Harmonogram klíčových aktivit a kap. 11 Hlavní plánované výstupy/produkty Studie proveditelnosti.

AI_1.2	Správa dat a možnosti vyhledávání
AI_1.3	Integrace SW pro generování a správu provenance s repozitářem pro AI
AI_1.4	Analýza povahy AI modelů z právně-technologického pohledu
AI_1.5	Licenční modely a vzorové smlouvy pro AI modely
AI_1.6	Metadatový model pro popis licenčních aspektů AI modelů
AI_1.7	SW – validace vytvořených metadatových modelů a depozičních metodik pomocí AI modelů zpracovávajících citlivá medicínská data
AI_1.8	Analýza možností integrace datových rozhraní na evropské úrovni
AI_1.9	Analýza možností integrace s datovými zdroji na národní úrovni
AI_1.10	Metodika přípravy, anotaci a kurátorství dat
AI_1.11	Vzdělávací materiály/tutoriály

10.3.2. PODAKTIVITA 4.2 – PROSTŘEDÍ PRO INFERENCI GENERATIVNÍCH AI MODELŮ ULOŽENÝCH V DATOVÉM REPOZITÁŘI

Tato aktivita se zaměří zejména na pokrytí specifických potřeb pracovní skupiny Data management pro umělou inteligenci a strojové učení v oblasti uživatelské interakce s generativními AI modely uloženými v datovém repozitáři. Jelikož stažení a zprovoznění těchto modelů vyžaduje nemalé prostředky, je žádoucí zpřístupnit omezenou interakci s nimi pro přiblížení jejich možností a základní otestování uživateli přímo v repozitáři.

Vznikne prostředí pro jednoduchou interakci s generativními modely a bude propojeno s datovým repozitářem. S využitím specifických metadat o daném modelu se tento nainstaluje na prostředcích vhodných pro jeho spuštění ve správném softwarovém prostředí a zpřístupní uživateli přes grafický uživatelský interface typu „chatbot“.

Repozitářový systém CLARIN-DSpace obohatíme o možnost uživatelské interakce s generativními datovými modely pro AI ve formě chatbotu s omezenými schopnostmi. Tato možnost bude jakýmsi ekvivalentem „náhledu“ na dataset, protože bude stejně jako náhled u obrazových dat poskytovat základní možnost posouzení vhodnosti daného modelu pro uživatele, aniž by si jej musel stáhnout a spustit na vlastních prostředcích.

Aktivita je plánována v těchto krocích.

Nejprve zanalyzujeme existujících SW řešení pro efektivní inferenci generativních modelů, jejich nahraditelnost a praktické testy vybraných řešení a vytvoříme seznam SW řešení inference, které bude nástroj používat pro maximální dosažitelnou kompatibilitu s modely a jejich formáty a ověřené způsoby načtení modelů. Dále provedeme analýzu přenosu datových modelů z repozitáře do inferenčního prostředí, včetně návrhu autentizace a autorizace jak k modelům v repozitáři, tak k inferenčnímu prostředí.

Implementujeme datové přenosy a automatické spuštění modelu na optimálně zvolených hardwarových prostředcích s maximálním využitím metadat v repozitářovém záznamu datasetu. A implementujeme grafické uživatelské rozhraní prostředí pro komunikaci s modely, včetně implementace napojení přímo z repozitáře CLARIN-DSpace.

Výstupem bude SW definující prostředí pro inferenci generativních AI modelů uložených v datovém repozitáři a interakci s nimi s důrazem na efektivitu inference pro různé typy a velikosti modelů. Po žádosti o „náhled“ modelu v datovém repozitáři bude model zkopírován do výpočetního prostředí

pro inferenci, nasazen na vhodný hardware a softwarovou platformu a uživatel bude informován, že může zahájit komunikaci s modelem. K tomu účelu bude k dispozici webové rozhraní.

Navrhované prostředí umožní prvotní seznámení s modely, aniž by uživatel musel tyto provozní otázky řešit.

Mezi hlavní kvalitativní/kvantitativní přínosy aktivity bude zkrácení životního cyklu použití metod umělé inteligence a strojového učení díky optimalizaci datových přenosů a zvýšení uživatelského komfortu použitím sjednoceného řešení pro řízení práce s daty.

Díky vyvíjenému nástroji uživatelé budou moci testovat modely na lokální infrastruktuře, automatizovaně je přenášet blízko výpočetních prostředků a využívat zde rozsáhlé datové sady. Společný nástroj pro řízení dat v oblasti aplikace metod umělé inteligence a strojového učení rovněž usnadní mezioborovou spolupráci. Zvýší se také kvalita dat s ohledem na principy FAIR, díky sjednocenému způsobu přenosu dat a jejich metadat. AI_2

ČINNOSTI:

- analýza existujících SW řešení pro efektivní inferenci generativních modelů, jejich nahraditelnost a praktické testy vybraných řešení;
- analýza přenosu datových modelů z repozitáře do inferenčního prostředí;
- implementace datových přenosů a automatického spuštění modelu;
- implementace grafického uživatelského rozhraní prostředí pro komunikaci s modely;
- provoz nejméně jedné instance repozitářového systému CLARIN-DSpace, která bude mít integrované uživatelské prostředí pro komunikaci s generativními modely a bude funkční a uživatelsky použitelný celý řetězec od vyžádání spuštění kompatibilního modelu z repozitáře až po uživatelskou komunikaci s modelem. AI_2

10.4. KA 5 – TEMATICKÉHO CLUSTERU SOCIÁLNÍ VĚDY

ZAPOJENÍ PARTNEŘI: SOÚ AV ČR (Garant TKA), UK, MU, ZČU

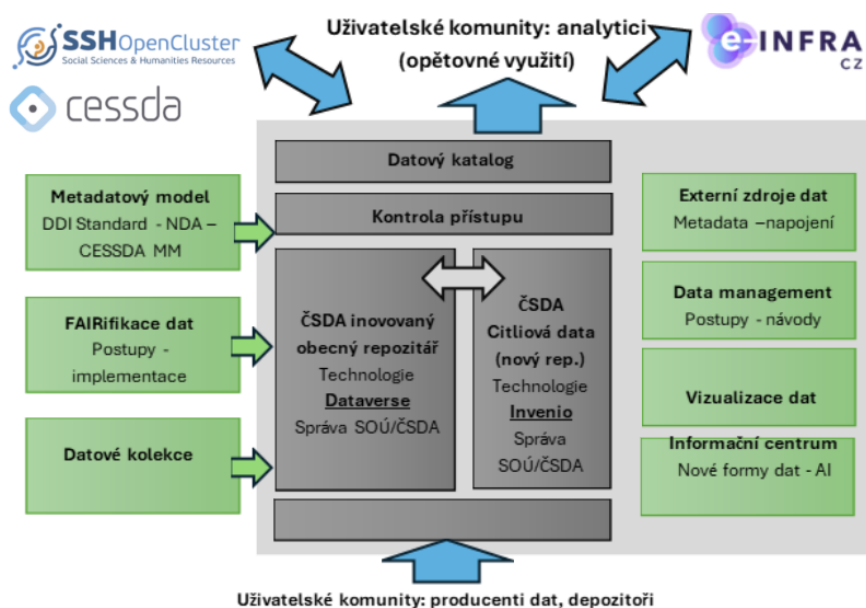
Vytvoříme oborovou repozitářovou platformu pro sociální vědy na základě inovace dvou existujících repozitářů ČSDA¹⁰³ (SOÚ) a DataHub¹⁰⁴ (PřF UK) doplněných o nově vybudovaný repozitář pro citlivá data. Efektivita dosažení cílů se opírá o zapojení existujících systémů do NDI EOSC v ČR a jejich doplnění o nezbytný, dosud chybějící prvek oborové infrastruktury. Komplexní prostředí pro implementaci politiky Open Science v sociálních vědách cílí jak na ukládání dat, tak na jejich opětovné využívání v sociálněvědním a mezioborovém výzkumu. Existující systémy jsou integrované do mezinárodního ekosystému datových služeb a zároveň tak bude realizováno propojení NDI na evropskou úroveň oborové datové infrastruktury.

Cílem inovace repozitářů ČSDA a DataHub je zajištění plné kompatibility s prostředím NDI a NMA a zároveň i mezinárodní infrastrukturou, integrace platformy do NDI, implementace principů FAIR data, vytvoření vazeb na uživatelské komunity producentů i analytiků dat a vytvoření nových služeb NDI, které podpoří sdílení dat s využitím NDI a využívání inovativních postupů ve výzkumu založeném na sdílení dat. Repozitář pro citlivá data v sociálních vědách v ČR chybí. Jeho vybudování je v této vědní oblasti, která často pracuje s osobními údaji nezbytné. Bez jeho existence nelze v českém prostředí

¹⁰³ Český sociálněvědní datový archiv (ČSDA), Sociologický ústav AV ČR, <https://archiv.soc.cas.cz/>.

¹⁰⁴ DataHub, Mapové a datové centrum, Přírodovědecká fakulta UK, <https://datahub.natur.cuni.cz/>.

citlivá sociálněvědní data uložit a zpřístupnit pro sekundární analýzu a zajistit naplňování principů FAIR. Efektivní konstrukce nového repozitáře se opírá o maximální využití prostředí NRP a výsledků TKA SENSI.



Obr 6: Inovovaný repozitář ČSDA a nově vytvořený repozitář citlivých dat

Obr. 6 shrnuje propojení aktivit projektu na repozitáře národního centra datových služeb ČSDA. Inovovaný obecný repozitář ČSDA bude přesunut do prostředí NRP, správu a péči o data zajistí SOÚ AV ČR s využitím technologie Dataverse a bude funkčně propojen s nově vybudovaným repozitářem ČSDA – Citlivá data. Do NDI budou integrována FAIRifikovaná data existujícího repozitáře. V rámci aktivit budou vytvořena východiska pro napojení na NDI a NMA (metadatový model, postupy FAIRifikace) a FAIRifikovány významné kolekce dat. Služby ukládání a poskytování přístupu k primárním datům doplní další služby NDI. Na systém budou napojeny externí zdroje dat (integrace metadat do datové knihovny, anotace, prolinkování dat). Maximální zpřístupnění těchto zdrojů je prioritou výzkumných komunit i ve vztahu k řešení aktuálních společenských výzev v decizní sféře. „Guidelines“ k problematice datového managementu přímo podpoří využívání služeb platformy, standardy sdílení dat a jejich kvalitu a prosazování politiky Open Science. Implementace nástrojů vizualizace dat rozšíří uživatelské komunity, zvýší informační hodnotu dat (vytváření časových řad) a přispějí k využívání NDI. Informační centrum podpoří prosazování inovativních metodologických postupů (nové formy dat, AI).

Systém zajistí napojení na ESFRI¹⁰⁵ výzkumnou infrastrukturu CESSDA ERIC¹⁰⁶ a tím i na ekosystém evropských datových služeb¹⁰⁷ Social Sciences & Humanities Open Cluster (SSHOC) a evropskou úroveň EOSC. Propojení se opírá o využití mezinárodních standardů a nástrojů, zejm. mezinárodní metadatový standard DDI, CESSDA Metadata Model (CMM), technologii Dataverse a principy Otevřeného archivačního a informačního systému (OAIS)¹⁰⁸ atd. Standardní řešení také pomůže v procesu získání mezinárodní certifikace Core Trust Seal (CTS). Přidanou hodnotou je zvýšení kvality dat a výzkumu

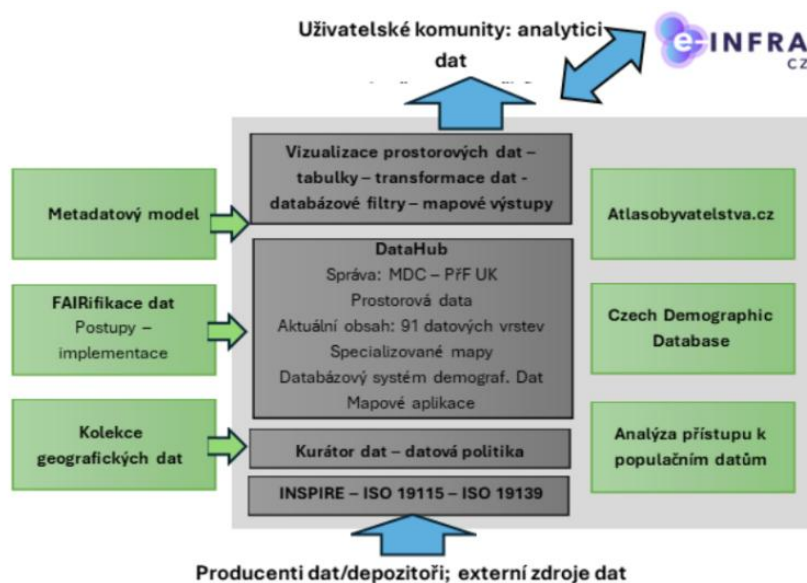
¹⁰⁵ Evropské strategické fórum pro výzkumné infrastruktury (ESFRI), <https://www.esfri.eu/>. Myšleny jsou infrastruktury zařazené na strategický plán Evropské komise pro evropské výzkumné infrastruktury tzv. ESFRI Roadmap.

¹⁰⁶ Consortium of European Social Science Data Archives European Research Infrastructure Consortium (CESSDA ERIC), <https://www.cessa.eu/>.

¹⁰⁷ Projekt má vazbu též na další ESFRI infrastruktury, a to ESS ERIC (European Social Survey, <https://www.europeansocialsurvey.org/>) a GGP (Generations & Gender Programme, <https://www.ggp-i.org/>).

¹⁰⁸ Open Archival Information System Model (OAIS).

a jeho mezinárodní konkurenceschopnosti i s ohledem na význam mezinárodního komparativního výzkumu pro vědeckou excelenci v oboru.



Obr. 7: Inovovaný repozitář DataHub

Obr. 7 popisuje provázání aktivit na inovovaný repozitář DataHub, který poskytuje data zejména v oblasti sociální geografie a demografie ve specifickém prostředí pro prezentaci prostorových dat. Pro připojení na NDI a NMA systém projde výraznou aktualizací. V rámci aktivit budou vytvořena východiska včetně specifického metadatového modelu a postupů FAIRifikace a FAIRifikována kolekce geografických dat. Zaměříme se také na systemizaci a zpřístupnění demografických databází a populačních dat prostřednictvím systému a nově vytvořených platform (Atlasobyvatelstva.cz, Czech Demographic Database), čímž odpoví na dlouhodobé požadavky výzkumné komunity.

Výsledky a navrhovaná dílčí řešení projektu TKA SOC budou konzultována s relevantními oborovými partnery, dále budou prezentována na odborných konferencích v České republice i v zahraničí. Pro posílení prezentace výsledků, získání zpětné vazby a konzultací řešení bude realizována řada zahraničních pracovních cest, zejména členy týmu Sociologického ústavu AV ČR a Západočeské univerzity v Plzni. Předpokládá se účast na odborných konferencích, workshopech, seminářích, příp. letních nebo zimních školách z oborů týkající se zaměření projektu, tj. sociální vědy, práce s odbornými daty a informacemi (např. management dat, FAIR data), problematika zapojení do EOSC, posílení spolupráce atd.

Zaměříme se také na systematické šíření výstupů projektu a rozvoj vzdělávacích kapacit v oblasti datového managementu nejen pro odbornou komunitu v sociálních vědách. Jejím cílem je prostřednictvím cílené komunikační strategie a nově vytvořené webové platformy zvýšit povědomí o nových službách, analýzách a vzdělávacích materiálech vzniklých v projektu.

10.4.1. PODAKTIVITA 5.1 – INOVACE REPOZITÁŘE ČSDA A FAIRIFIKACE DAT¹⁰⁹

Český sociálněvědní datový archiv (ČSDA) v současnosti zpřístupňuje data převážně ze sociologických výzkumů, ale i výzkumů některých dalších oborů (nyní 742 datových souborů) včetně bohaté

¹⁰⁹ V popisu KA jsou plánované aktivity propojeny na realizované výstupy. Výstupy jsou označeny kódem, pro TKA SOC v podobě S_x (klíčový), S_x.x (dílčí, tzn. náležící ke klíčovému).

dokumentace (metadata) za účelem sekundární analýzy v sociálněvědních výzkumech. V rámci projektu dojde k rozšíření kapacity datového repozitáře, a tím bude možné ukládat a zpřístupňovat FAIRifikovaná data i pro další obory, jako jsou výzkumy z experimentální ekonomie nebo nové typy dat, např. syntetická data, která umožňují inovativní přístupy k analýze dat. Realizace aktivity podpoří jak rozšíření komunit využívajících repozitář, tak vyšší využitelnost pro komunity stávající včetně podpory pro využívání inovativních, vysoce konkurenceschopných metodologických postupů analýzy dat. Zároveň bude vybudován repozitář pro zpřístupnění citlivých dat, která nelze uložit v nezabezpečeném prostředí a volně zpřístupnit.

Pro obor experimentální ekonomie mají výsledky tohoto projektu zásadní dopad, protože dosud nemá oborový repozitář, kam by bylo možné data ukládat, a ani vypracované metodiky pro práci s daty. Specifická je různorodost dat, která v rámci oboru vznikají, ať způsobem sběru, laboratorní, online, terénní nebo získaná z různých softwarů (např. Tree, oTree, Qualtrics) případně fyziologická data (z eye-trackingu, EEG, fMRI atp.).

S ohledem na technický vývoj bude věnována pozornost využití nástrojů AI a jazykových modelů pro práci s daty a možností ochrany identity respondentů. Pro práci s daty to může být významný proces, aby je bylo možné po skončení výzkumu uložit do repozitáře a patřičně zpřístupnit podle principů FAIR.

Pro inovaci datového úložiště ČSDA bude využita nově vznikající struktura a kapacity NRP a stávající systém ČSDA vybudovaný na platformě Dataverse pro otevřené datové archivy. Systém bude přesunut ze serveru SOÚ¹¹⁰ do prostředí NRP. Budou tak využity pokročilé systémy prostředí NRP (AAI, přidělování licencí, kybernetická bezpečnost, systém replikace). Bude zachován systém Dataverse realizující napojení repozitáře na (a) existující datové služby ČSDA, (b) mezinárodní prostředí datových služeb CESSDA ERIC a jejím prostřednictvím na evropskou úroveň EOSC zahrnující sociálněvědní cluster SSHOC, kde je CESSDA vůdčí organizací. Budou nastaveny veškeré podmínky a postupy za účelem zajištění technické interoperability s NRP a celkové standardizace v rámci NDI jako napojení na NMA a NKR. Dále budou vytvořeny pracovní pozice správce repozitáře a datového kurátora, kteří budou zajišťovat správu, rozvoj a příjem dat do inovovaného repozitáře ČSDA. **s_1**

Pro integraci existujícího repozitáře do NDI připravíme metadatové modely kompatibilní s NMA. Pro účely zpřístupnění a zpracování různých typů dat ze sociálněvědní oblasti bude vypracován obecný metadatový model pro sociální vědy, založený na mezinárodních oborových standardech. Vycházíme z mezinárodního standardu DDI¹¹¹, který je komplexní, generický a kontinuálně rozvíjený a v sociálních vědách běžně používaný. DDI umožní pořizovat bohatá metadata různých typů dat pro účely interoperability a sekundární analýzy ve výzkumu. V rámci projektu budou definovány povinné elementy z hlediska kompatibility s metadatovým modelem NMA a metadatovými modely mezinárodní datové infrastruktury (CESSDA CMM aj.). **s_1.1**

Budou vypracovány metodiky práce se sociologickými daty a dále s daty z experimentů z oblasti ekonomie. Obě metodiky se budou zabývat správou výzkumných dat, jakým způsobem je zpracovat, aby je bylo možné uložit do repozitáře, zpřístupnit s důrazem na principy FAIR a interoperability a jak s těmito daty v sekundární analýze pracovat. Data ze sociologických výzkumů, typicky se jedná o dotazníková šetření – kvantitativní, ale i kvalitativní, se často používají k sekundární analýze, proto je důležité mít data dobře a podrobně popsána. Experimentální data z oboru ekonomie představují specifický datový typ, pro jehož využití je potřeba mít znalosti nad rámec těch, které jsou v oboru

¹¹⁰ Sociologický ústav AV ČR.

¹¹¹ <https://ddialliance.org/>.

rozšířeny. Metodika bude sloužit k rozšíření těchto znalostí, aby ekonomická experimentální data mohla být využívána v sekundárních analýzách a replikacích. Jedná se o pilotáž a ověření začlenění specifického typu dat do datové infrastruktury v oblasti sociálních věd. **S_1.3, S_1.4**

Pro FAIRifikaci dat bude vytvořeno několik FAIR Implementation Profile (FIP), které budou definovat a konkrétně navrhovat, jakým způsobem plnit principy FAIR pro konkrétní vědeckou komunitu v rámci sociálněvědní oblasti. **S_1.3, S_1.4**

Veškeré připravované metodiky a metadatové modely budeme konzultovat na mezinárodní úrovni, s ohledem na interoperabilitu a zajištění propojení dat i služeb do mezinárodního prostředí. Z tohoto důvodu budou pro výsledky projektu přínosné účasti na mezinárodních fórech, konferencích, seminářích, zabývajících se správou dat, FAIR, interoperabilitou atd. Důležité bude i navázání úzké spolupráce s významnými mezinárodními výzkumnými infrastrukturami a dalšími významnými iniciativami (např. EOSC, Research Data Alliance, CESSDA). **S_1.3, S_1.4, S_1.1, S_1.2**

FAIRifikované datové sady budou ukládány v repozitáři do kolekcí jednotlivých oborů nebo významnějších výzkumů atd. Předpokladem je, že nových datasetů bude v rámci projektu uloženo v repozitáři ČSDA velké množství. Z toho důvodu jsou zde uváděny jen významné kolekce (pro data z experimentální ekonomie, sociologie a syntetická data). Budou upravena a FAIRifikována data již zveřejněná v repozitáři ČSDA. Z důvodu interoperability a FAIRifikace budou provedeny nutné práce na překladech slovníků, čištění dat, úpravě metadat. Data budou opatřena metadaty podle patřičného metadatového modelu. K tomu budou realizovány překlady dokumentace (součást metadat), implementace mezinárodního tezauru ELST atp. za účelem integrace do mezinárodních datových služeb = FAIRifikace dat. Proběhne kontrola a zpracování dat pro účely depozice, poskytování a sdílení dat. U již uložených dat dojde k jejich čištění, doplnění či úpravě metadat podle požadovaného standardu. **S_1.7, S_1.6, S_1.8**

Samostatně bude vypracována metodika pro vytváření a práci se syntetickými daty. Syntetická data v sociologii vznikají jako uměle vytvořená data, která obsahují pro výzkumníky stejnou informaci o celku společnosti jako získaná empirická data, ale již neobsahují původní informace od původních respondentů. Můžeme říct, že původní data od respondentů jsou promíchána tak, aby nebylo možné původní respondenty dohledat, ale aby o celku společnosti vypovídala stále stejně dobře. Na základě reálných empirických dat se nasimulují noví „virtuální“ respondenti, kteří nikdy neexistovali, ale chovají se statisticky stejně jako ti v původním datasetu. Ověřuje se, zda syntetická data odpovídají charakteristikám reálných empirických dat, ale zároveň nezpřístupňují informace o konkrétních jednotlivcích. Dále se používají při testování analytických metod, algoritmů nebo nástrojů bez toho, aby poškodila nějaký výzkum nebo respondenty. Často se také využívají ve výuce a školení. Umožňují simulaci hypotetických scénářů, například lze vytvořit model společnosti s určitými vlastnostmi a sledovat, co by se stalo, kdyby se změnila míra nezaměstnanosti nebo volební systém. Jedná se o nový typ dat, proto dosud neexistují postupy, jak s těmito daty pracovat z pohledu správy dat, jejich FAIRifikace, zpřístupnění v datových repozitářích. Tato data je náročná (finančně i časově) vytvořit, proto je jejich sdílení přínosem pro vědeckou a výzkumnou práci. Informace nutné k vypracování metodiky budou získány také v rámci konzultací s odborníky, na odborných fórech a zpětnou vazbou při prezentování dílčích výsledků. **S_1.5**

S rozvojem umělé inteligence se zvyšuje riziko odhalení identity respondentů i v případech, kdy jsou data formálně anonymizována. Moderní AI nástroje totiž dokážou z kombinace zdánlivě již anonymizovaných dat (např. věk, pohlaví, lokalita, povolání) odvodit nebo odhadnout konkrétní osoby. Jedná se o problematiku, která je velmi aktuální a do budoucna bude zásadní a není v současnosti dostatečně řešena. Analýza se bude věnovat využití AI nástrojů pro aplikaci procesů, které mohou

anonymizovat nebo naopak deanonymizovat sociálněvědní data. K realizaci analýzy možnosti odhalování identit respondentů s využitím AI nástrojů a popsání možností ochrany dat proti odhalení identity s důrazem na možnosti jejich znovupoužitelnosti budou také využity prostředky AI. Součástí bude kritická rešerše volně dostupných zdrojů, které mohou být zneužity k identifikaci respondentů. Sledujeme tedy dva cíle: 1) za pomoci AI ověřit, zda je bezpečné nadále sdílet data, která podle současných měřítek považujeme za anonymizovaná a deidentifikovaná, nebo mohou vést k identifikaci respondentů po spojení s jinými veřejně dostupnými daty; 2) najít za asistence AI způsoby, jak data nově spolehlivě anonymizovat a deidentifikovat. **S_1.10**

Většina souborů kvantitativních a kvalitativních dat je vědecky přínosná a znovuvyužitelná i v případech, kdy neobsahují osobní a/nebo citlivé údaje. Proto je nutná analýza vhodných prostředků anonymizace datových souborů obsahujících osobní a citlivá data, která umožní zveřejnění datových souborů, které původně obsahovaly osobní a citlivé údaje. Pomocí služeb jazykových modelů bude navržen postup, jak anonymizovat citlivá kvalitativní data, tak aby byla využitelná pro sekundární analýzy. Dodatečným zdrojem pro získání informací k vytvoření analýzy budou konzultace s odborníky a prezentace dílčích výsledků na odborných fórech, která poskytnou zpětnou vazbu. **S_1.9, S_1.10**

ČINNOSTI:

- napojení repozitáře ČSDA do NRP a jeho rozvoj a provoz; **S_1**
- příprava obecného modelu pro sociální vědy a metadatového modelu pro demografická data; **S_1.1 a S_1.2**
- tvorba několika FAIR Implementation Profile (FIP); **S_1.3, S_1.4**
- vypracování metodiky možností vytváření syntetických dat a práce se syntetickými daty; **S_1.5**
- FAIRifikace kolekcí dat (pro experimentální ekonomii, sociologická data, syntetická data); **S_1.6, S_1.7, S_1.8**
- analýza možností postupu kódování citlivého textu a jeho následného zveřejnění bez citlivých údajů; **S_1.9**
- analýza možností odhalování identit respondentů s využitím AI nástrojů a popsání možností ochrany dat proti odhalení identity s důrazem na možnosti jejich znovupoužitelnosti. **S_1.10**

KÓDY VÝSTUPŮ PODAKTIVITY¹¹²:

S_1	Napojení repozitáře ČSDA do NRP a jeho rozvoj a provoz
S_1.1	Obecný metadatový model pro sociální vědy
S_1.2	Metadatový model pro demografická data
S_1.3	Metodika/standard práce s daty ze sociologických výzkumů
S_1.4	Metodika/standard práce s daty z experimentů z oblasti ekonomie
S_1.5	Metodika možností vytváření syntetických dat a práce se syntetickými daty
S_1.6	Kolekce FAIRifikovaných dat pro experimentální ekonomii
S_1.7	Kolekce FAIRifikovaných dat: sociologická data
S_1.8	Kolekce FAIRifikovaných dat: syntetická data
S_1.9	Analýza možností postupu kódování citlivého textu a jeho následného zveřejnění bez citlivých údajů

¹¹² Kódy výstupů podaktivity jsou provázány na přílohu č. 3 Harmonogram klíčových aktivit a kap. 11 Hlavní plánované výstupy/produkty Studie proveditelnosti.

S_1.10	Analýza možností odhalování identit respondentů s využitím AI nástrojů a popsání možností ochrany dat proti odhalení identity s důrazem na možnosti jejich znovupoužitelnosti
--------	---

10.4.2. PODAKTIVITA 5.2 – TVORBA NOVÉHO REPOZITÁŘE ČSDA PRO CITLIVÁ DATA A FAIRIFIKACE

V ČR v sociálních vědách nyní neexistuje infrastruktura pro ukládání a sdílení citlivých dat, která by vyhověla nárokům vyhověla nárokům v rozvíjejícím se prostředí Open Science, legislativy a etickým standardům. Aktivita ji vytvoří, ověří její funkčnost a zařadí do prostředí oborových datových služeb v ekosystémů EOSC-CZ a evropského CESSDA, SSHOC a EOSC. Výzkumná komunita bude mít možnost ukládat a sdílet tato data standardním způsobem odpovídajícím principům FAIR data a požadavkům

Open Science, a to v širším multioborovém a mezinárodním prostředí. Aktivita přinese kvalitativní posun pro efektivní využívání dat v kvalitativním a ostatním výzkumu založeném na analýze citlivých dat. Zároveň zvýší standardy ochrany subjektů výzkumu (GDPR a etika), zajistí vysoké standardy zabezpečení, včetně kybernetické bezpečnosti. To bude mít dopad i na možnosti mezinárodní spolupráce (dostupnost dat, komparace), a tedy konkurenceschopnost výzkumu. Vznikne prostředí vhodné pro zpřístupňování citlivých dat veřejné sféry, což je prioritní požadavek komunity i vědních politik. Součástí projektu je uložení a zpřístupnění citlivých dat (dosud 0, nově minimálně 10 datasetů).

Bude vytvořen nový repozitář pro citlivá data, tj. taková, která potřebují zvláštní ochranu osobních údajů subjektů výzkumu a jiná data vyžadující zvláštní ochranu. Repozitář bude vytvořen s využitím softwarového systému INVENIO s příslušnými modifikacemi pro citlivá data. e. SOÚ/ČSDA zajistí kurátorskou správu, stanoví podmínky přístupu a nastavení ve spolupráci s NRP. Dále bude zajištěna role správce repozitáře, který se bude úzce podílet na přípravě repozitáře a poté na jeho správě. Budou vypracovány nebo inovovány zásady archivace dat, datová politika a archivační řád ČSDA na základě implementace. Nastavení Acces Control bude provedeno na základě výstupů z TKA Citlivá data a podle oborových standardů. Pro popis ukládaných dat bude využit upravený obecný metadatový model, popsáný výše v textu. S_2, S_1.1

Sociální vědy často vytvářejí znalosti o společnosti na základě zpracovávání individuálních údajů o osobách. Ne vždy je možné je zpracovávat a analyzovat v anonymizované podobě. Archivace a zpřístupňování takových dat buď vyžaduje de-identifikaci subjektů údajů při zachování požadované výpovědní hodnoty dat (metody anonymizace dat a minimalizace rizika odkrytí identity). Pokud to nejde, je třeba nastavit specifické postupy a organizační a technická opatření k zabezpečení osobních údajů při jejich používání ve výzkumu. Bude vytvořena metodika, jak postupovat při přípravě citlivých dat pro repozitář a při jejich zpřístupňování. Při vývoji bude metodika průběžně procházet konzultacemi s odbornou komunitou a bude zpracovávat dílčí výstupy TKA SENSI. S_2.1

Navržená metodika bude ověřena při zpracování modelových datasetů s citlivými daty v nově vybudovaném repozitáři s vysokým stupněm kontroly přístupu a specifickými postupy při rozhodování o jeho přidělení. Pro testování procesů depozice, zpracování a distribuce (zpřístupnění) dat budou nejprve využita data, která má datový archiv již k dispozici, ale má jen velmi omezené možnosti jejich zpřístupnění. Existující data budou upravena a FAIRifikována, zejména bude provedena komplexní kontrola obsahu dat, včetně analýzy rizika odkrytí identity (disclosure risk assessment). Bude připraven FAIR Implementation Profile (FIP), který definuje, jak aplikovat a plnit FAIR principy při depozici, zpracování a distribuci tohoto typu dat v sociálních vědách. Dále budou zkontrolována a připravena metadata podle odpovídajícího metadatového modelu. Nový repozitář bude registrován v NKR. Předpokládáme, že na konci projektu bude zveřejněno 10 datasetů s citlivými daty. S_2.2

ČINNOSTI:

- vybudování repozitáře pro citlivá data a jeho provoz; S_2
- tvorba metodiky pro data management, archivaci a sdílení citlivých dat v sociálních vědách, zpřístupnění citlivých dat v repozitáři; S_2.1
- testování procesů depozice, zpracování a distribuce dat. S_2.2

KÓDY VÝSTUPŮ PODAKTIVITY¹¹³:

S_2	Nový repozitář citlivých dat pro sociální vědy
S_2.1	Metodika pro data management, archivaci a sdílení citlivých dat v sociálních vědách, zpřístupnění citlivých dat v repozitáři
S_2.2	Pilotní kolekce citlivých dat a metadat pro uložení v repozitáři

10.4.3. PODAKTIVITA 5.3 – INOVACE REPOZITÁŘE DATAHUB A FAIRIFIKACE DAT

DataHub Mapového a datového centra slouží jako centrální platforma Přírodovědecké fakulty Univerzity Karlovy pro sdílení, vizualizaci a analýzu geografických, zejména prostorových dat. Nabízí nástroje pro práci s daty, otevřené datasety, interaktivní mapové aplikace, analytické dashboards i prohlížeč mapových děl. Zpřístupňuje vědecké informace z oblastí sociální a fyzické geografie a demografie. Prostorová data jsou poskytována v podobě vrstev a geodatabází, formátů tzv. geografického informačního systému (GIS). V současnosti DataHub obsahuje 91 datových vrstev/geodatabází. Používáno je komerční, avšak mezinárodní geografickou komunitou standardně používané prostředí ArcGIS Online s nadstavbou ArcGIS Hub (umožňující tvorbu geoportálu s metadatovým katalogem). Licence je hrazena Univerzitou Karlovou, využívána mnoha pracovišti a je s ní počítáno v dlouhodobém horizontu. Tvorba vlastního prostředí portálu mimo ArcGIS by vyžadovala řádově větší investice do technického zázemí i kvalifikovaného personálu. Veškerá činnost vedoucí k napojení repozitáře DataHub do NDI/NMA je proveditelná v rámci stávající licence v kombinaci s otevřenými postupy a nevyžaduje náklady navíc. Jelikož se jedná o specifický typ dat, nelze je zpřístupnit v běžném datovém repozitáři (respektive pouze omezeně a bez přidání hodnoty prostorové informace, kterou lze zobrazit v integrované mapové aplikaci). Přesto se jedná o významné datové zdroje, proto jejich FAIRifikace a zpřístupnění metadat v prostředí NDI bude přínosné nejen pro výzkumnou a vědeckou komunitu, ale vzhledem k formě bude atraktivní i pro širokou veřejnost.

Zároveň budou zkoumány možnosti zpřístupnění populačních dat, která jsou důležitou součástí řady výzkumů, a integrování do NDI.

Platforma DataHub bude rozšířena o napojení na systém NDI/NMA. Bude vyřešena automatizace nahrávání dat, získávání identifikátorů (persistentní link, DOI) a správa metadat pomocí vlastního programu, který bude využívat REST API. Ten bude přístupný správci a datovým kurátorům Mapového a datového centra, kteří budou zajišťovat nahrávání a správu dat. Bude realizováno napojení na společnou autentizační a autorizační infrastrukturu NDI. Protokoly systému, standardy a datové politiky budou inovovány na základě implementace. Dále bude naprogramován databázový systém, který umožní rozšíření služeb o specifické datové výstupy z oboru demografie (neprostorová

¹¹³ Kódy výstupů podaktivity jsou provázány na přílohu č. 3 Harmonogram klíčových aktivit a kap. 11 Hlavní plánované výstupy/produkty Studie proveditelnosti.

databázová data). Budou vytvořeny pracovní pozice správce repozitáře a datového kurátora, kteří budou zajišťovat správu, rozvoj a příjem dat do inovovaného repozitáře. **s_3**

Pro integraci existujícího repozitáře do NDI/NMA budou připraveny metadatové modely pro geografická a demografická data zajišťující kompatibilitu s tímto systémem. Dále bude repozitář registrován v NRK. Z důvodu zajištění interoperability bude první model vypracován s přihlédnutím ke specifikům geografických dat a požadavků prostředí GIS. Vycházet se bude z mezinárodních standardů týkajících se geografických dat – ISO 19115, 19139 a INSPIRE. Pro demografická data budou vyhledány mezinárodní standardy a normy, ze kterých bude možné vycházet. **s_1.2, s_3.1**

Geografická data, která jsou nebo budou zveřejněna v inovovaném repozitáři DataHub, budou FAIRifikována. Z důvodu interoperability a FAIRifikace budou nutné práce s překladem slovníků, čištění dat, úprava metadat. U již uložených dat (91 datových vrstev) dojde k jejich čištění a doplnění metadat podle požadovaného standardu. Proběhne kontrola a zpracování dat pro účely depozice, poskytování a sdílení dat. Předpokládá se zpřístupnění více než jednoho datasetu v průběhu projektu. **s_3.2**

Dostupnost detailních populačních dat je nezbytná pro řadu oblastí výzkumu (například demografii, sociologii, epidemiologii, ekonomii či regionální plánování), neboť umožňuje konstrukci ukazatelů a provádění analýz, které z běžně publikovaných agregovaných dat není možné realizovat. Ve srovnání s jinými zeměmi přetrvávají v České republice značné rezervy v přístupu k mikrodatům (tj. údaje o jednotlivcích v populaci jako věk, pohlaví, vzdělání atd. na individuální úrovni, nikoli agregace v rámci skupin) z administrativních zdrojů a v možnostech jejich využití. Analýza mezinárodních standardů a přístupů k poskytování detailních populačních dat a mikrodat bude aplikována na situaci v ČR včetně doporučení zaměřených na zjednodušení přístupu k potřebným datům, jejich integraci do prostředí NDI a rozšíření možností jejich využití v souladu s principy Open Science. Informace nutné k vypracování analýzy budou také získány v rámci konzultací s odborníky, na odborných fórech a zpětnou vazbou při prezentování dílčích výsledků. **s_3.3**

ČINNOSTI:

- napojení repozitáře DataHub do NRP a jeho rozvoj a provoz; **s_3**
- zajištění kompatibility se systémem repozitáře (metadatové modely); **s_3.1, s_1.2**
- FAIRifikace kolekce geografických dat České republiky; **s_3.2**
- analýza mezinárodních standardů a způsobů poskytování populačních dat. **s_3.3**

KÓDY VÝSTUPŮ PODAKTIVITY¹¹⁴:

s_3	Inovovaný repozitář DataHub
s_3.1	Metadatový model pro geografická data
s_3.2	Kolekce FAIRifikovaných dat: podrobná geografická data ČR
s_3.3	Analýza mezinárodních standardů a způsobů poskytování populačních dat

¹¹⁴ Kódy výstupů podaktivity jsou provázány na přílohu č. 3 Harmonogram klíčových aktivit a kap. 11 Hlavní plánované výstupy/produkty Studie proveditelnosti.

10.4.4. PODAKTIVITA 5.4 – VÝVOJ A PILOTNÍ ZAVEDENÍ SLUŽEB A NÁSTROJŮ PRO SOCIÁLNĚVĚDNÍ OBLAST

Aktivita završuje konstrukci oborové repozitářové platformy v ekosystému EOSC v ČR na základě vybudování a integrace nastavbových služeb datové infrastruktury. Na základě výsledků výše uvedených aktivit budou postaveny konkrétní služby ČSDA a DataHub pro výzkumnou komunitu a napojeny na národní i mezinárodní ekosystém EOSC. Budou vytvořeny a implementovány nástroje pro vizualizaci dat (analýza časových řad). Budou vytvořeny zcela nové služby pro DataHub v oblasti demografických dat (Atlasobyvatelstva.cz). Pozornost bude věnována i významným oborovým zdrojům, jako jsou data z mezinárodních velkých výzkumných infrastruktur s českou účastí a možnostem jejich sdílení a využívání. Všechny služby rozšíří možnosti mezioborového využití a zvýší nebo rozšíří jejich využitelnost.

Bude vytvořen vizualizační softwarový webový nástroj sociálněvědních dat z výběrových šetření zobrazující vývoj postojů, názorů a přesvědčení cílové populace v čase. Nástroj bude umožňovat filtrování výsledků dle sociodemografických a dalších podstatných ukazatelů. Je tak možné sledovat dynamiku vývoje jednotlivých skupin v populaci. Nástroj bude zaměřen na jednu ze základních metod sociálních věd. Časová srovnání, rozšíří uživatelskou skupinu datových služeb nad rámec analytiků ovládajících statistické metody (tj. přiblíží datové služby též řadě expertů přímo v decizní sféře). Nástroj bude cíleně vyvíjen jako flexibilní s možností dodávání dalších měřených konceptů a třídících znaků. V pilotním provozu bude nasazen na časové řady z repozitáře ČSDA z dat Centra pro výzkum veřejného mínění (vzhledem k vysokému využití dat CVVM se tak zvýší dosah na cílovou skupinu) s možností rozšíření na další sociálněvědní datové zdroje. s_4

Bude provedena integrace platformy Atlasobyvatelstva.cz pro specializované mapy do datových služeb DataHub. Jedná se o unikátní data využitelná i v jiných oborech, v rámci předchozí aktivity dojde k jejich FAIRifikaci, ale zároveň potřebují k jejich zobrazení a následnou práci s nimi speciální rozhraní, které bude v rámci projektu vyvinuto. Dojde tedy k revitalizaci platformy Atlasobyvatelstva.cz pro specializované mapy. s_5

V projektu rozšíříme stávající DataHub (PřF UK) o novou inovativní databázi zaměřenou na podrobná demografická data České republiky. Databáze nabídne harmonizovaná, vysoce kvalitní data o plodnosti, úmrtnosti, migraci a struktuře obyvatelstva na různých geografických úrovních (ČR, kraje, okresy) a časových osách. Díky volné dostupnosti a podpoře pokročilých analýz se stane klíčovým zdrojem pro demografický a mezioborový výzkum, včetně tvorby podkladů pro státní správu a samosprávu. K integraci budou potřeba extra programátorské práce, aby došlo k efektivnímu zpřístupnění a zobrazení dat. s_6

S ohledem na praxi, kdy sekundární analýza dat je v sociálněvědní oblasti běžnou činností, je důležitá znalost o zdrojích dat. V souladu s Výzvou projektu budou za účelem zvyšování využívání repozitářů provedeny landscape analýzy zaměřené na významné producenty dat a datových zdrojů, budou vybrány datové zdroje, jejichž data nelze nebo není vhodné umístit přímo do repozitářové platformy. Zpřístupněné budou uživatelům následujícím způsobem: a) harvestování dostupných metadat, FAIRifikace a převod do metadatového modelu repozitářové platformy, zařazení do katalogu ČSDA (Dataverse) a zveřejnění, prolinkování na datový zdroj (data bude možné získat z původního zdroje), b) vytvoření popisů, přehledu a Webového rozcestníku zdrojů dat zcela mimo repozitářovou platformu, který bude zveřejněn na webových stránkách ČSDA. Dojde tak k rozšíření služeb repozitáře pro potřeby vědecké komunity. s_7

Bude vypracována analýza, která se zaměří na způsoby sdílení a praktického využívání dat z mezinárodních výzkumných infrastruktur jako ISSP, EVS, SHARE ERIC, ESS ERIC a GGP, s důrazem na to, jak lze data získat pro výzkumné a studijní účely. Tyto mezinárodní infrastruktury provádí sběry dat i na úrovni ČR, jsou tedy velmi důležitým zdrojem pro výzkum. Vypracují se proto podklady pro přehledné návody a podpůrné materiály k tomu, jak získat přístup k mezinárodním datům, jak s nimi pracovat a jak je využít ve výzkumné praxi. Ty budou zveřejněny na webových stránkách ČSDA a stanou se součástí vzdělávacích aktivit. Součástí analýzy bude i posouzení přístupových politik, licenčních režimů a příkladů dobré praxe, které mohou sloužit jako inspirace sdílení dat. Budou hledány cesty, jak navázat užší spolupráci v rámci omezení, která v současnosti existují z hlediska pravidel pro zveřejnění. Problematika a případná řešení budou diskutovány s experty, na odborných fórech. **s_8**

ČINNOSTI:

- tvorba vizualizačního softwarového webového nástroje sociálněvědních dat; **s_4**
- integrace externích zdrojů dat; **s_7**
- integrace platformy Atlastobyvatelstva.cz pro specializované mapy do DataHub; **s_5**
- rozšíření stávající DataHub o demografickou databázi integrovaných do datových služeb; **s_6**
- analýza sdílení a praktického využívání dat z mezinárodních výzkumných infrastruktur. **s_8**

KÓDY VÝSTUPŮ PODAKTIVITY¹¹⁵:

s_4	Nástroj pro vizualizaci dat
s_5	Platforma Atlasobyvatelstva.cz pro specializované mapy
s_6	Demografická databáze integrovaná do datových služeb DataHub
s_7	Integrace externích zdrojů dat
s_8	Analýza sdílení a praktického využívání dat z mezinárodních výzkumných infrastruktur

10.4.5. PODAKTIVITA 5.5 – VZDĚLÁVACÍ PLATFORMA A STRATEGIE KOMUNIKACE, ŠÍŘENÍ VÝSTUPŮ PROJEKTU UVNITŘ ODBORNÉ KOMUNITY V OBORECH SOCIÁLNÍCH VĚD

Následující aktivity se zaměřují na systematické šíření výstupů projektu a rozvoj vzdělávacích kapacit v oblasti datového managementu nejen pro odbornou komunitu v sociálních vědách. Jejím cílem je prostřednictvím cílené komunikační strategie a nově vytvořené webové platformy zvýšit povědomí o nových službách, analýzách a vzdělávacích materiálech vzniklých v projektu. Aktivita podporuje praktické dovednosti a teoretické znalosti v oblasti práce s daty v souladu s principy FAIR a reaguje na potřebu systematické podpory při práci s různými typy sociálněvědních dat. Významně přispívá k rozvoji datové kultury, podpoře mezioborové spolupráce.

V rámci projektu bude dosaženo řady unikátních výstupů využitelných pro sociálněvědní odbornou komunitu. Tato aktivita bude cílit na rozšíření výstupů (nově zřízené služby, vzdělávací materiály, vypracované analýzy) do české sociálněvědní komunity. Cílem navržené aktivity je realizace zacílené komunikační strategie a šíření výstupů projektu uvnitř odborné komunity v oborech sociálních věd (stakeholders, depozitoři, členové uživatelských komunit, výzkumníci, projektoví manažeři atp.). S tím souvisí diseminace informací o nově zřízených službách, vytvořených analýzách a vyprodukovaných vzdělávacích materiálech a rozcestnících a poukazovat na jejich praktickou

¹¹⁵ Kódy výstupů podaktivity jsou provázány na přílohu č. 3 Harmonogram klíčových aktivit a kap. 11 Hlavní plánované výstupy/produkty Studie proveditelnosti.

využitelnost. Realizace této strategie se opírá o dlouhodobé zkušenosti týmu a kontakty v odborné komunitě a bude uskutečněna v návaznosti na rozsáhlé služby poskytované směrem k sociálním vědcům. Zacílená komunikační strategie vyústí v podstatné zvýšení povědomí o datových službách a také přispěje ke zlepšení kultury sdílení a managementu dat v českých sociálních vědách prostřednictvím celé řady diseminačních aktivit, které budou klást důraz na širší dosah v rámci komunit uživatelů a také na interdisciplinaritu ve společenských vědách (např. zvyšování povědomí o FAIR datech a znovuvyužitelnosti dat, informování o vzdělávacích aktivitách zaměřených na uživatele v hlavních oblastech datového managementu). Za účelem plnění strategie bude zřízena speciální sekce webu ČSDA, jehož prostřednictvím bude částečně probíhat komunikace směrem ke stakeholderům (texty informující o nových službách, publikace materiálů vzniklých v projektu, promo týkající se aktivit projektu atp.). Obsah a forma strategie bude cílit také na stakeholders v sociálních vědách (adekvátně zvolený jazyk a odborná terminologie, využití existujících komunikačních kanálů). Navržená komunikační strategie bude realizovaná v návaznosti na další projektové aktivity, bude odrážet zájmy a potřeby hlavních cílových skupin a v průběhu projektu bude průběžně vyhodnocován její přínos. **s_9**

Dále bude vytvořen oborový portál pro podporu vzdělávání v oblasti DMP (Data Management Plan) a data managementu. Tato vzdělávací webová platforma bude obsahovat vzdělávací materiály týkající se managementu různých typů sociálněvědních dat. Na této platformě bude vytvořen obecný tutorial o datovém managementu sociálněvědních dat a několik sad vzdělávacích materiálů pro jednotlivé typy dat (sociologická data a citlivá data, syntetická data, demografická data, populační data, data z experimentální ekonomie), které vzniknou na základě výstupů z dalších aktivit tohoto projektu. A dále budou postaveny na odborné rešerši již existujících materiálů pro management sociálněvědních dat a na znalostech a zkušenostech členů týmu projektu, kteří vyučují sociálněvědní datový management na vysokých školách a specializovaných workshopech. Součástí platformy bude také rozcestník relevantních odborných zdrojů pro management dat v sociálních vědách.

V materiálech zveřejněných prostřednictvím platformy bude kladen důraz na hlavní oblasti datového managementu, tedy vyhledávání dat určených pro sekundární analýzu, plánování práce s daty ve výzkumném projektu (DMP), dokumentace a zpracování dat, etické a legální aspekty produkce dat a práce s daty, ochrana dat, krátkodobé uložení dat a jejich dlouhodobá archivace a kurátorství a veřejné sdílení dat. Znalosti datového managementu nejsou primárně technické (ovládání softwarů pro data management), ale praktické a teoretické. Je nezbytné, aby producenti dat (výzkumníci, projektoví vedoucí) a uživatelé sekundárních dat disponovali znalostí datového managementu, což jim umožní ve spolupráci s datovými repozitáři a dalšími stakeholders produkovat a využívat data splňující principy FAIR.

Webová platforma obsahující výše popsané vzdělávací materiály bude primárně určena a) výzkumníci a výzkumníkům, kteří produkují a/nebo vědecky využívají data pro vlastní výzkum, a b) studujícím v oborech sociálních věd. Dále bude využitelná pro vyučující datového managementu a metodologie sociálněvědního výzkumu jako podklad k výuce. **s_10.1, s_10.2**

V návaznosti na vytvoření analýz, metodik a dalších materiálů k problematice nových forem dat a využití AI ve výzkumu v rámci dalších aktivit této části projektu, budou na webu systematicky a v ucelené podobě zpřístupněny sady informačních materiálů, zdrojů a postupů, které budou zařazeny do struktury služeb ČSDA, případně i DataHub.

Toto informační centrum ve formě vzdělávací webové platformy podpoří rozvoj využívání nových vysoce aktuálních postupů v sociálněvědním výzkumu, otevře významné otázky v oblasti etiky, technologií a právního prostředí a poskytne informace o ověřených řešeních. Zároveň se jedná

o východisko pro širší zapojení do mezinárodní spolupráce výzkumných infrastruktur v této oblasti. Žádná systematicky vytvářená infrastruktura v této oblasti dosud neexistuje. **s_10**

Ve spojitosti s vytvořenými vzdělávacími materiály počítáme s uspořádáním několika školení. Ta budou zejména určena pro poskytovatele dat, studenty, vědecké pracovníky, vyučující a další uživatele datových služeb sociálních věd.¹¹⁶

KÓDY VÝSTUPŮ PODAKTIVITY¹¹⁷

s_9	Strategie komunikace a šíření výstupů projektu uvnitř odborné komunity v oborech sociálních věd
s_10	Informační centrum k problematice nových forem dat a využití AI ve výzkumu do datových služeb
s_10.1	Materiály pro vzdělávání datového managementu a metodologie sociálněvědního výzkumu
s_10.2	Vytvoření guidelines materiálů pro vzdělávání datového managementu a metodologie sociálněvědního výzkumu

10.5. KA 6 – TEMATICKÉHO CLUSTERU FYZIKÁLNÍ VĚDY

ZAPOJENÍ PATRNEŘI: FZU (Garant TKA), UK, UPOL

1Současná fyzikální věda se opírá o stále větší a rozmanitější objemy dat – od masivních datových toků částicové fyziky na LHC (již nyní přes 1,5 petabajtů dat denně) a v projektech DUNE či Pierre Auger až po terabajtové série měření v materiálových a mikroskopových laboratořích. Evropská iniciativa EOSC, česká Národní datová infrastruktura (NDI) a připravovaná legislativa o otevřeném přístupu vyžadují, aby tato data byla **FAIR** – tedy dohledatelná, přístupná, interoperabilní a opětovně využitelná. Bez vybudování nástrojů a standardů by hrozila ztráta vědeckého i ekonomického potenciálu těchto unikátních souborů a Česko by nebylo schopno se plně zapojit do mezinárodních infrastruktur a projektů s dlouhodobým horizontem.

KA 6 proto **buduje specializovaný oborový repozitář „Fyzika“ založený na implementaci systému Invenio** v Národní repozitářové platformě (NRP). Repozitář nabídne robustní úložiště, jasně definované metadatové modely a automatizované nástroje pro hromadný přenos a verifikaci velkých datových balíčků z experimentů ATLAS ITk, DUNE, CTAO SST-1M, Auger i z krystalografické strukturní analýzy. Tím zajistí, že data vznikající dnes budou již od začátku ukládána v souladu s FAIR zásadami a budou dlouhodobě dohledatelná a použitelná pro další generace vědců.

Na technickou část navazují **inovační služby**:

- nástroj pro **automatizovanou tvorbu metadat** propojený s Elektronickými laboratorními deníky (ELN);
- webová komponenta pro **přímou vizualizaci multidimenzionálních dat** (HDF5/NeXus, FITS aj.) v prohlížeči;

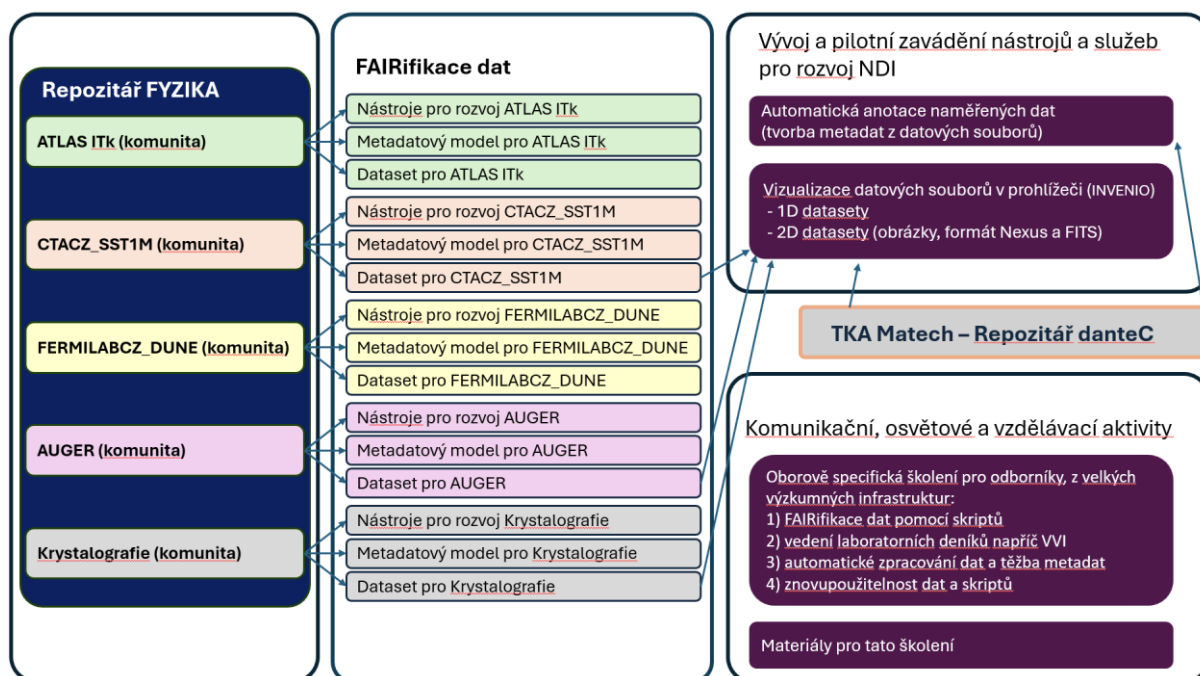
¹¹⁶ Školení budou realizována ve spolupráci se Školicím centrem EOSC CZ, které zajistí odpovídající kapacity, technické zázemí a organizační podporu a nejsou tak zahrnuta do rozpočtu OS II.

¹¹⁷ Kódy výstupů podaktivity jsou provázány na přílohu č. 3 Harmonogram klíčových aktivit a kap. 11 Hlavní plánované výstupy/produkty Studie proveditelnosti.

- a **e-learningové moduly a materiály pro workshopy**, které rychle rozšíří dovednosti FAIR správy dat mezi českými i mezinárodními týmy. Tyto open-source nástroje budou okamžitě využitelné také v dalších clustrech NDI a posílí interoperabilitu celé české infrastruktury.

Klíčové experimenty právě v této době vstupují do fáze upgradu nebo ostrého provozu (ATLAS Run 4 v roce 2030, první data z DUNE v roce 2031). Pokud architekturu ukládání, řízení přístupových práv a FAIR procesy nastavíme už během přípravy, ušetříme nemalé finanční prostředky (v řádu milionů Kč) za pozdější konverze dat, omezíme riziko jejich ztráty a získáme konkurenční výhodu při zapojení do projektů ESFRI a Horizon Europe. Vytvořením jednotného fyzikálního repozitáře navíc poskytneme českým výzkumníkům silnou vyjednávací pozici při sdílení dat a výpočetních zdrojů s mezinárodními partnery a podpoříme transfer znalostí do průmyslu, medicíny a energetiky.

Tímto clustrem tak plníme strategický cíl Výzvy: pokrýt celý životní cyklus fyzikálních dat – od jejich vzniku přes standardizované ukládání a otevírání světu až po cílené vzdělávání komunity. Výsledkem bude trvale udržitelný ekosystém, který promění objemná fyzikální data v reálně znovupoužitelný zdroj inovací pro vědu i společnost.



Obr. 8: Grafické zobrazení TKA Fyzika

10.5.1. PODAKTIVITA 6.1 – TVORBA NOVÉHO REPOZITÁŘE FYZIKA A FAIRIFIKACE DAT¹¹⁸

V návaznosti na projekt NRP vytvoříme nový repozitář pro oborový cluster Fyzika. Repozitář bude využívat službu úložiště poskytovanou a spravovanou projektem NRP, bude se jednat o standardní implementaci poskytovanou NRP na platformě Invenio. Podaktivita zajistí funkčnost, bezpečnost a použitelnost repozitáře pro uživatele, kteří budou data vkládat, a také pro uživatele, kteří budou k datům přistupovat. Připravena bude specifická dokumentace pro tento repozitář s konkrétními příklady postupů pro vkládání dat a jejich popis pomocí metadat. F_1

¹¹⁸ V popisu KA jsou dále plánované aktivity propojeny na realizované výstupy. Výstupy jsou označeny kódem, pro TKA FYZIKA v podobě F_x (klíčový), F_x.x (dílčí, tzn. náležící ke klíčovému).

Vnitřní členění repozitáře Fyzika bude organizováno s ohledem na komunity a datové sady nových velkých fyzikálních experimentů:

- **ATLAS ITk**: Nově budovaný vnitřní detektor (inner tracker, ITk) experimentu ATLAS (A Toroidal LHC Apparatus) na Velkém hadronovém urychlovači LHC (Large Hadron Collider) Evropské organizace pro jaderný výzkum (CERN);
- **FERMILAB CZ_DUNE**: experiment DUNE (Deep Underground Neutrino Experiment) ve FermiLab v USA;
- **CTACZ_SST1M**: funkční prototyp teleskopu pro Cherenkov Telescope Array Observatory (CTAO);
- **AUGER**: observatoř Pierra Augera (PAO – Pierre Auger Observatory);
- **Krystalografie**: pokročilá krystalografická strukturní analýza (ASTRA, tj. Advanced Structure Analysis).

Tomu odpovídají jednotlivé dílčí logické celky, které jsou detailně popsány níže.

S důrazem na FAIRifikaci budou v rámci činností jmenovaných dílčích celků vytvářeny odpovídající metadatové modely, nástroje pro automatické vkládání dat, nástroje pro verifikaci uložených dat a metodiky pro užití dat v dalším výzkumu.

Očekáváme, že TKA FYZIKA tak vytvoří vzory pro rozvoj Open Science i v širší fyzikální komunitě, a to zejména ve vazbě na projekty velkých výzkumných infrastruktur Cestovní mapy ČR, skrze které MŠMT podporuje otevřený přístup k unikátním zařízením.

Na mezinárodní úrovni bude tímto usnadněno zapojení do Cestovní mapy Evropského strategického fóra pro výzkumné infrastruktury (ESFRI). Konkrétní příklady zamýšlené diseminace tvoří program COMPASS Upgrade a Badatelské centrum PALS (Prague Asterix Laser System), na kterých FZU spolupracuje s Ústavem fyziky plazmatu AV ČR, ale i Extreme Light Infrastructure (ELI).

Přímo do TKA FYZIKA jsou zapojeny přes své mateřské instituce FZU a MFF UK také výzkumné infrastruktury MGML (Materials Growth & Measurement Laboratory) a CzechNanoLab, které blízce spolupracují s TKA MATECH.

Úspěch TKA FYZIKA záleží i na adopci založených komunit a vyvinutých nástrojů na mezinárodní úrovni, tedy ve velkých kolaboracích experimentů. Proto budou výsledky prezentovány a nabídnuty uživatelům formou aktivní vystoupení, která jsou plánována např. na následujících konferencích:

- Accelerating the Adoption of Open Science (CERN, Ženeva);
- Open Science Fair: Fusing Forces – Accelerating Open Science through Collaboration (CERN, Ženeva).

Tyto konference jsou oborově zaměřené, přesto je jejich dosah značný, i vzhledem k velikosti kolaborací.

Kromě výše uvedených oborově zaměřených konferencí zajistíme i přesah do širší komunity ve fyzikálních vědách, a to např. na konferencích TRIPLE (Transforming Research through Innovative Practices for Linked Interdisciplinary Exploration) „Improving Discovery and Collaboration in Open Science“ nebo obdobných.

1/ ATLAS ITk

Vytvoříme novou komunitu v repozitáři Fyzika se specifickými metadaty pro data ze sensorů pro ATLAS ITk. Do repozitáře Fyzika převedeme již zaznamenaná data dosud uchovávaná na lokálním diskovém poli. Data budou převedena do vhodného formátu a popsána metadaty. Přístup k datům bude

po nezbytnou dobu vyhrazen pouze pro členy kolaborace ATLAS, jejichž priorita je odůvodněna podílem na vybudování či provozování experimentu ATLAS. Řešení přístupu bude pomocí nástrojů AAI vyvinutých projektem NRP. Virtuální organizace ATLAS, která registruje oprávněné pracovníky, již existuje, využívá se pro gridové počítání. Ve spolupráci s NRP vyvineme a implementujeme řešení pro využití existence této Virtuální organizace pro řízení přístupu v národním repozitáři. **F_1.1**

V laboratoři ve FZU se proměňují senzory pro upgrade vnitřního detektoru ATLAS, který je nyní v plné přípravě tak, aby byl uveden do provozu s počátkem Run 4 plánovaným na rok 2030. Získaná data o celkové velikosti několika desítek TB je nutné uchovávat dlouhou dobu (po dobu nabírání dat pomocí těchto senzorů a jejich analýzy, tedy až po roce 2040) tak, aby mohla být zpětně kontrolována kvalita jednotlivých senzorů. Data obsahují číselné výsledky měření i fotografie senzorů. Kolekci FAIRifikovaných dat vložíme do repozitáře Fyzika pro potřeby oborové komunity. Data budou opatřena specifickými metadaty, vytvoříme pro ně metadatový model a vhodné webové rozhraní či skripty pro přístup k nim. Pro jejich nahrání a verifikaci vytvoříme automatizované nástroje. **F_1.2, F_1.1**

Využijeme nástroje pro hromadný přenos dat z lokálního diskového serveru do repozitáře. Protože se jedná o velké množství souborů s celkovou velikostí několik desítek TB, budou muset být tyto nástroje dostatečně robustní, aby vyprodukovaly konzistentní záznamy v repozitáři i při náhlých výpadcích sítě, problému při čtení z lokálního úložiště či nestabilitě některých komponent repozitáře v důsledku přetížení nebo z jiných důvodů (update verze Invenia, údržba serverů, upgrade operačního systému apod.). Pokud to bude možné, budou nástroje využívat již existující softwarové knihovny vyvinuté projektem NRP nebo jinými skupinami tak, aby údržba tohoto software nevyžadovala velké úsilí při změnách API repozitáře nebo jiných změnách.

2/ FERMILAB CZ_DUNE

DUNE (Deep Underground Neutrino Experiment) bude proměřovat vlastnosti neutrin produkovaných na urychlovači v laboratoři Fermilab. V rámci výzkumné infrastruktury Fermilab-CZ na tomto projektu participuje i česká skupina FZU AV ČR, UI AV ČR a MFF UK. Obří detektor se bude skládat ze 4 modulů, každý s obsahem 17 000 tun tekutého argonu. K vyčítání signálu se bude používat 300 000 křemíkových fotonásobičů (SiPM – Silicon based Photo Multipliers). Tyto detektory fotonů se aktuálně proměřují v několika laboratořích, aby je bylo možné nainstalovat pro plánované zahájení nabírání dat v roce 2031. Výsledky měření jsou nyní uchovávány na diskových serverech u měřicí aparatury. Vlastnosti tak velkého vzorku SiPM jsou velmi zajímavé a využitelné i pro další projekty. Výsledky měření budou popsány metadaty, uloženy do nové komunity repozitáře a zpřístupněny na základě politiky otevřených dat projektu DUNE. Vzhledem k plánovanému nabírání dat nejméně do roku 2042 je nutné data spolehlivě uchovat i pro studium dlouhodobé spolehlivosti SiPM.

Vytvoříme novou komunitu v repozitáři Fyzika se specifickými metadaty pro data z měření křemíkových fotonásobičů pro experiment DUNE. Do repozitáře převede již zaznamenaná data dosud uchovávaná na lokálním úložišti u laboratorního setupu. Bude vytvořen dataset z dat měření křemíkových fotonásobičů, data budou nahrávána průběžně do konce projektu 12/2028. Data budou převedena do vhodného formátu a popsána metadaty v souladu s FAIR principy. Přístup k datům bude po nezbytnou dobu vyhrazen pouze pro členy kolaborace DUNE. Řešení přístupu bude pomocí nástrojů AAI vyvinutých projektem NRP. Virtuální organizace DUNE (zahrnuje pracovníky z více než 35 zemí), která registruje oprávněné pracovníky, již existuje a využívá se pro gridové počítání. Ve spolupráci s NRP vyvineme a implementujeme řešení, jak existenci Virtuální organizace DUNE využít pro řízení přístupu v národním repozitáři. **F_1.4, F_1.3**

Pomocí nástrojů/služeb poskytovaných projektem NRP zkontrolujeme konzistenci datasetů a opatříme je požadovanými přístupovými právy. Po zpracování již existujících dat budeme průběžně doplňovat nové záznamy z nových měření.

Budeme potřebovat nástroje pro hromadné převedení dat do vhodného formátu a velikosti a pro přenos z lokálního úložiště do repozitáře. Protože se jedná o velké množství souborů, budou muset být tyto nástroje dostatečně robustní, aby vyprodukovaly konzistentní záznamy v repozitáři i při náhlých výpadcích sítě, problému při čtení z lokálního úložiště či nestabilitě některých komponent repozitáře v důsledku přetížení nebo z jiných důvodů (updaty verze Invenia, údržba serverů, upgrade operačního systému apod.). Kde to bude možné, využijeme již existující software vyvinutý projektem NRP nebo jinými skupinami tak, aby údržba těchto nástrojů nevyžadovala velké úsilí při změnách API repozitáře nebo jiných změnách.

3/ CTACZ_SST1M

ERIC CTAO (Cherenkov Telescope Array Observatory) buduje na dvou místech (ve Španělsku na Kanárských ostrovech a v poušti Atacama v Chile) rozsáhlé sítě teleskopů pro pozorování gama záření přicházející z kosmu. Česká skupina spolu s dalšími partnery připravila dva prototypy tzv. malých teleskopů SST-1M (SST – Small-Sized Telescopes, průměr zrcadlové plochy 4 m) a oba provozuje na pozemcích Astronomického ústavu v Ondřejově. Nabíraná data již přesáhla 100 TB. Pro tato data vypracujeme metadatový model, vytvoříme novou komunitu v obecném repozitáři Fyzika a data uložíme. S pomocí expertů na AAI nastavíme přístupová práva pro sdílení nejdříve jen pro členy kolaborace s plánem data plně otevřít po uplynutí embargo. Vytvořený metadatový model budeme konzultovat a sdílet s celou kolaborací CTA (která připravuje další typy teleskopů) a prozkoumáme možnosti ukládání vybraných dat pro celou kolaboraci. **F_1.5, F_1.6**

Vytvoříme novou komunitu v repozitáři Fyzika se specifickými metadaty pro data z měření teleskopů SST1M. Do repozitáře převede a FAIRifikuje již zaznamenaná data dosud uchovávaná na lokálním úložišti a na zálohách v objektovém úložišti poskytovaném službou Datových úložišť CESNET. Data budou převedena do vhodného formátu a popsána metadaty. Přístup k datům bude po nezbytnou dobu vyhrazen pouze pro členy kolaborace SST-1M. Řešení přístupu bude pomocí nástrojů AAI vyvinutých projektem NRP. **F_1.6**

4/ AUGER

Observatoř Pierra Augera nabírá data z křemíkových fotonásobičů detekujících gama záření od roku 2004 s plánem nabírat ještě po roce 2030. Postupně tato data uvolňuje pro veřejný přístup ve vhodném formátu. Observatoř zatím nemá stálý repozitář pro primární naměřená data ani pro rozsáhlejší kolekci dat z Monte Carlo simulací. Naše skupina zajistí FAIRifikaci a uložení těchto cenných dat, na jejichž nabírání se podílí velká mezinárodní kolaborace 500 pracovníků z 18 zemí, do oborového repozitáře vytvořením nové komunity s patřičným popisem pomocí metadat na základě specifického metadatového modelu pro data a simulace (v souladu s PAO specifikací). Naměřená data budou postupně přenášena a FAIRifikována z dosavadního úložiště ve výpočetním centru CC IN2P3 v Lyonu založeného na systému iRODS. Vzhledem k počtu souborů je nutné použít nástroje pro spolehlivý přenos dat a kontrolu integrity dat v cílovém repozitáři. Naměřená data budou doplněna o simulace metodou Monte Carlo. Protože objem nasimulovaných souborů je velmi velký (přes 1 PB) a jsou uloženy distribuovaně v různých střediscích zapojených do gridu EGI (European Grid Infrastructure), bude nutné vybrat data vhodná pro dlouhodobé uložení a ta data, na která se odkazují publikované výsledky. Kolaborace Pierra Augera naměřená data postupně uvolňuje pro veřejnost. Ve spolupráci s AAI specialisty bude řízen přístup k různým datovým sadám na základě politiky otevřených dat projektu Auger (licenční podmínky se řídí standardy CC BY-SA 4.0). **F_1.8, F_1.7**

Využijeme nástroje pro hromadný přenos dat z gridového úložiště do repozitáře. Aktuálně jsou data uložena v systému iRODS (Integrated Rule-Oriented Data System) v CC IN2P3 v Lyonu, ale plánuje se přesun do gridového střediska v Bologni. Nástroje tedy budou muset podporovat různá vstupní úložiště.

5/ Krystalografie

Oddělení strukturní analýzy ASTRA (Advanced Structure Analysis) je součástí CzechNanoLab a v rámci Open Access poskytuje strukturní analýzu krystalických a polykrystalických materiálů uživatelům z celého světa (počet určovaných struktur se blíží 1000 ročně). Určovány jsou struktury jak anorganických, tak organických vzorků, včetně biologicky významných struktur a léčiv, a to jak metodami rentgenovské difrakce, tak i difrakcí elektronového svazku. Oddělení je také původcem a správcem analytického softwaru JANA, který je celosvětovým standardem pro analýzu a zpřesňování struktur.

Strukturní analýza je příkladem oboru, který již tradičně vede databáze získaných struktur, má standardizovaný formát dat CIF (Crystallographic Information File) a data jsou před publikací validována. Existuje velké množství struktur v několika uznávaných databázích, včetně Open Access Crystallography Open Database (COD). Obvykle však nejsou k dispozici původní data vzhledem k jejich velikosti (ASTRA ukládá původní data na diskových polích s kapacitou 100 TB). Údaje o měřených vzorcích jsou spravovány prostřednictvím on-line systému ASTRA¹¹⁹, který ale nesplňuje současné standardy FAIR a neposkytuje přístup k datům prostřednictvím persistentních datových identifikátorů.

Krystalografie reprezentuje oblast fyzikálního výzkumu, která se přístupem liší od předchozích dílčích podaktivit. Na rozdíl od velkých kolaborací čítajících stovky až tisíce výzkumníků spolupracujících na velkých experimentech je krystalografie využívána velkým množstvím menších výzkumných skupin. Navíc typicky představuje jeden z mnoha přístupů používaných ke studiu materiálových vzorků a bez mikroskopické charakterizace, spektroskopických a dedikovaných experimentů by měla jen omezený smysl. Reprezentuje tedy i tu část fyzikálního výzkumu, která souvisí s materiálovým výzkumem. Její hlavní výzvou je provázání dat z různých experimentů i materiálových simulací prostřednictvím tzv. linked data přístupu, viz TKA MATECH.

Připravíme v oborovém repozitáři Fyzika vhodné podmínky pro mezioborovou komunitu krystalografie a metadatový model. Hlavní činnosti budou zaměřeny na FAIRifikaci (v maximální možné míře, kterou dovolí stávající struktura metadat) datových sad originálních měření strukturní analýzy. Ve vyvineme vhodné nástroje pro hromadný přenos dat a připravíme metadatový model pro jejich popis. **F_1.10, F_1.9**

Vyvineme nástroje pro hromadný přenos dat a nástroje pro selekci a vizualizaci krystalografických dat.

KÓDY VÝSTUPŮ PODAKTIVITY¹²⁰

F_1	Vytvoření a provoz repozitáře pro obor Fyzika
F_1.1	Metadatový model pro ATLAS_ITk
F_1.2	Nová komunita ATLAS_ITk
F_1.3	Metadatový model pro FERMILABCBZ_DUNE

¹¹⁹ <https://astra.fzu.cz/sampleman/public/login>.

¹²⁰ Kódy výstupů podaktivity jsou provázány na přílohu č. 3 Harmonogram klíčových aktivit a kap. 11 Hlavní plánované výstupy/produkty Studie proveditelnosti.

F_1.4	Nová komunita FERMILAB CZ_DUNE
F_1.5	Metadatový model pro CTACZ_SST1M
F_1.6	Nová komunita CTACZ_SST1M
F_1.7	Metadatový model pro AUGER
F_1.8	Nová komunita pro observatoř AUGER
F_1.9	Metadatový model pro Krystalografie
F_1.10	Nová komunita pro krystalografická data FZU – Krystalografie

10.5.2. PODAKTIVITA 6.2 – VÝVOJ A PILOTNÍ ZAVEDENÍ SLUŽEB/ NÁSTROJŮ PRO ROZVOJ NDI

Repozitářové nástroje tematického clustru Fyzika reagují na potřeby výzkumných týmů pracujících s experimentálními daty v oblasti fyzikálních věd, zejména v laboratořích, kde vznikají datové sady komplexní struktury a různých formátů. Výchozím bodem je situace, kdy vědecká data jsou sice elektronicky ukládána, ale často bez potřebného kontextu, systematických metadat či vazby na experimentální workflow. To snižuje jejich znovuvyužitelnost a dlouhodobou interpretovatelnost.

Prostřednictvím dvou dílčích podaktivit budou vyvinuty nástroje, které tento problém systematicky řeší: 1) umožní automatické generování metadat během měření a jejich propojení s Elektronickým laboratorním deníkem (ELN); 2) zajistí přímou vizualizaci dat uložených v repozitářích v prostředí běžného webového prohlížeče.

Vzniklé komponenty podpoří interoperabilitu v rámci Národní datové infrastruktury (NDI), budou přizpůsobeny fyzikálním datovým formátům jako HDF5, NeXus či FITS a budou vyvíjeny jako open-source, aby mohly sloužit i dalším výzkumným komunitám.

Nástroj pro automatizovanou tvorbu metadat F_2

Cílem dílčí podaktivity je vývoj a pilotní zprovoznění nástroje pro automatizovanou tvorbu metadat z měření fyzikálních veličin a jejich propojení s Elektronickým laboratorním deníkem (ELN). Jedná se o oborově specifický nástroj pro fyziku experimentálního materiálového výzkumu, která snižuje administrativní zátěž vědců, zvyšuje kvalitu metadat a podporuje interoperabilitu v rámci NDI.

Nástroj bude zajišťovat:

- automatické přihlášení vědců do ELN (napojení na AAI);
- načítání informací o vzorcích prostřednictvím QR kódů;
- extrakci metadat přímo z výstupních datových souborů měřicích přístrojů (např. HDF5, NeXus, CSV, ASCII).

Systém bude napojen na open-source systém Python Research Infrastructure User System (PRIUS) a vyvíjen modulárně tak, aby mohl být přizpůsoben různým typům experimentů. Nástroj bude implementován do velké výzkumné infrastruktury MGML na MFF UK.

ČINNOSTI:

- návrh architektury systému a vazby na ELN, AAI a repozitářová API; F_2.1

- vývoj parserů metadat a jejich integrace s ELN; F_2.2
- napojení systému na QR identifikaci vzorků; F_2.3
- vytvoření uživatelské a instalační dokumentace; F_2.4
- pilotní nasazení nástroje na pracovištích MFF UK;
- finalizace a zveřejnění nástroje v open-source repozitáři.

Navrhne celkovou architekturu nástroje pro automatickou tvorbu metadat z měření. Architektura bude zahrnovat komponenty pro integraci s Elektronickým laboratorním deníkem (ELN), napojení na AAI infrastrukturu (e-INFRA CZ) pro autentizaci uživatelů a rozhraní umožňující export metadat do repozitářů v rámci NDI. Návrh bude projednán s vývojáři systému PRIUS a vývojáři platformy INVENIO v rámci NRP. F_2.1

Vyvineme sadu parserů umožňujících extrakci metadat z běžně používaných datových formátů v experimentech (např. CSV, HDF5, NeXus). Tyto moduly budou integrovány do ELN prostředí tak, aby automaticky předvyplňovaly relevantní pole z datových souborů. Bude kladen důraz na modularitu, škálovatelnost a možnost adaptace na různé experimentální protokoly. F_2.2

Zprovozníme systém pro identifikaci vzorků prostřednictvím QR kódů. Každý vzorek bude označen unikátním QR štítkem, jehož načtením dojde k automatickému přiřazení vzorku k měření. Tato funkce bude navržena s důrazem na uživatelskou přívětivost a návaznost na evidenci vzorků v rámci institucionálních databází. F_2.3

Připravíme dokumentaci pokrývající instalaci, konfiguraci a běžné použití systému. Dokumentace bude obsahovat technické požadavky, scénáře nasazení v různých prostředích a návody pro administrátory i koncové uživatele. Součástí budou i ukázky datových vstupů a výstupů s ohledem na standardy FAIR. F_2.4

Provedeme nasazení a provozní testování systému v reálném prostředí dvou laboratorních pracovišť infrastruktury MGML v rámci MFF UK a FZÚ. Budou ověřeny všechny klíčové funkce, včetně správnosti generovaných metadat, propojení s AAI a propojení s evidencí vzorků. Shromáždíme zpětnou vazbu od uživatelů, která bude sloužit k doladění nástroje před jeho zveřejněním.

Po pilotním ověření nástroj finalizujeme a zveřejníme jako open-source balíček. Bude připraven ke snadnému nasazení i v dalších infrastrukturách, které využívají ELN nebo obdobné systémy. Kód bude zveřejněn s otevřenou licencí.

Nástroj pro vizuální zobrazení datových souborů F_3

Cílem dílčí podaktivity je vývoj nástroje pro vizuální zobrazení datových souborů měření fyzikálních veličin přímo v internetovém prohlížeči. Výstup umožní uživatelům NDI efektivní a uživatelsky přívětivé nahlížení na výzkumná data bez nutnosti instalace specializovaného software. Dílčí podaktivita odpovídá potřebám oborů pracujících s multidimenzionálními daty a podporuje opětovnou použitelnost dat v souladu s FAIR principy. Navíc nástroj zpřístupní prohlížení dat široké veřejnosti, která nemá specializované nástroje na jejich prohlížení.

Vyvineme nástroj postavený na moderních webových technologiích (Vue), který bude umožňovat zobrazování datových souborů ve formátech HDF5/NeXus (1D a 2D), FITS (2D obrazová data) a relací ve formátu Metadata4Ing. Nástroj bude podporovat interaktivní zoom, výběr datových řezů, zobrazení základních metadat a bude integrován s repozitářovými API INVENIO. Kód bude distribuován jako open-source komponenta vhodná pro integraci do repozitářů NRP i oborových instancí.

Navrhneme strukturu a vzhled webového rozhraní s důrazem na srozumitelnost pro uživatele z oblasti fyzikálních věd. Rozhraní bude umožňovat zobrazení různých typů dat a bude optimalizováno pro běžné prohlížeče bez nutnosti instalace dalšího software. Architektura bude oddělovat prezentační vrstvu od zpracování dat, což usnadní údržbu a rozšiřování funkcí. **F_3.1**

Vyvineme komponenty umožňující zobrazení fyzikálních dat uložených ve formátech HDF5 a NeXus. Uživatel bude moci procházet strukturu souboru, vybírat 1D datové řady (např. spektra) nebo 2D obrazová data (např. mapy, difrakční vzory) a upravovat zobrazení (měřítka, barevná schémata, výřezy). Vizualizace bude postavena na knihovnách jako h5web nebo Plotly. **F_3.2**

Zajistíme zobrazování 2D obrazových dat ve formátu FITS, často používaném v astrofyzice a dalších zobrazovacích metodách. Nástroj umožní základní interaktivní práci s obrazem. Důraz bude kladen na jednoduché použití i pro uživatele mimo astronomickou komunitu. **F_3.3**

Vyvineme modul pro vizualizaci relací mezi jednotlivými experimentálními kroky dle metadatového formátu Metadata4Ing. Využijeme vizualizační knihovny (např. D3.js) k zobrazení experimentálních workflow ve formě interaktivních grafů, umožňujících přehlednou orientaci v datech a jejich kontextu. **F_3.4**

Nástroj bude nasazen v prostředí repozitáře Fyzika **F_1** a DANTE^c **M_1**.

Po dokončení vývoje a otestování bude nástroj zveřejněn jako open-source software s otevřenou dokumentací. K dispozici bude zdrojový kód, instalační skripty a ukázková data. Nástroj bude připraven k využití dalšími repozitáři na platformě Invenio.

ČINNOSTI:

- návrh uživatelského rozhraní a technické architektury; **F_3.1**
- implementace podpory pro HDF5 a NeXus formáty; **F_3.2**
- implementace podpory pro FITS; **F_3.3**
- implementace podpory pro Metadata4Ing; **F_3.4**
- nasazení a provozu v rámci MFF UK;
- zveřejnění nástroje jako open-source balíčku s dokumentací.

KÓDY VÝSTUPŮ PODAKTIVITY¹²¹

F_2	Nástroj pro automatizovanou tvorbu metadat
F_2.1	Návrh architektury systému a vazby na ELN, AAI a repozitářová API
F_2.2	Vývoj parserů metadat a jejich integrace s ELN
F_2.3	Napojení systému na QR identifikaci vzorků
F_2.4	Vytvoření uživatelské a instalační dokumentace
F_3	Nástroj pro vizuální zobrazení datových souborů
F_3.1	Návrh uživatelského rozhraní a technické architektury
F_3.2	Implementace podpory pro HDF5 a NeXus formáty

¹²¹ Kódy výstupů podaktivity jsou provázány na přílohu č. 3 Harmonogram klíčových aktivit a kap. 11 Hlavní plánované výstupy/produkty Studie proveditelnosti.

F_3.3	Implementace podpory pro FITS
F_3.4	Implementace podpory pro Metadata4Ing

10.5.3. PODAKTIVITA 6.3 – TVORBA E-LEARNINGOVÝCH KURZŮ A MATERIÁLŮ PRO VZDĚLÁVÁNÍ

Podaktivita reaguje na specifické potřeby výzkumníků a technických pracovníků ve fyzice, zejména těch, kteří působí ve VVI. V oblasti fyzikálních věd je běžné, že infrastruktury (např. synchrotrony, neutronové zdroje, velké detektory) automaticky publikují surová (RAW) data do repozitářů, často bez dostatečného kontextu nebo metodické podpory pro jejich FAIRifikaci a další využití. Tato situace klade zvýšené nároky na uživatele, kteří musí data doplňovat o metadata, strukturovat je a zajišťovat jejich srozumitelnost a dlouhodobou interpretovatelnost.

E-learningové kurzy proto nabídnou přehled postupů, jak automatizovaně zpracovávat data z VVI, včetně tvorby Data Management Planů (DMP), které budou navázány na reálné datové toky a Elektronické laboratorní deníky (ELN). Obsahem kurzů bude i práce se skripty pro FAIRifikaci a opětovné využití dat v dlouhém časovém horizontu, typickém pro fyzikální experimenty. Paralelně budou realizovány školicí akce zaměřené na pokročilé využití ELN napříč infrastrukturami a přímou práci s nástroji EOSC. Výstupy budou připraveny v návaznosti na školicí rámec národního školicího střediska EOSC a zařazeny do vzdělávací nabídky NDI¹²².

Vytvoříme sérii interaktivních e-learningových modulů, které budou dostupné online a pokrývají témata:

- zpracování a správa dat z VVI;
- tvorba a údržba Data Management Planů (DMP) s využitím ELN ve VVI;
- využívání nástrojů EOSC pro správu, FAIRifikaci a sdílení výzkumných dat.

Kurzy budou připraveny s důrazem na srozumitelnost pro cílovou skupinu, doplněny o praktická cvičení, příklady z výzkumné praxe a návody. Výstupem bude plně funkční a veřejně dostupný vzdělávací obsah, který bude možné využívat i po skončení projektu. F_4.1

Vytvoříme materiály pro čtyři tematicky zaměřené workshopy/školení pro výzkumníky a technický personál VVI. Témata jednotlivých workshopů budou:

- FAIRifikace dat pomocí skriptů;
- vedení elektronických laboratorních deníků v prostředí více infrastruktur;
- automatické zpracování dat a těžba metadat;
- znovupoužitelnost dat a skriptů v dlouhém časovém horizontu.

Vzdělávací materiály budou kombinovat teoretický výklad s praktickými úlohami a budou navržena tak, aby účastník získal konkrétní dovednosti využitelné ve své instituci. Vytvořené materiály, které budou zpřístupněny online.¹²³ F_4.2

¹²² Školení budou realizována Školicím centrem EOSC CZ, které zajistí odpovídající kapacity, technické zázemí a organizační podporu.

¹²³ Vzdělávací materiály jsou výstupem projektu OS II. Samotná školení/workshopy budou realizována ve spolupráci se Školicím centrem EOSC CZ, které zajistí odpovídající kapacity, technické zázemí a organizační podporu, a nejsou tak zahrnuta do rozpočtu OS II.

KÓDY VÝSTUPŮ PODAKTIVITY¹²⁴

F_4.1	Série tří e-learningových kurzů FYZIKA
F_4.2	Vzdělávací materiály pro čtyři workshopy FYZIKA

10.6. KA 7 – TEMATICKÉHO CLUSTERU HUMANITNÍ VĚDY a umění

ZAPOJENÍ PATRNEŘI: ÚČL AV ČR (Garant TKA), MU, UK, ZČU

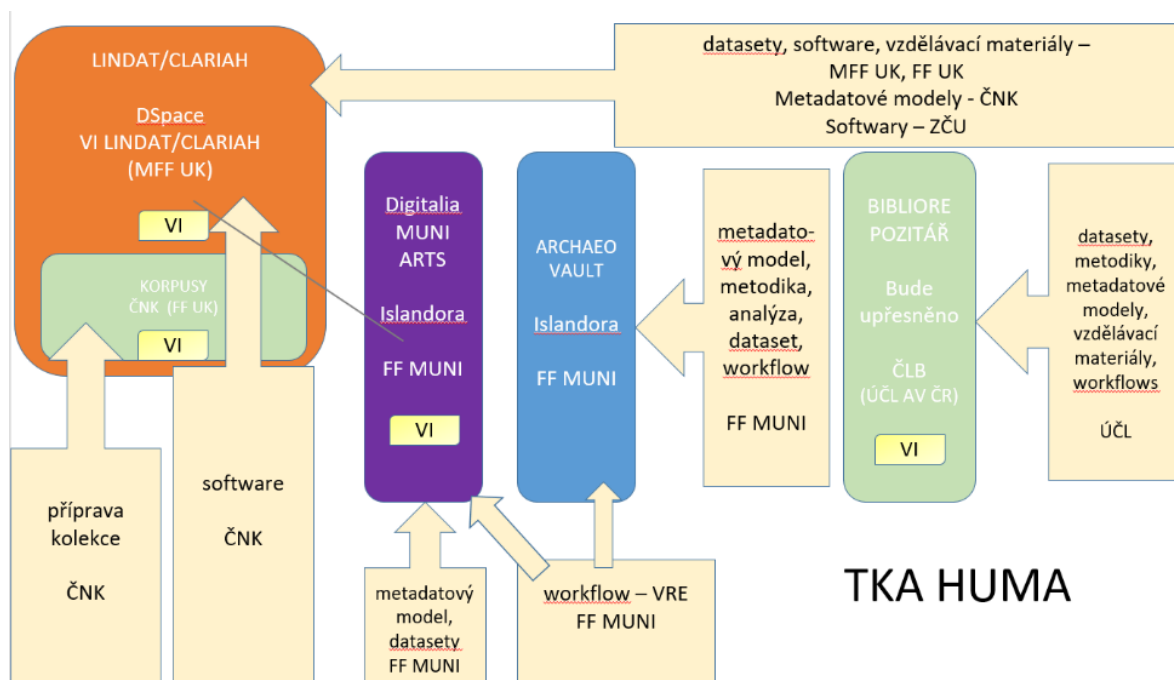
Aktivity TKA HUMA se zaměří na inovaci stávajících a rozvoj nových repozitářů v oborovém clusteru humanitních věd a dále pak na vývoj nástrojů a služeb pro rozvoj NDI. Repozitářem přitom v závislosti na konkrétních technických podmínkách a oborové a institucionální politice při popisu TKA myslíme (a) konkrétní instanci daného repozitářového systému s možností dílčích specificky vymezených jednotek nebo (b) dílčí logicky vymezenou jednotku v rámci takovéto instance ve formě specifické komunity či kolekce se specifickými kurátorskými postupy včetně případných individualizovaných variant metadatových modelů. Zmiňované repozitáře jsou soustředěny okolo čtyř existujících velkých výzkumných infrastruktur pro humanitní vědy jako ústředních producentů výzkumných dat a kompetenčních center v oblasti otevřené vědy v rámci clusteru (Archeologický informační systém ČR – AIS ČR, Česká literární bibliografie – ČLB, Český národní korpus – ČNK a LINDAT/CLARIAH-CZ). Budou při tom využita již existující repozitářová řešení, která budou u vybraných oborů s výraznou produkcí vědeckých dat případně doplněna o nové repozitáře či repozitářové komunity.

Implementace projektu OS II bude tvořena celkem pěti dílčími podaktivitami. Jednotlivé dílčí podaktivity budou provázány na celkem čtyři repozitáře, přičemž každá se zaměřuje na jeden z nich: inovovány budou repozitáře (1) LINDAT/CLARIAH-CZ (systém DSpace; provozovatel MFF UK), který zapojením dat výzkumné infrastruktury ČNK, na jejichž základě vznikne samostatná kolekce/komunita repozitáře, rozšíří svou uživatelskou komunitu o obor korpusové lingvistiky a dále bude repozitář LINDAT/CLARIAH-CZ doplněn o FAIRifikované datasety spolupracujících institucí (MFF UK, FF UK). Inovací pak dále projdou repozitáře FF MU, a to konkrétně (2) institucionální repozitář Digitalia MUNI ARTS a (3) repozitář pro archeologii ArchaeoVault (oba na systému Islandora). Uživatelské komunity obou těchto repozitářů budou díky jejich propojení na NRP rozšířeny o uživatele mimo mateřskou instituci. U repozitáře ArchaeoVault je navíc plánována koordinace aktivit s existující výzkumnou infrastrukturou AIS. Jako nový repozitář pak vznikne (4) repozitář pro bibliografická data provozovaný výzkumnou infrastrukturou Česká literární bibliografie (ÚČL AV ČR), a to v závislosti na výsledcích interní analýzy formou samostatné kolekce/komunity v rámci některého ze stávajících repozitářů, popř. jako samostatná instance některého z repozitářových systémů NRP (ASEP, DSpace, Invenio). Poslední z podaktivit tvoří (5) podaktivita Vývoj a pilotní zavedení nadstavbových nástrojů pro repozitáře TKA HUMA, jejíž náplní bude příprava celkem 4 takovýchto výstupů pro rozšíření služeb a zvýšení uživatelského komfortu jednotlivých repozitářových platforem existujících při TKA HUMA (ČNK, FF MU, ZČU).

Výsledky TKA HUMA budou konzultovány s relevantními oborovými partnery a prezentovány na odborných konferencích v ČR i v zahraničí. Pro posílení prezentace výstupů projektu je proto plánována řada zahraničních pracovních cest, realizovaných členy týmu za ÚČL AV ČR, MFF UK a FF UK, jejichž náplní bude buď prezentace použitých řešení zahraničním partnerům a konzultace s nimi v rámci výměnných pobytů, a/nebo konferenční vystoupení na relevantních odborných akcích. Počítáme

¹²⁴ Kódy výstupů podaktivity jsou provázány na přílohu č. 3 Harmonogram klíčových aktivit a kap. 11 Hlavní plánované výstupy/produkty Studie proveditelnosti.

přítom s prezentací výstupů projektu na prestižních zahraničních konferencích, zejména z oboru (matematické) lingvistiky či digital humanities, ale i dalších humanitních disciplín. Na úrovni celé TKA předpokládáme alespoň 10 aktivních vystoupení formou konferenčního příspěvku či posteru.



Obr. 9: Grafické schéma aktivit

10.6.1. PODAKTIVITA 7.1 – INOVACE REPOZITÁŘE LINDAT/CLARIAH-CZ A FAIRIFIKACE SOUVISEJÍCÍCH DAT¹²⁵

Repozitář LINDAT/CLARIAH-CZ je dlouhodobě provozován stejnojmennou výzkumnou infrastrukturou v systému DSpace. Správu a rozvoj repozitáře kontinuálně zajišťuje Ústav formální a aplikované lingvistiky při MFF UK. Jde o aktuálně největší a nejkomplexnější repozitář existující v clusteru HUMA. V návaznosti na rozšiřování velké výzkumné infrastruktury LINDAT/CLARIAH-CZ se původní lingvistické zaměření repozitáře proměňuje a nově zahrnuje i další vědní obory a instituce, které se staly součástí konsorcia LINDAT/CLARIAH-CZ (celkem 15 partnerských institucí), pro něž nabízí možnost ukládání výzkumných dat, popř. vytváření dílčích oborových či institucionálních kolekcí/komunit. V rámci projektu OS II bude repozitář LINDAT/CLARIAH-CZ inovován vytvořením specifické kolekce/komunity Český národní korpus, již bude kurátorsky zajišťovat stejnojmenná výzkumná infrastruktura, a uživatelská komunita repozitáře LINDAT/CLARIAH-CZ tak bude rozšířena i korpusovou lingvistikou jako clusterově významného producenta oborových dat. Pro potřeby zapojení dat ČNK (a obecněji pro data podobného typu přístupná přes vzdálené služby) bude vyvinuto samostatné metadatové schéma, které zohlední jejich specifické charakteristiky tak, aby byla umožněna jejich snadná integrace do repozitáře. Obsah repozitáře LINDAT/CLARIAH-CZ bude rozšířen o související výstupy z oboru korpusové lingvistiky, a to konkrétně o FAIRifikovaný dataset InterCorp (komunita ČNK). FAIRifikací pak projdou i další datasety jednotlivých institucí konsorcia LINDAT/CLARIAH, zapojených do řešení projektu OS II, a to konkrétně vybrané datasety MFF UK a FF UK. Vedle FAIRifikovaných datasetů pak bude dále pro potřeby snadnějšího zapojení nových i stávajících uživatelských komunit připraven soubor

¹²⁵ V popisu KA jsou plánované aktivity propojeny na realizované výstupy. Výstupy jsou označeny kódem, pro TKA HUMA v podobě H_x (klíčový), H_x.x (díleč, tzn. náležící ke klíčovému).

vzdělávacích materiálů pro práci s repozitářem a soubor softwarů pro zpracování vystavených dat. Implementací projektu OS II tak bude repozitář LINDAT/CLARIAH-CZ obohacen o významnou uživatelskou komunitu korpusových lingvistů a zároveň zásadním zhodnocením projdou, popř. zcela nově při zohlednění FAIR principů budou zařazeny významné oborové datasety (mj. korpus InterCorp, aktualizované verze lingvistických datasetů MFF UK či digitální edice Vělefovy bible) a softwarové nástroje pro jejich zpracování.

Výsledky této aktivity budou prezentovány na relevantních oborových konferencích typu Association for Computational Linguistic Conference (ACL), International Conference on Language Resources and Evaluation (LREC), Conference on Empirical Methods in Natural Languages Processing (EMNLP) atp. Plánovány jsou též návštěvy předních oborových pracovišť rozvíjejících stávající spolupráci, během nichž by byly prezentovány a dále konzultovány výsledky projektu (v předběžném plánu mj. Oxford University, University of Colorado, Brandeis University či Brown University). MFF UK v rámci této aktivity plánuje 14 a FF UK 8 zahraničních cest. V rámci podaktivity budou vykonávány též metodické činnosti související s provozem repozitáře LINDAT/CLARIAH-CZ a implementací Open Science v prostředí humanitních věd a umění.

Inovace repozitáře

V souladu s Výzvou bude repozitář v průběhu řešení projektu (a) integrován s NMA, kam budou též poskytována povinná metadata, (b) bude napojen na společnou autentizační a autorizační infrastrukturu EOSC, (c) bude napojen na technické nástroje přesunu dat, metodiky a standardy, které vytváří IPs EOSC-CZ a projekt zajišťující budování NRP, (d) bude registrován v NKR. Dle možností a oborových zvyklostí bude repozitář zapojován též do mezinárodního výzkumného prostředí. H₂

Vývoj metadatového modelu

V rámci projektu bude vyvinuto metadatové schéma pro data přístupná přes vzdálené služby. To je potřeba zejména s ohledem na specifickou situaci jazykových korpusů, které často z důvodu copyrightu nemohou zpřístupňovat zdrojová data v podobě plných textů či mimo své vlastní servery. H₅₈

Příprava materiálů pro vzdělávání

V rámci projektu bude připravený soubor oborově specifických vzdělávacích materiálů popisujících praktické možnosti vystavení a dokumentace jednotlivých datasetů v repozitáři DSpace tak, aby byly naplněny FAIR principy. H_{1.9}

Vývoj software

V rámci projektu bude vyvinut softwarový balík s konvertory a validátory pro jednotlivé datasety PDT-C-EOSC (The Prague Dependency Treebank – Consolidated) tak, aby byla zajištěna konzistence a koherence vystavených dat. H_{1.2}

FAIRifikace datasetů

V rámci projektu budou FAIRifikovány vybrané klíčové datasety institucí zapojených do VI LINDAT/CLARIAH a VI ČNK. Data jednotlivých datasetů budou vyčištěna, harmonizována, dle potřeby obohacena o perzistentní identifikátory a prověřena z hlediska konzistence a integrity a následně vystavena k využití v repozitáři.

Předpokládáme tyto datasety: Digitální kritická edice Vělefovy bible, Lexikálně-sémantická databáze češtiny, Databáze Manuscripta.cz – Databáze středověkých rukopisů v ČR, PDT-C-EOSC (The Prague Dependency Treebank – Consolidated), Databáze CoCzeFLA – Korpusy dětské češtiny v přirozeném prostředí, Korpus InterCorp. H_{1.3–1.8}

ČINNOSTI:

- inovace repozitáře LINDAT/CLARIAH-CZ s novou kolekcí Český národní korpus;
- vývoj metadatového modelu, zajištění jeho dostupnosti na veřejném úložišti;
- příprava materiálů pro vzdělávání, zpřístupněný na veřejném úložišti;
- vývoj SW balíku s konvertory a validátory pro jednotlivé datasety PDT-C-EOSC FAIRifikace datasetů dostupných v repozitáři LINDAT/CLARIAH;
- metodické a podpůrné činnosti související s provozem repozitáře LINDAT/CLARIAH-CZ a implementací Open Science v prostředí humanitních věd a umění.

PŘEHLED KÓDŮ VÝSTUPŮ PODAKTIVITY¹²⁶

H_1	Repozitář LINDAT/CLARIAH-CZ – zapojení jazykových korpusů infrastruktury ČNK jako samostatné kolekce repozitáře
H_1.1	Vývoj metadatového schématu pro data přístupná přes vzdálené služby
H_1.2	Softwarový balík (LINDAT-KVS) s konvertory a validátory
H_1.3	FAIRifikovaný dataset Digitální kritická edice Vilemova bible
H_1.4	FAIRifikovaný dataset Lexikálně-sémantická databáze češtiny
H_1.5	FAIRifikovaný dataset Databáze Manuscripta.cz – databáze středověkých rukopisů v ČR
H_1.6	FAIRifikovaný dataset PDT-C-EOSC
H_1.7	FAIRifikovaný dataset Databáze CoCzeFLA – korpusy dětské češtiny v přirozeném prostředí
H_1.8	FAIRifikovaný dataset korpusu InterCorp
H_1.9	FAIR data při nasazení a provozu instance repozitáře DSpace v rámci EOSC CZ – vzdělávací materiály

10.6.2. PODAKTIVITA 7.2 – INOVACE REPOZITÁŘE DIGITALIA MUNI ARTS A FAIRIFIKACE SOUVISEJÍCÍCH DAT

Repozitář Digitalia MUNI ARTS funguje jako institucionální repozitář FF MU. V rámci projektu OS II bude zajištěna jeho zásadní inovace, spočívající v jeho provázání na NRP, čímž bude významně rozšířena stávající uživatelská komunita, která nově zahrne i výzkumníky mimo FF MU. Za těmito účely budou vyvinuty postupy a nástroje pro dlouhodobou archivaci stávajících dat. Obsah repozitáře pak bude obohacen o tři FAIRifikované datasety FF MU se zásadní hodnotou pro výzkum v rámci clusteru HUMA. Jmenovitě půjde o Slovník českých filozofů, dataset vzácných fotodokumentů Diatheca a bázi soudobé soukromé korespondence. Mezioborová interoperabilita vystavených dat bude zajištěna díky využití metadatového modelu pro kulturněhistorická data CIDOC-CRM, který bude aplikován pro vybrané pilotní datasety.

V souladu s Výzvou bude repozitář v průběhu řešení projektu (a) integrován s NMA, kam budou též poskytována povinná metadata, (b) bude napojen na společnou autentizační a autorizační infrastrukturu EOSC, (c) bude napojen na technické nástroje přesunu dat, metodiky a standardy, které vytváří IPs EOSC-CZ a projekt zajišťující budování NRP, (d) bude registrován v NKR. Dle možností a oborových zvyklostí bude repozitář zapojován též do mezinárodního výzkumného prostředí. V rámci

¹²⁶ Kódy výstupů podaktivity jsou provázány na přílohu č. 3 Harmonogram klíčových aktivit a kap. 11 Hlavní plánované výstupy/produkty Studie proveditelnosti.

podaktivity budou vykonávány též metodické činnosti související s provozem repozitáře Digitalia MUNI ARTS a implementací Open Science v prostředí humanitních věd a umění. **H_2**

V rámci projektu bude aplikována CIDOC-CRM ontologie na vybrané datasety repozitáře Digitalia MUNI ARTS. Tyto datasety budou následně zpřístupněny v RDF formátu, čímž bude značně posílena jejich interoperabilita. **H_2.1**

V rámci aktivity budou FAIRifikovány tři jiné datasety z provenience FF MU. U těchto datasetů bude zrevidována datová struktura, poté projdou vyčištěním a prověřením konzistence dat. Výsledné soubory budou následně zpřístupněny v repozitáři. Jedná se o Dataset Slovník českých filozofů, Dataset Diatheca a Dataset Soukromá korespondence 20. a 21. stol. **H_2.2, H_2.3, H_2.4.**

ČINNOSTI:

- inovovaný repozitář Digitalia MUNI ARTS s rozšířenou uživatelskou komunitou;
- zpřístupnění datasetů obohacených o využití CIDOC-CRM ontologie;
- zpřístupnění FAIRifikovaných datasetů v repozitáři Digitalia MUNI ARTS;
- metodické a podpůrné činnosti související s provozem repozitáře Digitalia MUNI ARTS a implementací Open Science v prostředí humanitních věd a umění.

PŘEHLED KÓDŮ VÝSTUPŮ PODAKTIVITY:¹²⁷

H_2	Repozitář Digitalia MUNI ARTS
H_2.1	Metadatový model CIDOC-CRM ontologie
H_2.2	FAIRifikovaný dataset Slovník českých filozofů
H_2.3	FAIRifikovaný dataset Diatheca
H_2.4	FAIRifikovaný dataset Soukromá korespondence 20. a 21. stol.

10.6.3. PODAKTIVITA 7.3 – INOVACE REPOZITÁŘE ARCHAEOVAULT A FAIRIFIKACE SOUVISEJÍCÍCH DAT

Repozitář ArchaeoVault je vytvářen jako repozitář určený cíleně pro archeologická data. Vzhledem ke značně interdisciplinární povaze moderního archeologického výzkumu, který úzce kombinuje poznatky humanitních a přírodních věd (např. datování jednotlivých dochovaných artefaktů pomocí pokročilých fyzikálních metod – izotopová měření, dendrochronologické analýzy atp.), se existence samostatného archeologického repozitáře jeví jako velmi funkční a žádoucí. Repozitář ArchaeoVault je připravován v úzké kooperaci s dalšími oborovými pracovišti, zejména Archeologickým ústavem AV ČR Praha a Brno, které společně provozují velkou výzkumnou infrastrukturu Archeologický informační systém (její vlastní repozitář AIS shromažďuje archeologické výzkumné zprávy jako materiál, který tato pracoviště musí shromažďovat ze zákona, a vzhledem k tomuto specifickému účelu zůstává mimo rámec projektu OS II). Příprava repozitáře ArchaeoVault zahrnuje významný podíl metodické práce: postupně tak budou ve spolupráci s odbornou komunitou vyvinuty a připraveny výzkumná zpráva/analýza, identifikující best practices při zpracování a digitalizaci archeologických dat, a dva metadatové modely pro popis archeologických dat. Na jejich základě pak vznikne metodika zpracování archeologických dat pro potřeby repozitáře ArchaeoVault. Praktický proces zpracování archeologických dat pak bude komplexně popsán formou vzorových workflows přístupných uživatelské

¹²⁷ Kódy výstupů podaktivity jsou provázány na přílohu č. 3 Harmonogram klíčových aktivit a kap. 11 Hlavní plánované výstupy/produkty Studie proveditelnosti.

komunitě. V praxi bude celý postup demonstrován na datasetu RES-HUM, který se zároveň stane modelovým příkladem pro tvorbu budoucího obsahu repozitáře ArchaeoVault. Inovace repozitáře ArchaeoVault tak proběhne v úzké interakci s klíčovými oborovými institucemi při zohlednění všech zásadních aspektů otevřené vědy i FAIR principů a ukáže možnosti jejich využití v kurátorské i badatelské praxi. V rámci podaktivity budou vykonávány též metodické činnosti související s provozem repozitáře ArchaeoVault a implementací Open Science v prostředí humanitních věd a umění.

Inovace repozitáře

V souladu s Výzvou bude repozitář v průběhu řešení projektu (a) integrován s NMA, kam budou též poskytována povinná metadata, (b) bude napojen na společnou autentizační a autorizační infrastrukturu EOSC, (c) bude napojen na technické nástroje přesunu dat, metodiky a standardy, které vytváří IPs EOSC-CZ a projekt zajišťující budování NRP, (d) bude registrován v NKR. Dle možností a oborových zvyklostí bude repozitář zapojován též do mezinárodního výzkumného prostředí. **H_3**

Vývoj metadatových modelů

Oba proponované metadatové modely počítají s identifikací jednotlivých datových objektů a jejich následnou analýzou. Na jejím základě bude vyvinuta a interaktivně prověřena standardizovaná metadatová struktura, která bude výsledně zdokumentována a veřejně prezentována. **H_3.1, H_3.2**

Příprava metodiky

V rámci projektu bude vyvinuta metodika zpracování archeologických dat. V první fázi s definicí požadavků na data a metadata a jejich typologie. Po zapracování technických požadavků repozitáře bude metodika uveřejněna. **H_3.3**

Příprava analýzy

Identifikace best practices při zpracování a digitalizaci archeologických dat. V úvodní fázi práce na projektu bude připravena analýza best practices při zpracování archeologických dat. Budou sesbírány a následně analyzovány kurátorské praktiky jednotlivých pracovišť. Na základě jejich komparace pak budou navrženy standardizované postupy s potenciálem univerzálního uplatnění. **H_3.4**

FAIRifikace datasetů

Dataset RES-HUM projde během řešení projektu optimalizací ukládacího formátu, bude vyčištěn a prolinkován na související entity. Následně budou data prověřena z hlediska konzistence a budou odstraněny případné chyby. Výsledný, značně rozsáhlý dataset bude zveřejněn v repozitáři. **H_3.5**

Příprava workflows

Na základě předchozích výstupů budou navrženy workflows pro zpracování dat. Ta počítají mj. s definicí vstupních funkcionalit, vývojem a testováním souvisejících softwarových nástrojů a jejich iterativním testováním a optimalizací. Finalizované workflows budou zdokumentovány a následně zveřejněny k využití. **H_3.6**

ČINNOSTI:

- inovace repozitáře Archaeovault, jeho rozšíření o komunitu archeologů mimo FF MU;
- vývoj metadatových modelů;
- tvorba metodiky pro zpracování archeologických dat;
- analýza best practices pro zpracování a digitalizaci archeologických dat;
- FAIRifikace datasetu RES-HUM a jeho zveřejnění v repozitáři ArchaeoVault;
- zpřístupnění workflows pro archeologická data;

- metodické a podpůrné činnosti související s provozem repozitáře ArchaeoVault a implementací Open Science v prostředí humanitních věd a umění.

PŘEHLED KÓDŮ VÝSTUPŮ PODAKTIVITY¹²⁸

H_3	Repozitář ArchaeoVault
H_3.1	Archeologické metadatové modely
H_3.2	Metadatový model Archeologické ontologie
H_3.3	Metodika zpracování archeologických dat
H_3.4	Analýza: identifikace best practices při zpracování a digitalizaci archeologických dat
H_3.5	FAIRifikovaný dataset RES-HUM
H_3.6	Workflows pro archeologická data

10.6.4. PODAKTIVITA 7.4 – VYTVOŘENÍ REPOZITÁŘE PRO BIBLIOGRAFICKÁ DATA A FAIRIFIKACE SOUVISEJÍCÍCH DAT

Výzkumná infrastruktura Česká literární bibliografie během doby řešení projektu připraví repozitář pro bibliografická data jako významný typ výzkumných dat nejen v rámci clusteru HUMA. Tento repozitář vznikne buď jako samostatná instance některé ze softwarových platform podporovaných NRP, nebo jako samostatná dílčí komunita/kolekce uvnitř některé z nich. Toto rozhodnutí vyplyne z výsledků interní analýzy, která zahrne první rok řešení projektu.

ČLB zároveň jako výstupy projektu připraví dvě metodiky pro FAIRifikaci základních typů bibliografických a souvisejících dat, a to konkrétně: (a) metodiku pro úpravu bibliografických dat existujících v analogové formě (tištěné bibliografie, lístkové kartotéky) a (b) obdobnou metodiku pro zpracování dat biografických slovníků (primárně přehledových částí). Tyto metodiky budou ověřovány na vybraných zdrojových datech, jež budou následně zpřístupněna formou tří pilotních datasetů, na nichž budou testovány možnosti převodu původně analogových dat do strukturované databázové podoby. Jako pilotní příklady byly předběžně vybrány klíčové zdroje dat s významným přesahem i do jiných disciplín v rámci clusteru: Retrospektivní bibliografie české literatury jako největší oborová bibliografie ve formě lístkové kartotéky v ČR, Lexikon české literatury jako klíčové dílo domácí humanitněvědné lexikografie a Bibliografický katalog ČS(S)R. Články jako ústřední národní zdroj bibliografických dat daného typu pro období 2. pol. 20. století. Testovací datasety vzniklé při pracích na projektu pak budou veřejně zpřístupněny a vedle nich vznikne též FAIRifikovaný dataset databáze literárních cen. Na základě obou metodik budou následně připraveny samostatné workflows pro potřeby tzv. SSH Open Marketplace jako ústřední clusterové platformy pro podporu otevřené vědy na evropské úrovni a budou prostřednictvím této platformy k dispozici v angličtině i pro mezinárodní publikum. Za účelem zajištění standardizace v oboru literárněvědné bibliografie pak vzniknou dva metadatové modely, a to metadatový model pro věcný popis literárněvědných dat a model pro popis a klasifikaci literárních děl. Výstupy této podaktivity pak budou prezentovány formou vzdělávacího materiálu (online tutoriály pro práci s nově vytvářeným repozitářem atp.).

¹²⁸ Kódy výstupů podaktivity jsou provázány na přílohu č. 3 Harmonogram klíčových aktivit a kap. 11 Hlavní plánované výstupy/produkty Studie proveditelnosti.

Pro samotný repozitář pro bibliografická data pak budou připravena základní kurátorská nastavení a budou definovány provozní politiky, které budou v souladu s FAIR principy a požadavky Open Science a zároveň zohlední specifické parametry bibliografických dat (verzování, perzistentní identifikátory, propojení bibliografických a autoritních dat). Repozitář bude do budoucna koncipován jako otevřený a bude nabídnut k využití i dalším bibliografickým pracovištím. Díky zařazení datasetů z projektu European Literary Bibliography, který společně vyvíjejí Česká a Polská literární bibliografie a jehož náplní je extrakce, harmonizace a následná společná prezentace bibliografických dat pro jednotlivé národní literatury (momentálně data česká, polská, finská a španělská) bude zajištěn též přesah repozitáře pro mezinárodní publikum.

Pro kvalitní diseminaci výstupů podaktivity pro mezinárodní publikum je plánováno celkem 6 zahraničních cest, jejichž náplní primárně bude účast na prestižních oborových akcích typu výročních konferencí iniciativy EOSC, Association of Digital Humanities Organisations (ADHO) či DARIAH Annual Events a související prezentace výstupů projektu. V rámci podaktivity budou vykonávány též metodické činnosti související s provozem Repozitáře pro bibliografická data a implementací Open Science v prostředí humanitních věd a umění.

Vytvoření repozitáře

V první, analytické fázi přípravy repozitáře pro bibliografická data budou analyzovány požadavky uživatelské komunity pro specifický bibliografický repozitář (metadatové schéma, uživatelská nastavení atp.) a zároveň budou srovnány možnosti existujících softwarových řešení. Na základě těchto prací bude následně zvoleno konkrétní technologické řešení. V druhé, konstrukční fázi bude zvolené řešení připraveno pro ostré nasazení: repozitář bude propojen s NMA a zaregistrován v NKR; bude nastavena odpovídající autentizační a autorizační infrastruktura a dojde k implementaci technických nástrojů, metodik a standardů platformy EOSC CZ. **H_4**

Vývoj metadatových modelů

V rámci projektu vzniknou dosud neexistující metadatová schémata pro potřeby literární vědy, a to konkrétně pro věcný (tematický) popis a klasifikaci forem a žánrů. Paralelně vznikne i metadatový model pro popis literárních děl včetně standardů pro popis vzájemných vazeb a závislostí mezi nimi. **H_4.1, H_4.2**

Příprava metodiky

V rámci projektu budou připraveny metodiky pro převod analogových dat do strukturované podoby. První z nich se bude věnovat datům bibliografickým, druhá biografickým. Obě metodiky budou v průběhu řešení projektu iterativně testovány na pilotních datasetech (srov. FAIRifikace datasetů dále) tak, aby navržené metodické postupy zohledňovaly modelové situace vycházející z reálné praxe. **H_4.3, H_4.4**

Příprava materiálů pro vzdělávání

V rámci projektu budou připraveny vzdělávací materiály pro práci s repozitářem pro bibliografická data, které uživatelům prakticky nastíní možnosti jeho využití. **H_4.10**

FAIRifikace datasetů

V rámci projektu budou připraveny celkem čtyři datasety. První tři budou koncipovány jako pilotní a budou určeny primárně k testování a ověřování funkčnosti výše proponovaných metodik. Vedle nich

bude kompletně vyčištěn, do standardizovaného formátu převeden a o soubor perzistentních identifikátorů obohaten dataset databáze literárních cen, který dnes již zdaleka nevyhovuje standardům moderní informační vědy. H_4.5, H_4.6, H_4.7, H_4.8

Příprava workflows

Na základě obou proponovaných metodik vzniknou související workflows, které budou publikovány v angličtině na platformě SSH Open Marketplace jako ústřední evropské platformě pro otevřenou vědu pro cluster HUMA. Zpřístupněním těchto workflows v angličtině bude zajištěno budoucí využití výstupů projektu i pro zahraniční publikum. H_4.9

ČINNOSTI:

- tvorba repozitáře pro bibliografická data připravený k nasazení do ostrého provozu;
- vývoj metadatových modelů;
- příprava metodik pro převod analogových dat do strukturované podoby;
- tvorba vzdělávacích materiálů pro práci s repozitářem pro bibliografická data;
- FAIRifikace souboru datasetů dostupných v repozitáři pro bibliografická data;
- příslušná workflows zpřístupněná na platformě SSH Open Marketplace;
- metodické a podpůrné činnosti související s provozem Repozitáře pro bibliografická data a implementací Open Science v prostředí humanitních věd a umění.

PŘEHLED KÓDŮ VÝSTUPŮ PODAKTIVITY¹²⁹

H_4	Repozitář pro bibliografická data
H_4.1	Metadatový model pro literární vědu
H_4.2	Metadatový model pro popis literárních děl
H_4.3	Metodika pro převod oborových bibliografií v analogové formě do standardizované a strukturované FAIR podoby za využití nástrojů AI
H_4.4	Metodika pro převod oborových lexikografických prací do strukturované a počítačem čitelné podoby
H_4.5	FAIRifikovaný dataset Retrospektivní bibliografie české literatury v převodu pomocí nástrojů umělé inteligence
H_4.6	FAIRifikovaný dataset Pilotní převod oborových bibliografií ve formě tištěných bibliografií do strukturované podoby za využití nástrojů umělé inteligence
H_4.7	FAIRifikovaný dataset Pilotní převod oborové lexikografické práce do strukturované a počítačem čitelné podoby na příkladu Lexikonu české literatury
H_4.8	FAIRifikovaný dataset Databáze literárních cen
H_4.9	Zpřístupnění bibliografických metodik v podobě workflows na platformě SSH Open Marketplace
H_4.10	Vzdělávací materiály pro práci s repozitářem pro bibliografická data

¹²⁹ Kódy výstupů podaktivity jsou provázány na přílohu č. 3 Harmonogram klíčových aktivit a kap. 11 Hlavní plánované výstupy/produkty Studie proveditelnosti.

10.6.5. PODAKTIVITA 7.5 – VÝVOJ A PILOTNÍ ZAVEDENÍ NADSTAVBOVÝCH NÁSTROJŮ PRO REPOZITÁŘE TKA HUMA

Podaktivita bude zaměřena na technologický vývoj a zavádění nástrojů pro rozvoj výše jmenovaných repozitářů tak, aby bylo docíleno vyššího uživatelského komfortu a zkvalitněna úroveň služeb pro koncové uživatele. Pokud to bude možné, budou jednotlivé výstupy vyvíjeny s univerzálnějším přesahem i pro uživatelské komunity mimo vlastní cluster HUMA. V projektu OS II tak plánujeme vyvinout softwarové nástroje a související workflow pro inkorporaci funkcí virtuální reality jako klíčový výstup „Soubor nadstavbových nástrojů pro rozvoj repozitářů TKA HUMA“.

ČNK v této KA vyvine software pro vizualizaci geolokačních dat. Ten bude uzpůsobený potřebám repozitářového systému DSpace a laděný primárně na jazyková či šířeji humanitněvědná data, bude však komponován s vizí případného využití i mimo tyto oblasti. ZČU v rámci této KA připraví dva softwarové nástroje: software pro strojové zpracování audiodat obsáhne mj. vývoj modulů pro automatické rozpoznávání řeči, diarizaci a případně dalších navazujících nástrojů pro zpracování textu a vedle něj vznikne i softwarový nástroj pro harvesting webových dokumentů. V rámci podaktivity budou vykonávány též metodické činnosti související s vývojem nadstavbových nástrojů pro humanitněvědné repozitáře a implementací Open Science v prostředí humanitních věd a umění.

Workflows pro VRE funkce repozitářů FF MU budou stavěny primárně nad archeologickými kolekcemi. Připravované řešení by mělo zahrnovat mj. online vizualizace, 3D modelování či GIS funkce včetně možnosti uložení výsledků těchto operací přímo do repozitáře.

V rámci projektu budou vyvinuty tři softwarové nástroje pro aplikaci v repozitářích oborového clusteru HUMA. Prvním z nich bude nástroj pro vizualizaci geolokačních dat primárně vyvíjený pro repozitářový systém DSpace. Zároveň budou vyvinuty dva nástroje pro zpracování jazykových dat: prvním z nich bude online nástroj pro rozpoznávání řeči a její následný převod do textové podoby, dostupný přes webové rozhraní, druhým z nich pak bude software pro automatické sklizení (harvesting) textů z webu, zejména pro potřeby lingvistického výzkumu. H_5.1, H_5.2, H_5.3

V rámci projektu budou pro repozitář ArchaeoVault implementovány funkce pokročilé virtuální reality (mj. pro vizualizaci, práci s mapou či 3D objekty). Díky jejich začlenění přímo do zdrojového repozitáře nebudou muset jednotliví uživatelé instalovat samostatné nástroje, čímž bude dosaženo lepšího uživatelského komfortu pro práci s repozitářem. H_5.4

ČINNOSTI:

- vývoj software;
- soubor zdokumentovaných softwarových nástrojů dostupných na veřejném úložišti;
- příprava workflows;
- metodické a podpůrné činnosti související s přípravou nadstavbových nástrojů pro humanitněvědné repozitáře a implementací Open Science v prostředí humanitních věd a umění.

PŘEHLED KÓDŮ VÝSTUPŮ PODAKTIVITY¹³⁰

H_5	Soubor nadstavbových nástrojů pro rozvoj repozitářů TKA HUMA
H_5.1	Software pro vizualizaci geolokačních dat v repozitářích DSpace

¹³⁰ Kódy výstupů podaktivity jsou provázány na přílohu č. 3 Harmonogram klíčových aktivit a kap. 11 Hlavní plánované výstupy/produkty Studie proveditelnosti.

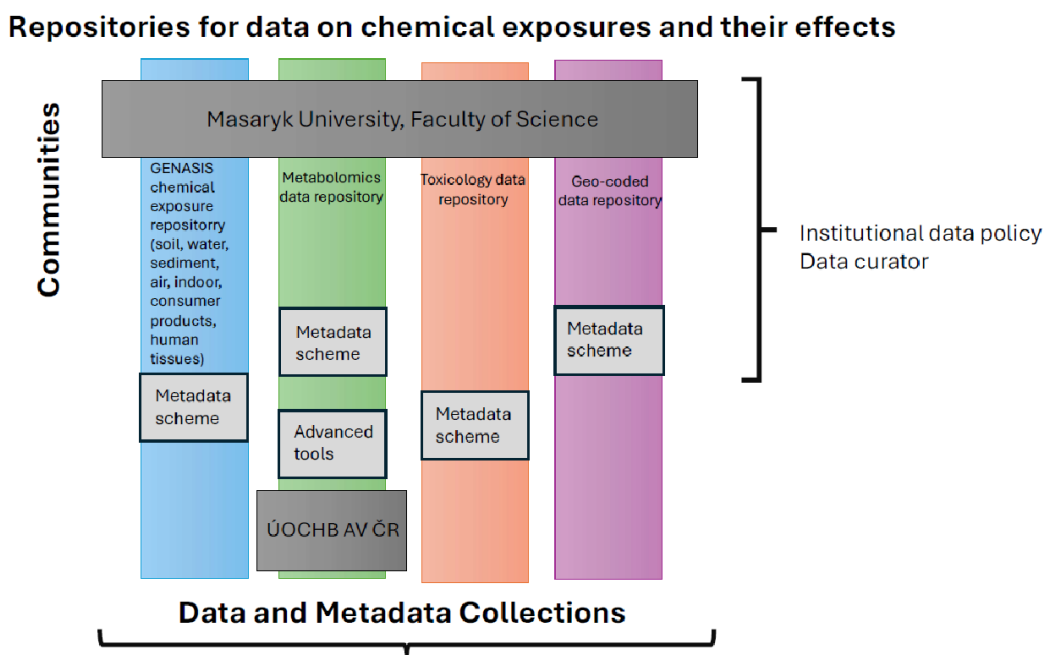
H_5.2	Software pro rozpoznávání řeči přes webové rozhraní
H_5.3	Software pro harvesting textů z webu
H_5.4	Workflow VRE funkce pilotního repozitáře ArchaeoVault

10.7. KA 8 – TEMATICKÉHO CLUSTERU ENVIROMENTÁLNÍ VĚDY

ZAPOJENÍ PATRNEŘI: MU (Garant TKA), ÚOCHB AV ČR, UK

U nových repozitářů s environmentální tematikou bylo identifikováno hned několik potřeb výzkumné komunity.

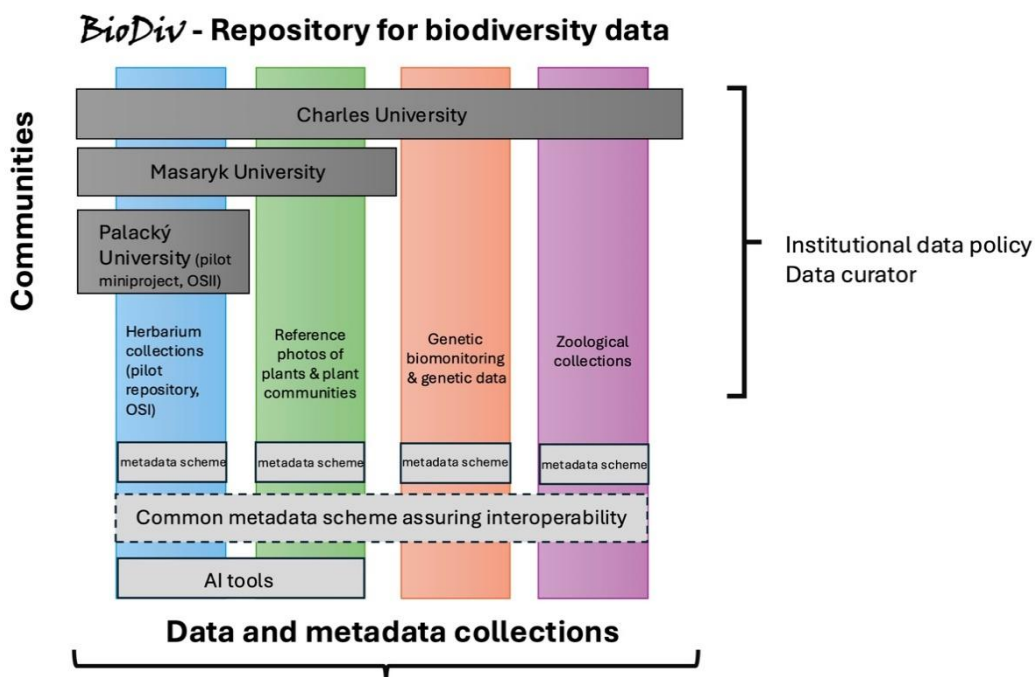
V oblasti chemických rizik se jedná o tři nové repozitáře. Cílené analýzy vybraných chemických látek jsou totiž v posledních letech nahrazovány nebo doplňovány necílenými analýzami, které uchovávají informace o všech (endogenních i exogenních) chemických látkách. V takovém případě je však nutno uchovávat rozsáhlá primární data z hmotnostní spektrometrie s vysokým rozlišením (repozitář pro data z necílené metabolomiky a exposomiky). Abychom mohli analyzovat chemická rizika, je nutné data o chemické expozici doplnit informacemi o chemických hazardech, tedy daty toxikologickými a ekotoxikologickými. A pokud chceme studovat vliv kontaminovaného prostředí na konkrétní populaci, je třeba provázat informace o kontaminaci prostředí na různých místech prostřednictvím geokódování všech informací. Tento přístup nám zároveň umožní propojit environmentální data s daty o zdraví i o dalších faktorech, které mohou na člověka působit (např. ekonomická či sociální deprivace).



Obr. 10: Grafické schéma podaktivit 8.1–8.4

Nové přístupy ke studiu prostředí jsou však ještě komplexnější. Tzv. trojitá environmentální krize, které čelíme, totiž zahrnuje nejen chemizaci prostředí, ale i klimatické změny a ztrátu biodiverzity, přičemž všechny tyto tři aspekty spolu úzce souvisí. Při tvorbě nových repozitářů se proto zaměříme také na nové repozitáře pro zoologické sbírky, genetické vzorky i referenční obrazová data. Vědeckými sbírkami zoologické povahy se rozumějí kolekce lihových, dermoplastických a osteologických preparátů nejrůznějších druhů obratlovců i bezobratlých živočichů potřebných pro interpretaci a reprodukovatelnost vědeckých dat v oblasti taxonomie a diverzity. Repozitáře pro genetický

biomonitoring umožní efektivní sdílení biodiverzitních dat ze sekvenování směsných environmentálních vzorků a genetických dat volně žijících organismů, vyhodnocení výskytu taxonů a jejich porovnání s referenčními daty. Referenční obrazová data jsou zásadní pro základní výzkum v oblasti botaniky, lichenologie, paleoekologie, pro ochranu biodiverzity i pro trénování klasifikačních modelů. Vytvořením příslušných metadatových modelů zajistíme nejen interoperabilitu dat v rámci clusteru repozitářů pro biodiverzitní data, ale i jejich napojení na mezinárodní úroveň (ESFRI infrastruktura DiSSCo.cz nebo agregátor biodiverzitních dat GBIF).



Obr. 11: Grafické schéma podaktivit 8.5–8.7

V klíčové aktivitě se zaměříme na:

- tvorbu metadatových modelů, standardů a metodik umožňujících zpracovat hmotnostně spektroskopická data, fotografické záznamy pro účely studia biodiverzity, genetickou banku volně žijících organismů, zoologické sbírky, toxikologická a ekotoxikologická data a GIS data;
- zavedení metadatových standardů k zajištění sémantické interoperability (provázanost na řízené slovníky, thesaury a ontologie);
- přiřazení citovatelných persistentních identifikátorů k datovým objektům či extraktům;
- přípravu aplikačního rozhraní pro výměnu metadat spravovaných datasetů (DCAT) a rozhraní pro výměnu dat;
- přípravu podpůrných materiálů včetně návodů k tvorbě a validaci FAIRifikovaných datasetů;
- analýzu dalších potřeb výzkumné komunity.

Ve všech případech bude řešena otázka zajištění kvality a interoperability repozitářů a dat v mezinárodním kontextu, otázka licenčních a právních modelů a napojení na externí zdroje dat.

Hlavními přínosy budou:

- standardizované metadatové modely, jež přispějí ke konsolidaci výzkumných domén i uživatelských komunit, které zatím operovaly bez řádně spravovaných repozitářů;
- zvýšení dostupnosti existujících oborově specifických dat (včetně jejich opatření metadaty);

- motivace výzkumné komunity k zapojení do EOSC a mobilizace dat uchovávaných mimo veřejná úložiště;
- lepší spolupráce mezi institucemi a týmy na národní úrovni v dotčených oblastech;
- viditelnost české vědy a jejích výsledků, silnější zapojení do mezinárodních projektů a partnerství;
- posílení pozice ČR jako potenciální hostitelské země EIRENE RI.

10.7.1. PODAKTIVITA 8.1 – INOVACE A ROZŠÍŘENÍ REPOZITÁŘE GENASIS PRO PREZENTACI DAT O CHEMICKÉ KONTAMINACI JEDNOTLIVÝCH MATRIC ŽIVOTNÍHO PROSTŘEDÍ A CHEMICKÉ EXPOZICI POPULACE¹³¹

Environmentální informační systém GENASIS¹³² byl v posledních dvou dekadách postupně vyvíjen v centru RECETOX MU v reakci na potřebu repozitáře, který by společně prezentoval výzkumná data o chemické kontaminaci jednotlivých složek životního prostředí (volné ovzduší, srážky, povrchová voda, sediment, půda atd.). Dostupnost takových dat je ovšem nezbytně nutná nejen pro environmentální výzkum, ale i pro tvorbu mezinárodních, národních a regionálních politik (například implementaci mezinárodních dohod o ochraně prostředí a lidského zdraví před účinky chemických látek). Systém byl proto postupně rozšiřován jak o další matrice (například vnitřní ovzduší nebo prach), tak o nově sledované emergentní látky, a jeho využití bylo nabízeno i dalším výzkumným institucím produkujícím data o kontaminaci prostředí v rámci společných (i mezinárodních) projektů. Metadatový model GENASIS byl následně využit i na mezinárodní úrovni Programem pro životní prostředí OSN (UNEP) pro přípravu a implementaci Globálního monitorovacího plánu Stockholmské úmluvy¹³³. V tomto případě slouží GENASIS jako primární úložiště individuálních datových bodů k následné agregaci a reportingu.

Hlavní přidaná hodnota systému GENASIS tkví v jednotné databázi harmonizovaných dat více poskytovatelů. Při vkládání dat do databáze probíhá jejich kontrola, čištění a případná transformace do harmonizovaného formátu buď na straně poskytovatele dat, nebo na straně datového manažera GENASIS. Sjednocením do harmonizované struktury je zajištěna srovnatelnost dat napříč databázemi bez ohledu na to, z jakého zdroje pocházejí. Tím se rapidně zvýší jejich využitelnost pro další aplikace, aniž by tato práce byla přenášena na koncové konzumenty dat, kteří budou vždy limitováni ve svých možnostech a znalostech, jak takovou harmonizaci provést.

Specifickým rysem GENASIS je to, že není určen pouze vědcům, ale široké škále uživatelů, včetně veřejné správy, aplikačního sektoru a široké veřejnosti. Pro tyto účely byly nad databází GENASIS postupně vystavěny nástroje na vizualizaci, analýzu a interpretaci dat včetně mapových výstupů, statistických modulů nebo analýzy dlouhodobých trendů, i nástroje na agregaci dat a jejich přenos do mezinárodních systémů. Z tohoto důvodu je vhodné celý systém zachovat na dosavadním úložišti, ale investovat do jeho aktualizace tak, aby plně odpovídal současným potřebám odborné komunity i FAIR principům a byla dořešena otázka licenčních a právních režimů.

Modernizace repozitáře GENASIS plánovaná v rámci projektu OS II zahrnuje jeho rozšíření směrem k novým typům matric (např. lidských tkání, jako je krev, moč nebo mateřské mléko) i nově sledovaných látek (ftaláty, bisfenoly, UV filtry, mykotoxiny atd.). Problém chemických rizik řeší v současné době na evropské úrovni HEU Partnerství PARC (Partnership for assessment of risks

¹³¹ V popisu KA jsou plánované aktivity propojeny na realizované výstupy. Výstupy jsou označeny kódem, pro TKA ENVIRO v podobě E_x (klíčový), E_x.x (dílčí, tzn. náležící ke klíčovému).

¹³² www.genasis.cz.

¹³³ www.pops-gmp.org.

from chemicals), kterého se centrum RECETOX aktivně účastní. Při tvorbě ontologií a metadatových modelů inovovaného repozitáře bude proto kladen důraz na jejich plnou kompatibilitu se standardy vytvořenými v PARC. K tomu bude třeba modifikovat datové struktury a popisná metadata a rozšířit metadatový model o atributy vyžadované v základním metadatovém modelu NMA. Pro zajištění flexibility při konverzích do různých výstupních formátů bude datový model zdokumentován ve strojově i lidsky čitelných formátech a vystavěn API pro předávání metadat do NMA. Zásadní bude zavedení sémantické interoperability, tedy napojení metadat na řízené slovníky a ontologie. E_1.1

Pro zajištění citovatelnosti bude dalším zásadním krokem přiřazení perzistentních identifikátorů k datovým objektům či extraktům. Budou využity globálně uznávané perzistentní identifikátory u důležitých entit typu CAS a InChI pro chemické látky, IGSN ID pro vzorky nebo RAiD pro projekty a studie. Problémem je, že databáze procházející kontinuálními updaty a inkrementy dat je těžko citovatelná. Jednotlivé datové záznamy ale můžeme považovat za neměnné (změny jsou možné pouze v případě nalezení chyby), stejně jako importované datasety. Perzistentní identifikátory můžeme přidělovat také exportovaným extraktům z repozitáře. Bude tedy možné pravidelně publikovat ustálené datasety zahrnující průběžné updaty a datové přírůstky (1–4x ročně podle povahy a frekvence aktualizace vstupních dat), Tyto ustálené strojově vytvářené datasety/extrakty budou automaticky publikovány do NMA. Bude ale možné i vytvářet extrakty na základě individuálních žádostí o vydání dat (tematicky vydávané datasety budou vytvářeny podle typicky využívaných kritérií).

V rámci inovace repozitáře GENASIS plánujeme také změnit mechanismus vydávání dat tak, aby veškeré extrakty dat byly registrovány v systému pro management Open Access přístupů do infrastruktury EIRENE/RECETOX. Ve všech případech bude redistribuce dat ošetřena s ohledem na jednotlivé kombinace licenčních podmínek. Licenční režim pak vytvoří jasná pravidla pro práci s daty externích poskytovatelů pro všechny uživatele systému (toto je důležité zejména vzhledem k tomu, že data externích poskytovatelů, včetně zahraničních institucí, mohou být předávána do dalších programů či projektů). E_1

V první fázi budou v souladu s novým metadatovým modelem harmonizována a opatřena metadaty stávající data koordinující instituce, v dalších fázích pak budou připojena prostřednictvím „minizáměrů“ i data dalších institucí. Všechny datové soubory budou doplněny příslušnými sety metadat. Předpokládáme, že inovovaný repozitář a všechny jeho nástroje budou v budoucnu využívány mezinárodně (jako součást datových služeb ESFRI výzkumné infrastruktury EIRENE i jako mechanismus sběru a prezentace dat Globálního monitorovacího plánu), což podpoří dlouhodobou udržitelnost inovovaného repozitáře GENASIS. E_1.2

ČINNOSTI:

- rozšíření repozitáře GENASIS pro prezentaci dat o chemické kontaminaci matric životního prostředí znamená nově naprogramovat strukturu inovovaného repozitáře a jeho napojení na NRP; E_1
- rozšíření metadatového modelu repozitáře dat o chemické expozici GENASIS, který umožní ukládání dat z nových matric i emergentních polutantů v souladu se standardy definovanými v Partnerství PARC; E_1.1
- vytvoření metadatových setů doprovázejících všechna data o chemickém znečištění matric životního prostředí v repozitáři GENASIS; E_1.2
- vytvoření harmonizovaných data setů o chemickém znečištění matric životního prostředí v souladu s nově definovaným metadatovým modelem repozitáře GENASIS; E_1.3
- příprava podpůrných materiálů (standardních operačních postupů a tutoriálů) k tvorbě a validaci FAIRifikovaných datasetů. E_1.4

PŘEHLED KÓDŮ VÝSTUPŮ PODAKTIVITY¹³⁴

E_1	Inovovaný repozitář GENASIS
E_1.1	Inovovaný metadatových model repozitáře dat o chemické expozici GENASIS, který umožní ukládání dat z nových matic i emergentních polutantů
E_1.2	Metadatové sety doprovázející všechna data o chemickém znečištění matic životního prostředí v repozitáři GENASIS
E_1.3	FAIRifikované data sety o chemickém znečištění matic životního prostředí v repozitáři GENASIS
E_1.4	Standardní operační postupy a tutoriály k tvorbě a validaci FAIRifikovaných datasetů

10.7.2. PODAKTIVITA 8.2 – REPOZITÁŘ PRO UKLÁDÁNÍ DAT Z NECÍLENÝCH HMOTNOSTNĚ SPEKTROSKOPICKÝCH ANALÝZ PRO ÚČELY HODNOCENÍ LIDSKÉHO EXPOZOMU

Environmentální analytická chemie se dlouhá desetiletí zaměřovala na cílenou analýzu několika vybraných skupin látek s prokazatelně škodlivými důsledky (výsledky takových cílených analýz jsou pak prezentovány například ve výše uvedeném repozitáři GENASIS). Jak ovšem rostl počet antropogenních látek používaných v průmyslové či zemědělské produkci a rozšiřoval se objem toxikologických informací o nich, rozšiřoval se i seznam chemických látek, které je nutno v prostředí či lidských maticích sledovat. Zároveň se ukázalo nepraktickým zaměřovat se jen na látky, o nichž máme aktuálně dostupné informace, a začaly se rozvíjet metody širokého necíleného screeningu. Tyto metody sbírají informace o přítomnosti celého spektra exogenních i endogenních látek v environmentálních maticích nebo lidských tkáních bez ohledu na to, zda se jedná o látky, které dokážeme v současné době identifikovat. Jedná se tedy o jakési digitální biobankování: generujeme úplný záznam signálů všech přítomných látek, který potřebujeme dlouhodobě uchovat a umožnit přitom jeho opakované prohledávání různými uživateli. Narozdíl od cílených analýz, kdy není třeba uchovávat primární data, ale jen výsledné naměřené koncentrace (viz GENASIS), v tomto případě jde skutečně o primární záznamy, na které budou nadále aplikovány různé metody analýzy, přičemž všechny meziprodukty těchto analýz je třeba zachovat k dalšímu využití.

Výsledný repozitář je vlastně zásadním funkčním rozšířením repozitáře GENASIS, proto je třeba jej stejně jako všechny ostatní repozitáře související s expozicí prostředí a člověka toxickým látkám – budovat jako jeden komplex tak, aby bylo možné všechna data propojovat směrem k analýze environmentálních a zdravotních rizik. Pro ilustraci celá kaskáda: Primární spektroskopická data představují obrovský objem dat srovnatelný s daty genomickými a podobně jako necílená data genomická jsou zdrojem pro dlouhodobý výzkum. Produktem prvotních analýz jsou tabulky jednotlivých signálů k budoucí potenciální identifikaci. Dalším krok je kvalitativní, tedy určení chemické identity signálu, a poslední krok kvantitativní, tedy určení koncentrace dané látky. Teprve tento výsledek může být uložen do repozitáře GENASIS, ale všechny ostatní musí být dostupné uživatelům jinde. Celý komplex těchto úložišť by se měl v budoucnu stát jádrem datových zdrojů výzkumné infrastruktury EIRENE.

Specifickým rysem přitom jde, že se nejedná o statický repozitář, ale zdroj pro neustálé další vytěžování, k čemuž je ale zapotřebí pokročilých bioinformatických technik. I tyto kaskády

¹³⁴ Kódy výstupů podaktivity jsou provázány na přílohu č. 3 Harmonogram klíčových aktivit a kap. 11 Hlavní plánované výstupy/produkty Studie proveditelnosti.

bioinformatických nástrojů je třeba neustále vyvíjet tak, aby se postupně zvyšoval podíl správně identifikovaných, případně kvantifikovaných signálů. Proto je součástí této klíčové aktivity i vývoj těchto nástrojů, jak bude popsáno dále. Na vývoji těchto softwarů se bude podílet mnoho světových týmů, proto je třeba je implementovat ve sdíleném prostředí, v tomto případě prostředí Galaxy spravovaném infrastrukturou ELIXIR. Právě propojení nově budovaných datových a výpočetních kapacit EIRENE a ELIXIR nabídne výzkumné komunitě dosud nedostupné zdroje.

U necílených analýz lidských vzorků je nespornou výhodou, že stejnou metodou můžeme měřit antropogenní polutanty i produkty metabolismu (a tedy spojit studium lidské expozice/expozomu a metabolomu) a přidat tak další vrstvy do multiomických analýz. Znamená to ovšem potřebu dlouhodobě ukládat velké objemy dat z vysokorozlišovací hmotnostní spektrometrie (primárních i odvozených) a umožňovat nejen jejich opakované prohledávání tak, jak roste objem našich poznatků o expozici populace škodlivým látkám a jejich škodlivých účincích, ale i jejich propojení s daty charakterizujícími lidský genom, transkriptom, proteom či mikrobiom. Jedině tak můžeme očekávat objevy nových biomarkerů účinku chemických expozic či biomarkerů rozvoje chronických stavů a významný pokrok personalizované medicíny a prevence.

Několik repozitářů pro ukládání dat z necílené metabolomiky ve světě existuje, jejich metadatová struktura ovšem neumožňuje třídění dostupných dat a opakované analýzy jejich relevantních podskupin. Proto se v této aktivitě zaměřujeme na vytvoření nového metadatového modelu, ve kterém bude výrazně rozšířen minimální soubor metadat. Cílem je vytvořit strukturu, která do budoucna otevře možnost integrace rozsáhlých dat z různých domén a analýzu genetických i negenetických faktorů, které stojí za rozvojem chronických onemocnění. Zároveň nemáme ambici budovat celosvětové úložiště, ale jen všeobecně akceptovaný model takového úložiště, který bude využíván primárně referenční laboratoří evropské infrastruktury EIRENE, případně clustrem referenčních laboratoří. V dalších krocích (mimo tento projekt) pak předpokládáme úzkou spolupráci mezi ESFRI infrastrukturou EIRENE a EMBL/EBI na vytvoření zrcadlového velkokapacitního repozitáře spravovaného EMBL/EBI a používaného mezinárodně. Tento postup zajistí mezinárodní interoperabilitu a možnost opakovaného integrovaného vytěžování metabolomických dat. Aby však bylo v budoucnu možné i propojení těchto dat s daty z genomiky a dalších „omic“, bude nutné nastavit bližší spolupráci s těmito odbornými komunitami a společně budovat datová rozhraní a cesty k propojení relevantních datasetů. Toto jde však za hranice OS II.

V rámci projektu OS II se zaměříme na vytvoření repozitáře, který umožní ukládání všech stupňů metabolomických dat (od primárních záznamů z plynové či kapalinové chromatografie a hmotnostní spektrometrie s vysokým rozlišením přes anotace, identifikace až po kvantifikace) i rozsáhlých metadat umožňujících nejen jejich opakované analýzy, ale i budoucí propojení na data z jiných domén.

Na základě našich zkušeností s dlouhodobým ukládáním a zpracováním těchto dat bude vytvořen formální datový model v návaznosti na využitou ontologii, zavedeny a zveřejněny metadatové standardy. Sémantická interoperabilita bude zajištěna provázaností na řízené slovníky, thesaury a ontologie (podstatné zejména pro chemické klasifikace, např. CAS, InChI). Jednotlivým datovým objektům budou přiřazeny citovatelné perzistentní identifikátory (v tomto případě jsou primární datové záznamy neměnné a každý nově vytvořený soubor sekundárních dat bude opatřen novým identifikátorem) a budou vytvořena aplikační rozhraní pro výměnu metadat spravovaných datasetů a pro výměnu samotných dat. Perzistentní identifikátory budou přidělovány i exportovaným výstupům z repozitáře (tedy sekundárním datům, např. tabulkám anotovaných signálů) pro zajištění citovatelnosti. Bude dořešen licenční model pro sdílení a využívání primárních i sekundárních dat, i s ohledem na citlivost potenciálního propojování s daty z jiných domén. Nový repozitář bude registrován v NKR, o datasetech spravovaných v tomto repozitáři budou poskytována povinná

metadata do NMA, autentizace a autorizace uživatelů bude využívat společnou autentizační a autorizační infrastrukturu (dále též „AAI“), kterou provozuje NRP a kdekoli bude možné, bude provedeno napojení na technické nástroje přesunu dat, metodiky a standardy NRP.

Repozitář bude v prvním kroku naplněn dostupnými daty a metadaty infrastruktury EIRENE/RECETOX, optimalizován a validován, v dalším kroku pak nabídnut k využití širokou komunitou. V souvislosti s touto aktivitou je plánována jedna zahraniční pracovní cesta na metabolickou konferenci z důvodu sdílení zkušeností s mezinárodní harmonizací standardů a diskuse s odbornou komunitou.

PŘEHLED KÓDŮ VÝSTUPŮ PODAKTIVITY¹³⁵

E_2	Repozitář pro ukládání dat z necílených hmotnostně spektroskopických analýz pro účely hodnocení lidského expozomu
E_2.1	Metadatový model pro primární hmotnostně spektroskopická data z necílené metaboliky a expozomiky lidských vzorků
E_2.2	Sety metadat příslušných k primárním hmotnostně spektroskopickým datovým sadám z necílených analýz pro studium expozomu a metabolomu
E_2.3	FAIRifikované datasety primárních hmotnostně spektroskopických dat z necílených analýz pro studium expozomu a metabolomu

10.7.3. PODAKTIVITA 8.3 – REPOZITÁŘ PRO UKLÁDÁNÍ TOXIKOLOGICKÝCH A EKOTOXIKOLOGICKÝCH DAT

Pro analýzu environmentálních a zdravotních rizik spojených s kontaminací prostředí chemickými látkami je nutné doplnit data o výskytu těchto látek v environmentálních matricích a lidských tkáních (uložená v repozitáři GENASIS) daty o jejich škodlivých účincích (hazardech), to znamená daty toxikologickými a ekotoxikologickými. Jednotný repozitář takových dat však celosvětově neexistuje a toxikologická data jsou publikována pouze ve vědeckých publikacích.

Věc se dále komplikuje tím, že na živé organismy nepůsobí nikdy pouze jedna látka (jak se běžně testuje v testech toxicity), ale komplexní směs. Při analýze chemických rizik bychom tedy měli optimálně brát v úvahu současné synergické či antagonistické působení všech přítomných látek. Jedním ze způsobů, jak se tomuto ideálu přiblížit, je koncept tzv. „adverse outcome pathways“ (AOP, dráhy škodlivého účinku). V takovém případě se sledují jednotlivé biologické dráhy, na které mohou toxické látky působit, a skládají se do nich informace o síle účinku všech relevantních stresorů. Celý koncept ale naráží na nedostatek nebo špatnou dostupnost toxikologických dat.

Analýzu potřeb mezinárodní výzkumné komunity provedeme ve spolupráci s evropským Partnerstvím pro analýzu chemických rizik (PARC), které sdružuje přední evropské toxikology. Na to bude navazovat návrh struktury toxikologického repozitáře a metadatového modelu, přičemž oba koncepty budou opět diskutovány s mezinárodní vědeckou komunitou. Mezinárodní konsensus a interoperabilita jsou klíčové pro budoucí úspěch projektu, protože repozitář bude, podobně jako dva předchozí, součástí evropské ESFRI infrastruktury EIRENE a bude využíván mezinárodně.

Návazně na vytvoření formálního datového modelu budou zavedeny a zveřejněny metadatové standardy. Podobně jako v případě expozičního repozitáře GENASIS bude sémantická interoperabilita

¹³⁵ Kódy výstupů podaktivity jsou provázány na přílohu č. 3 Harmonogram klíčových aktivit a kap. 11 Hlavní plánované výstupy/produkty Studie proveditelnosti.

zajištěna provázaností na řízené slovníky, thesaury a ontologie. Propojení dat o expozici (GENASIS) a hazardu (tento repozitář) bude možné díky jednotné chemické klasifikaci (CAS, InChI). Perzistentní identifikátory budou pro zajištění citovatelnosti přidělovány pravidelně publikovaným extraktům z repozitáře zahrnujícím průběžné updaty a datové přírůstky. Tyto ustálené strojově vytvářené datasety budou automaticky publikovány do NMA, bude ale možné vytvářet také extrakty na základě individuálních žádostí o vydání dat. Nový repozitář bude registrován v NKR a o datasetech spravovaných v tomto repozitáři budou poskytována povinná metadata do NMA. Bude nastaven licenční model a autentizace a autorizace uživatelů bude využívat společnou autentizační a autorizační infrastrukturu AAI. Repozitář bude pilotně testován na dostupných toxikologických a ekotoxikologických datech a metadatach centra RECETOX před tím, než bude otevřen mezinárodní vědecké komunitě (prostřednictvím „minizáměrů“ i infrastruktury EIRENE).

PŘEHLED KÓDŮ VÝSTUPŮ PODAKTIVITY¹³⁶

E_3	Repozitář pro ukládání toxikologických a ekotoxikologických dat
E_3.1	Analýza potřeb výzkumné komunity v oblasti správy toxikologických a ekotoxikologických dat a dostupných řešení pro přípravu nového repozitáře a jeho metadatového modelu
E_3.2	Metadatový model pro toxikologická a ekotoxikologická data
E_3.3	Sety metadat pro toxikologická a ekotoxikologická data
E_3.4	FAIRifikované pilotní datasety toxikologických a ekotoxikologických dat

10.7.4. PODAKTIVITA 8.4 – NOVÉ ÚLOŽIŠTĚ UMOŽŇUJÍCÍ PROPOJENÍ GEOKÓDOVANÝCH DAT Z RŮZNÝCH DOMÉN

Máme-li studovat vliv environmentálních faktorů na zdraví, potřebujeme propojit dlouhodobá data o zdraví člověka (z dlouhodobých sledování populačních kohort či zdravotních záznamů pacientů) s informacemi o všech možných typech environmentálních stresorů (chemickém znečištění ovzduší, půdy, pitné vody, přírodní katastrofy, horka, sucha, snížené biodiverzity) působících v místě jeho pobytu. Zároveň ale můžeme v konkrétních místech sledovat i vlivy ekonomické nebo sociální (např. sociálně vyloučené lokality). Současné hodnocení všech faktorů působících na člověka a jeho zdraví je principem konceptu lidského expozomu. Tento holistický přístup je ale přenositelný na studium zdraví jakékoli populace (tzv. eko-expozom nebo také princip „one health“ neboli jednoho zdraví).

Abychom mohli sledovat zdravotní dopady (zdravotnická data, výsledky analýzy biologického materiálu) všech faktorů působících v jednom místě ve stejném čase, je třeba tyto informace ukládat ve vazbě na geografický informační systém. Úložiště geokódovaných dat budované v rámci tohoto projektu umožní propojení GIS vrstev pocházejících z různých domén do jednoho funkčního systému, který umožní kvantifikovat environmentální rizika jednotlivých oblastí.

Podobná úložiště jsou budována i na jiných místech, například v rámci projektů Evropské sítě lidského expozomu (EHEN) nebo US projektu NEXUS. I v tomto projektu tedy budeme postupovat v úzké spolupráci s mezinárodními partnery a dbát na zachování mezinárodních standardů, interoperability a dlouhodobé využitelnosti dat. V tomto případě jde ale o vytvoření českého modelu, který shromáždí maximum lokálně relevantních dat, která nemohou být převzata z mezinárodního prostředí.

¹³⁶ Kódy výstupů podaktivity jsou provázány na přílohu č. 3 Harmonogram klíčových aktivit a kap. 11 Hlavní plánované výstupy/produkty Studie proveditelnosti.

Budeme provozovat mapový server s aplikačním rozhraním ve standardu WMS/WFS. Při popisu metadatových profilů každé mapové vrstvy navážeme na standard ČSN ISO 19107 (Geografické informace – Prostorové schéma). Analýza potřeb výzkumné komunity bude provedena ve spolupráci s projekty PARC, IHEN a NEXUS a bude klíčem k bližší specifikaci struktury repozitáře, jeho metadatového modelu a metadatových standardů. Nový repozitář bude registrován v NKR, o jednotlivých mapových vrstvách spravovaných v tomto repozitáři budou poskytována povinná metadata do NMA. Tyto datové extrakty budou opatřeny perzistentními identifikátory pro zajištění citovatelnosti. Autentizace a autorizace uživatelů bude využívat společnou AAI, kterou provozuje NRP.

Repozitář bude pilotován s využitím dlouhodobě budovaných GIS vrstev dostupných na MU. Další partneři budou motivováni ke sdílení nových typů dat, k mezioborové spolupráci a využívání dostupných informací v nejrůznějších výzkumných doménách i aplikačních sektorech. Umožní-li to licenční podmínky, budeme mapové vrstvy uživatelům nabízet ke stažení, ale zejména budeme přes tento mapový servis poskytovat přístup ke spravovaným mapovým vrstvám pro uživatelskou komunitu. Dále budeme vyvíjet geoprocesingové služby, které na základě zadaných prostorových souřadnic poskytnou informace odvozené z jednotlivých mapových vrstev (to umožní zpracovávat výsledky i jedincům, kteří neovládají ArcGIS).

PŘEHLED KÓDŮ VÝSTUPŮ PODAKTIVITY¹³⁷

E_4	Nové úložiště umožňující propojení geokódovaných dat z různých domén
E_4.1	Analýza potřeb výzkumné komunity v oblasti propojování geokódovaných dat z různých oblastí a dostupných řešení pro přípravu nového repozitáře a jeho metadatového modelu
E_4.2	Metadatový model pro úložiště geokódovaných dat

10.7.5. PODAKTIVITA 8.5 –REPOZITÁŘ PRO REFERENČNÍ OBRAZOVÁ DATA FLÓRY A ROSTLINNÝCH SPOLEČENSTEV

Referenční obrazová data jsou fotografie objektů (živých rostlin, jejich společenstev či dílčích orgánů, jako jsou třeba pylová zrna či mikroskopické řezy dřevem atp.), které jsou systematicky jednoznačně a přesně klasifikovány. Jsou zásadní jako referenční vzorky pro základní výzkum v oblasti botaniky, lichenologie, paleoekologie, ale i pro aplikovaný výzkum a ochranu biodiverzity. V oblastech, kde se ve větší míře uplatňuje strojová identifikace environmentálních vzorků (jako je oblast pylové analýzy) je pak jejich význam pro trénování klasifikačních modelů zcela nezastupitelný.

Tato data jsou již dnes využívána vědeckou komunitou doma i v zahraničí díky zpřístupnění v rámci aplikací, jako jsou např. český botanický portál Pladias¹³⁸, portál evropské vegetace¹³⁹, česká palynologická databáze¹⁴⁰ nebo paleoekologická databáze Neotoma¹⁴¹ či botanická fotogalerie¹⁴².

Partneři OS II projektu (MU, UK) společně s Botanickým ústavem AV ČR představují největší správce referenčních botanických dat v České republice. Spravovaná data jsou uložena fragmentovaně, tj. po jednotlivých institucích a dílčích projektech, z nichž byla data vytvořena, bez náležité integrace, jednotného metadatového popisu a jednotné správy dat tak, aby odpovídala FAIR principům. Převod

¹³⁷ Kódy výstupů podaktivity jsou provázány na přílohu č. 3 Harmonogram klíčových aktivit a kap. 11 Hlavní plánované výstupy/produkty Studie proveditelnosti.

¹³⁸ <https://pladias.cz>.

¹³⁹ <https://floraveg.eu>.

¹⁴⁰ <https://botany.natur.cuni.cz/palycz/>.

¹⁴¹ <https://www.neotomadb.org/>.

¹⁴² <https://www.botanickafotogalerie.cz>.

do repozitářové platformy je již úspěšně aplikován na skeny herbářových položek v rámci pilotních aktivit projektu NRP, a získané zkušenosti a aktivizovanou komunitu využijeme pro další typ dat, u kterých jsou přítomny oba základní požadavky pro repozitář: potřeba velké kapacity úložiště (zejména u strojově generovaných dat z mikroskopických preparátů) a zajištění konektivity a citovatelnost při jejich publikaci – vedle vědeckého užití jsou totiž referenční data také velmi potřebná a hojně užívána pro prezentaci botanických taxonů/dat širší veřejnosti.

Vytvoříme proto samostatný repozitář typu Invenio pro referenční obrazová data v rámci Národní repozitářové platformy se vhodným metadatovým modelem a naplněním všech požadavků (dokumentace, politika, metodická podpora) pro integraci v rámci NMA a přesuneme do něj existující data. Na úrovni přípravy dat pro import do repozitáře proběhne nejprve mobilizace týmů v rámci obou institucí, následně budou získaná data standardizovaná dle metadatového modelu a finálně publikována v rámci repozitáře.

Hlavními výstupy podaktivity budou: funkční repozitář s národním zaměřením a dostatečnou úložnou kapacitou pro referenční botanická obrazová data; metadatový model pro zahrnutá data, jeho integrace v rámci clusteru biodiverzitních repozitářů v rámci NRP s výhledem pro interoperabilitu i v rámci nadnárodních aktivit (zejména ESFRI DiSSCo.eu) a ostatní náležitosti provozování repozitáře. FAIRifikované datasety získané mobilizací široké vědecké komunity v rámci projektu OS II. Součástí aktivity jsou zahraniční služební cesty s cílem zabezpečení interoperability repozitářových platform a metadatových standardů s konsorciem Distributed System of Scientific Collections¹⁴³ a JACQ.

PŘEHLED KÓDŮ VÝSTUPŮ PODAKTIVITY¹⁴⁴

E_5	Repozitář pro referenční obrazová data živých rostlin a rostlinných společenstev
E_5.1	Metadatový model pro referenční obrazová data rostlin a rostlinných společenstev
E_5.2	Sety metadat pro referenční obrazová data rostlin a rostlinných společenstev
E_5.3	FAIRifikovaný dataset obrazové referenční sbírky flóry a vegetace
E_5.4	Metodika tvorby referenční sbírky obrazových dat flóry a vegetace, tvorby dat a metadat, jejich FAIRifikace, ukládání a správy

10.7.6. PODAKTIVITA 8.6 – REPOZITÁŘ PRO GENETICKÝ BIOMONITORING A GENETICKÁ DATA

Genetický biomonitoring umožňuje masivní analýzu a efektivní sdílení biodiverzitních dat z environmentálních vzorků za využití sekvenování směsných vzorků a vyhodnocení výskytu taxonů s porovnáním s referenčními daty. V posledních desetiletích se genetické metody monitoringu volně žijících organismů široce rozšířily, a přesto zatím postrádají sjednocenou národní platformu pro uložení dat. Vytvoříme proto nový repozitář pro data pocházející z genetického biomonitoringu a genetická data volně žijících organismů. Repozitář bude orientován napříč taxonomickými skupinami a bude zahrnovat zejména tyto typy dat:

a/ obrazová data genetického biomonitoringu organismů, která jsou získávána robotickým přístupem a jsou propojena s výsledky metabarkódového sekvenování volně žijících organismů získaných z environmentálních vzorků, jakými jsou například hromadné odchyty hmyzu;

¹⁴³ <http://discco.eu>.

¹⁴⁴ Kódy výstupů podaktivity jsou provázány na přílohu č. 3 Harmonogram klíčových aktivit a kap. 11 Hlavní plánované výstupy/produkty Studie proveditelnosti.

b/ data z měření velikosti genomu a ploidie, případní další charakteristiky buněčného jádra volně žijících organismů, získané z průtokové cytometrie, tedy z oblasti, kde ČR je světovým lídrem v aplikaci těchto metod na výzkum evoluce a biodiverzity;

c/ cytogenetická obrazová data z fluorescenční mikroskopie mitotických a meiotických jader buněk. V této oblasti je ČR rovněž jedním ze světových lídrů aplikace cytogenetických metod na výzkum diverzity a evoluce organismů. Výše zmíněné datasety jsou v současnosti uloženy na vícero akademických pracovištích a v muzeích v ČR – jednak u partnerů projektu (Univerzita Karlova, Masarykova Univerzita), ale i u mimopartnerských institucí (Botanický ústav AV ČR, Ústav experimentální botaniky AV ČR, Biologické centrum AV ČR, Ústav živočišné fyziologie a genetiky AV ČR, Ústav biologie obratlovců, Přírodovědné muzeum Národního muzea atp.) Hlavním výstupem podaktivity E03 budou: funkční repozitář s národním zaměřením a dostatečnou úložnou kapacitou pro genetický biomonitoring a genetická data; jeho integrace v rámci klusteru biodiverzitních repozitářů v rámci NRP s výhledem pro interoperabilitu i v rámci nadnárodních aktivit (zejména ESFRI DiSSCo.eu); řízený přístup, ukládání a správa dat na jednom místě prostřednictvím AAI v rámci NDI a ostatní náležitosti provozování repozitáře NRP; jednotná metodika a standardy. Podstatnými výstupy budou také publikované FAIRifikované datasety získané mobilizací široké vědecké komunity v rámci institucí zapojených v rámci projektu OS II.

PŘEHLED KÓDŮ VÝSTUPŮ PODAKTIVITY¹⁴⁵

E_6	Nový repozitář pro genetický biomonitoring a genetická data volně žijících organismů
E_6.1	Metadatový model pro genetickou banku volně žijících organismů
E_6.2	Sety metadat pro genetický biomonitoring a genetických dat volně žijících organismů
E_6.3	FAIRifikovaný dataset genetického monitoringu volně žijících organismů (rostlin a hub)
E_6.4	Metodika tvorby dat a metadat genetického monitoringu volně žijících organismů a jejich FAIRifikace

10.7.7. PODAKTIVITA 8.7 – REPOZITÁŘ PRO ZOOLOGICKÉ SBÍRKY

Vědecké sbírky jsou spravované akademickými (univerzity a ústavy Akademie věd) a muzejními institucemi (zejména Přírodovědné muzeum Národního muzea a Moravské zemské muzeum, ale i mnohá regionální muzea). Pod vědeckými sbírkami zoologické povahy se rozumějí kolekce lihových, dermoplastických a osteologických preparátů nejrůznějších druhů obratlovců i bezobratlých živočichů, které představují základní kameny pro interpretaci a reprodukovatelnost vědeckých dat v oblasti taxonomie a diverzity na národní a nadnárodní úrovni. Na jejich materiálním základě vznikají digitální popisy (koncept digital twins) a obrazová data (fotografie a skeny ve vysokém rozlišení, včetně 3D zobrazení) umožňující jejich zpřístupnění online, a zároveň jsou materiálem pro pokročilé zobrazovací a analytické techniky (genetické, tomografické atd.). Souběžně je každý sbírkový předmět také dokladem nálezu v daném místě a čase, a je tak zdrojem dat o rozšíření organismů.

Vytvoření nového repozitáře vychází z potřeb výzkumné komunity v ČR, která v současné době spravuje obrazová data a metadata zoologických sbírek velmi fragmentovaně, tj. po jednotlivých institucích (zejména Univerzita Karlova, Ústav biologie obratlovců, Národní muzeum), bez náležité integrace, jednotného metadatového standardu a jednotné správy dat tak, aby odpovídala FAIR

¹⁴⁵ Kódy výstupů podaktivity jsou provázány na přílohu č. 3 Harmonogram klíčových aktivit a kap. 11 Hlavní plánované výstupy/produkty Studie proveditelnosti.

principům. Vytvořením příslušného metadatového modelu zajistíme interoperabilitu dat jednak v rámci klustru repozitářů pro biodiverzitní data BioDiv a jednak na mezinárodní úrovni – propojení na ESFRI infrastrukturu DiSSCo.cz a celosvětově největší agregátor biodiverzitních dat – GBIF¹⁴⁶.

V rámci projektu vytvoříme samostatný repozitář typu Invenio pro referenční obrazová data zoologických sbírek v rámci Národní repozitářové platformy se vhodným metadatovým modelem a naplněním všech požadavků (dokumentace, politika, metodická podpora) pro integraci v rámci NMA a přesuneme do něj existující data. Na úrovni přípravy dat pro import do repozitáře budou získaná data standardizovaná dle metadatového modelu a finálně publikována v rámci repozitáře.

PŘEHLED KÓDŮ VÝSTUPŮ PODAKTIVITY¹⁴⁷

E_7	Nový repozitář pro zoologické sbírky
E_7.1	Metadatový model pro zoologické sbírky
E_7.2	Sety metadat zoologických sbírek
E_7.3	FAIRifikovaný dataset zoologických sbírek
E_7.4	Metodika tvorby dat a metadat zoologických sbírek, jejich FAIRifikace, ukládání a správy

10.7.8. PODAKTIVITA 8.8 – VÝVOJ A PILOTNÍ ZAVEDENÍ SLUŽEB PRO ANALÝZU ENVIRONMENTÁLNÍCH A BIODIVERZITNÍCH DAT

Jak bylo popsáno výše, nový repozitář dat z necílených spektroskopických analýz s vysokým rozlišením bude cenným zdrojem informací, které mohou být vytěžovány různými uživateli po velmi dlouhou dobu. Podmínkou je ovšem existence pokročilých softwarových nástrojů na prohledávání takových dat, tedy modulů na čištění signálu, identifikaci píků, predikci hmotnostních spekter, kvalitativní nebo kvantitativní analýzu. Nejedná se přitom o jeden software, ale o bohatou kaskádu nástrojů, které může uživatel podle potřeby kombinovat. Budou-li takové nástroje dostatečně kvalitní, dokumentované, validované a otevřeně dostupné, přispěje to ke standardizaci celého procesu vytěžování dat a důvěryhodnosti získaných informací.

Návazně na nově vybudovaný repozitář metabolomických dat budou proto v rámci tohoto projektu budovány pokročilé nástroje pro analýzu dat z necílených GC/MS (gas chromatography/mass spectroscopy) a LC/MS (liquid chromatography/mass spectroscopy) analýz pro metabolomiku a funkční expozomiku s cílem zlepšit dostupnost a opětovnou využitelnost výzkumných dat dle potřeb výzkumné komunity i aplikačního sektoru. Všechny tyto nástroje pro práci s dostupnými daty budou implementovány ve spolupráci s infrastrukturou ELIXIR (a jejím českým koordinačním uzlem na ÚOCHB AV ČR) v prostředí Galaxy, které zajišťuje jejich plnou dokumentaci a otevřený přístup pro všechny uživatele. To přispěje ke snižování bariér mezioborového využívání výzkumných dat v globálním měřítku, zajistí propojení s mezinárodními aktivitami, standardizaci všech nástrojů a jejich širokou dostupnost. Součástí budovaných služeb bude i příprava nových spektrálních knihoven, standardních operačních postupů pro produkci a zpracování dat, tréninkových materiálů a tutoriálů. V souvislosti s touto aktivitou je plánována jedna zahraniční pracovní cesta na konferenci Galaxy komunity z důvodu sdílení zkušeností s implementací metabolomických nástrojů a diskuse s odbornou komunitou.

¹⁴⁶ <https://www.gbif.org>.

¹⁴⁷ Kódy výstupů podaktivity jsou provázány na přílohu č. 3 Harmonogram klíčových aktivit a kap. 11 Hlavní plánované výstupy/produkty Studie proveditelnosti.

Pro potřeby repozitářů biodiverzitních dat budeme vyvíjet a aplikovat softwarové nástroje umělé inteligence týkající se (i) analýzy obrazu vzorků biodiverzitní povahy, (ii) aplikace velkých jazykových modelů (Large Language Models – LLM), které umožní automatizované čtení herbářových sched a záznamů zoologických sbírek za účelem spolehlivého, masivního a efektivního zpřístupňování metadat biologických sbírek široké odborné i laické veřejnosti a které budou součástí repozitářů biodiverzitních dat v rámci NRP.

ČINNOSTI:

- pipelines na analýzu hmotnostně spektroskopických dat z necílených metabolomických a expozomických analýz; E_8
- AI nástroje na analýzu obrazu a automatické čtení herbářových sched. E_9

PŘEHLED KÓDŮ VÝSTUPŮ PODAKTIVITY¹⁴⁸

E_8	Nové bioinformatické nástroje na analýzu hmotnostně spektroskopických dat z necílených metabolomických a expozomických analýz
E_9	AI nástroje na analýzu obrazu biologických objektů a automatické čtení herbářových sched

10. 8. KA 9 – TEMATICKÉHO CLUSTERU CITLIVÁ DATA

ZAPOJENÍ PARTNEŘI: CESNET (Garant TKA), MU, UPOL, ZČU, SÚ AV ČR, ÚFCH JH AV ČR

Tematický cluster Citlivá data přidá do projektu vrstvu správy citlivých dat v celém prostoru NRP a NDI. Naváže na podporu činností nezbytných pro FAIRifikaci dat v repozitářích a práci nad licenčními a dalšími právními režimy/modely (v NRP aktuálně pro data v režimu „necitlivá“, tedy veřejná). V NRP a NDI je infrastruktura umožňující správu dat na vysoké úrovni, ale bez potřebných nástrojů a funkcionalit nezbytných pro správu citlivých dat. TKA SENSÍ vytvoří detailní a návodné postupy FAIRifikace citlivých dat, nástroje či služby aplikovatelné v rámci životního cyklu správy citlivých dat a v neposlední řadě se zaměříme na správnou praxi sdílení a zpracování citlivých dat (např. v prostředích typu Trusted Research Environment¹⁴⁹) vyvstalých mj. i ze spolupráce mezi akademickou a soukromou sférou. Nově vznikající nebo inovované repozitáře, které budou v rámci OS II integrovány do NRP a mají ambici uchovávat citlivá data, budou ve spolupráci s odborníky TKA SENSÍ tyto politiky, postupy a služby či nástroje vyvíjet. Tematická klíčová aktivita si klade za cíl vytvořit jasná a harmonizovaná pravidla pro zpřístupnění citlivých dat, která za stávajících podmínek nemohou být sdílena. Inspirováni tzv. FAIR principy¹⁵⁰ vytvoříme metodiku FAIRifikace citlivých dat použitelnou v celé NDI a všemi repozitáři obsahujícími citlivá data. V návaznosti na technické nástroje vznikající v projektu NRP i NDI, jako je Fair Implementation Profile¹⁵¹ (včetně FIP Wizard¹⁵²) anebo vzdělávání pracovníků datové podpory a dalších odborníků v rámci komunity, naše aktivity budou zaměřeny na správce a datové kurátory repozitářů, poskytovatele dat a jejich spolupráce s žadateli o data. Metodika FAIRifikace citlivých dat vznikne na základě spolupráce s nově vznikajícími repozitáři a zkušeností expertů z TKA SENSÍ.¹⁵³ Metodika a vzniklá pravidla správy citlivých dat v NRP, resp. NDI budou

¹⁴⁸ Kódy výstupů podaktivity jsou provázány na přílohu č. 3 Harmonogram klíčových aktivit a kap. 11 Hlavní plánované výstupy/produkty Studie proveditelnosti.

¹⁴⁹ Zabezpečená platforma pro správu a analýzu citlivých dat, <https://ukhealthdata.org/developing-technology-services/trusted-research-environments/>.

¹⁵⁰ <https://opendata.muni.cz/fair-a-open-data/fair-data-a-jak-na-ne/hlavni-principy-fair>.

¹⁵¹ <https://www.go-fair.org/how-to-go-fair/fair-implementation-profile/>.

¹⁵² <https://fip-wizard.readthedocs.io/en/latest/about/about.html>.

¹⁵³ Viz kap. 9 Organizační struktura, řízení projektu a popis rolí realizačního týmu.

základem školení pro akademiky spravující citlivá data. Naše TKA připraví vzdělávací materiály pro tato školení a zajistí školitele z řad expertů naší TKA.¹⁵⁴

V rámci projektu vznikne v jednotlivých TKA sedm repozitářů obsahujících nějakým způsobem citlivá data. Se čtyřmi z nich bude tým TKA SENSI úzce spolupracovat a pravidelně konzultovat přípravu metodiky FAIRifikace citlivých dat, služeb i nástrojů. Se zbylými třemi repozitáři budeme spolupracovat v průběhu celého projektu a dle potřeby.

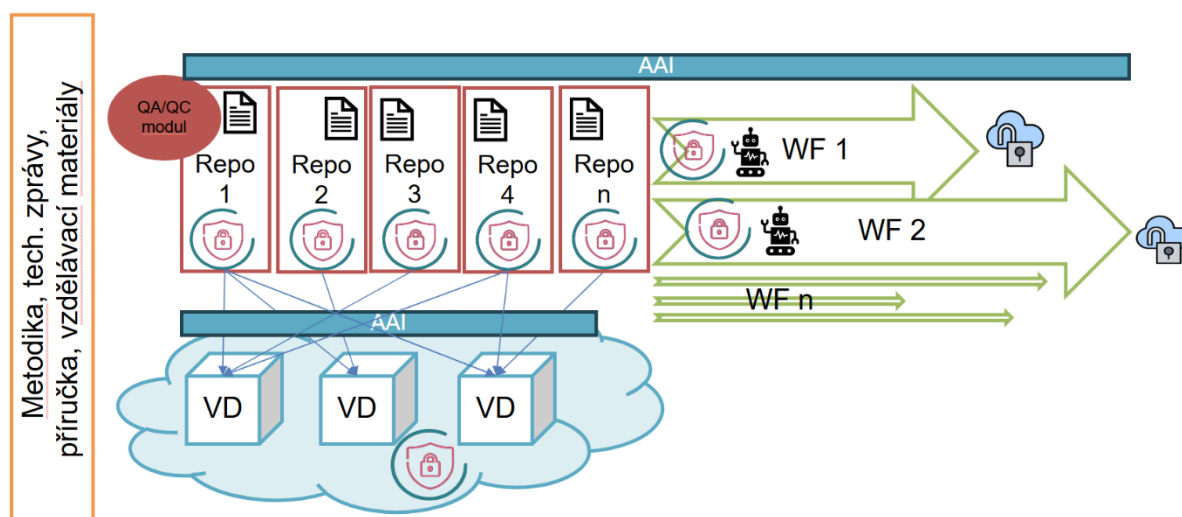
Zmiňované FAIR principy budou aplikovány týmem TKA v následujících aktivitách napříč NDI:

- **Dohledatelnost** – repozitáře NRP obsahující citlivá data budou obsahovat datasety řádně popsané metadaty; zacházení s metadaty citlivých datasetů se může lišit napříč NRP dle potřeb repozitářů; perzistentní identifikátory a metadata jsou základem dohledatelnosti dat a podporou reprodukovatelnosti výsledků získaných analýzou citlivých datasetů.
- **Přístupnost** – budou stanovena transparentní pravidla přístupů k datům a jasné podmínky užití dat nastavená dle potřeb každého repozitáře; vytvoříme službu (do jisté míry automatizovaných) workflows umožňujících analýzu citlivých dat v rámci NDI bez potřeby přístupu žadatelem nebo přenosem a zpracováním v zabezpečeném interaktivním výpočetním prostředí.
- **Interoperabilita** – z pohledu TKA SENSI je důležité, aby citlivá data v repozitářích byla ve standardních a otevřených formátech umožňujících propojování dat napříč datasety a současně byla nastavena bezpečnostní pravidla pro takové propojení, resp. následné sdílení.
- **Znovuvyužitelnost** – navážeme na kontrolu technické kvality v jednotlivých repozitářích a zaměříme se na kvalitu dat vycházející z kontroly obsahu datasetu (tzv. plausability); TKA SENSI vytvoří transparentní podmínky, za kterých je možné citlivá data nebo přístup k nim žádat, vč. strojově čitelné licence umožňující automatizovaným systémům, aby mohly provádět operace s daty v souladu s licenčními podmínkami bez nutnosti lidského zásahu.

Vedle obecné metodiky FAIRifikace citlivých dat a nastavování pravidel přístupů k datasetům a sdílení dat jednotlivými repozitáři zavedeme služby a nástroje pro rozvoj NDI v oblasti zpracování citlivých dat. Účelem je poskytnout uživatelům NDI (výzkumníkům, výzkumným skupinám apod.) bezpečné prostředí, kde mohou zpracovávat citlivá data. Tato citlivá data přitom mohou být získávána či ukládána přímo z repozitářů v NDI. Zaměřujeme se přitom na vývoj a zavádění služeb, které:

1. umožní implementaci bezpečného procesu příjmu a vydání citlivých dat do/z prostředí pro jejich zabezpečené zpracování (např. do/z úložišť a repozitářů v NDI);
2. umožní implementaci vhodného rozhraní pro uživatele pro práci s těmito daty v daném zabezpečeném prostředí (např. formou speciálních VPN, virtuálních desktopů (VD) pro interaktivní práci, rozhraním pro spouštění dávkových workflows (WF) apod.).

¹⁵⁴ Vzdělávací materiály jsou výstupem projektu OS II. Samotná školení budou realizována ve spolupráci se Školícím centrem EOSC CZ, které zajistí odpovídající kapacity, technické zázemí a organizační podporu, a nejsou tak zahrnuta do rozpočtu OS II.



Obr. 12: Grafické schéma aktivit a výstupů TKA SENSİ

Tato klíčová aktivita poskytne správcům a kurátorům dat repozitářů NRP softwarový modul pro kontrolu a zajištění kvality dat repozitářů (dále QA/QC modul¹⁵⁵). TKA SENSİ se rozhodla vyvíjet tento modul co nejuniverzálněji, avšak modul bude vyvíjen na základě zkušeností v oblasti QA/QC klinických studií a biomedicínského výzkumu. Do tohoto SW modulu budou mít správci, kurátoři i uživatelé přístup přes REST-API¹⁵⁶.

10.8.1. PODAKTIVITA 9.1 – TVORBA METODIKY A VZDĚLÁVACÍCH MATERIÁLŮ FAIRIFIKACE CITLIVÝCH DAT¹⁵⁷

Cílem této aktivity je zajistit, aby všechny repozitáře obsahující citlivá data splňovaly nejvyšší standardy ochrany dat a byly plně interoperabilní s NDI. Vytvoříme obecnou metodiku, pokrývající standardy a minimální požadavky na kvalitu (citlivých dat) v NRP, eticko-právní zajištění správy a zpracování citlivých dat pokrývající i duševní vlastnictví a spolupráci se soukromou sférou. Součástí této metodiky budou odkazy na nástroje a služby vytvořené touto TKA, uvedeny níže.

Správa a ukládání citlivých dat v NRP, resp. NDI není aktuálně vůbec řešena z eticko-právního pohledu i kvality dat. Očekáváme, že současná situace zaměřená na čistě technické řešení správy potenciálně citlivých dat se díky implementaci Metodiky změní na přehlednou a jasně uchopitelnou pro všechny subjekty, které budou chtít různé citlivá data do repozitářů ukládat i z nich žádat a zpracovávat je. Toto zahrne spolupráci s akademickými i soukromými subjekty.

Vypracování obecné metodiky správy citlivých dat v repozitářích budeme řešit v souladu s etickými principy/standardy pro výzkum a otevřenou vědu a s platnou evropskou a národní legislativou, včetně vytvoření pravidel a definice podmínek pro zpřístupnění a uvolnění dat k dalšímu využití ve vědě a výzkumu.

Mezi zásadní témata FAIRifikace citlivých dat zahrnujeme: bezpečné prostředí pro zpracování citlivých dat, doporučení bezpečnostních opatření, jako je šifrování dat při přenosu i v repozitáři, pravidelné audity, nastavení pravidel přístupu k datům a správu citlivých dat v bezpečném prostředí. Součástí

¹⁵⁵ QA = quality assurance; QC = quality control.

¹⁵⁶ REST = REpresentational State Transfer; API = Application Programming Interface; <https://restfulapi.net/>.

¹⁵⁷ V popisu KA jsou plánované aktivity propojeny na realizované výstupy. Výstupy jsou označeny kódem, pro TKA SENSİ v podobě SE_x (klíčový), SE_x.x (dílčí, tzn. náležící ke klíčovému).

bude popis nastavení dostatečných autentizačních mechanismů a technologií pro detekci a prevenci neoprávněného přístupu, v návaznosti na výstupy projektu NRP a také popis nastavení workflow pro správu citlivých dat s ohledem na jejich následné využití. Bude zde zpracováno téma skrytí citlivých metadat, uložení náležitých údajů do provenance informace a současně tím, jak existenci citlivých metadat dát vědět uživateli repozitáře. Zaměříme se na návrh obecného technického řešení repozitáře vhodného pro bezpečné uložení citlivých dat a v neposlední řadě eticko-právní návrhy pro sdílení citlivých dat včetně výběru licence pro datasety, propojování datasetů napříč repozitáři, kvalitu dat atp. Obecná metodika bude v rámci projektu konzultována a aplikována repozitáři NRP obsahující citlivá data.

Část metodiky zaměřené na etické a právní aspekty správy dat, bude zahrnovat:

- vzorové dokumenty pro sdílení dat (data transfer agreement, DTA) a politiky pro přístup (accessible usage policy, AUP), vzorové dotazy na žadatele o citlivá data ohledně plánovaného využití požadovaných dat;
- etické, právní a formální náležitosti pro dlouhodobé uchovávání výzkumných dat a metadat;
- etické, právní a formální náležitosti žádosti o znovuvyužití dat a jejich uvolnění ke znovuvyužití;
- zpřístupnění výzkumných dat v závislosti na typu prováděné analýzy (možnost využití anonymizovaných, pseudonymizovaných, de-identifikovaných a identifikovatelných dat) z eticko-právního náhledu;
- analýzu zapojení Datové komise (DAC, data access committee) do procesu uvolňování dat k jejich znovuvyužití včetně jejího složení, rozhodovacích kritérií, práv a povinností;
- návrh přístupové pipeline (workflow) na základě výše uvedených analýz;
- eticko-právní analýzu užití jednotného identifikátoru pro jednotlivce v rámci repozitářů spravujících citlivá (biomedicínská, technická, technologická, sociologická a další) data, tj. možnost propojení osobních údajů na úrovni jednotlivce mezi repozitáři;
- zhodnocení z etického hlediska – analýzu přínosu versus rizik s ohledem na základní lidská práva, zejména na právo na soukromí.

Finální dokument bude k dispozici v online podobě s možností (ne)pravidelných revizí a aktualizace v období po projektu.

Metodika bude provázána na konkrétní repozitáře oborových clusterů Bio/zdraví/potraviny, Sociální vědy, Enviromentální vědy, Datamanagement 4AI, Materiálové vědy, repozitářů vytvářených v projektu. Konkrétně jde o repozitáře, které zpracovávají citlivá data. Se všemi repozitáři pracujícími s citlivými daty bude dle potřeby v kontaktu bioetička, expertka na spolupráci se soukromou sférou, právníčka, AAI expert/ka, data steward/ka, vše za koordinace a pomoci garantky TKA¹⁵⁸.

- národní omický repozitář Czech Omics Node (OmiCZ);
- repozitář pro lidská i animální obrazová a fyziologická multimodální data;

¹⁵⁸ Tým TKA SENSI obsahuje expertízu na EHDS (Nařízení o evropském prostoru pro zdravotní údaje = European Health Data Space platné od března 2025, více na https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space-regulation-ehds_cs) resp. přípravu implementačních dokumentů tohoto nového nařízení a také GDI (Projekt datové infrastruktury pro genomická data, který se snaží vytvořit federovanou, udržitelnou a bezpečnou infrastrukturu pro přístup k genomickým a souvisejícím fenotypovým a klinickým datům napříč Evropou). Dále jsou členové TKA SENSI zapojeni v aktivitách VVI BBMRI-ERIC, EATRIS, EBRAINS, Bioimaging, ELIXIR, EOSC a dalších.

- nový repozitář ČSDA pro citlivá data pro sociální vědy;
- ClinData Repozitář;
- repozitář pro ukládání dat z necílených hmotnostně spektroskopických analýz pro účely hodnocení lidského expozomu;
- repozitář DM4AI;
- repozitář DANTE^c.

Analýza požadavků a příprava návrhu Metodiky bude zajištěna činností pozice datového analytika (experti s konkrétním zaměřením) a metodika (zná partnery, zná procesy a rozumí systémům a datům), bioetičky, právníčky a garantky TKA (specifické s ohledem na typ výzkumných dat, typ spolupráce a oborovost) na expertíze pro FAIRifikaci citlivých dat od životní fáze dat od přípravy pro zpřístupnění v repozitářích po sdílení za jasně definovaných podmínek. Budou zpracovávány podpůrné materiály a dokumentace pro vytváření bezpečných prostředí pro správu citlivých dat. Analýza správné praxe bude provedena v rámci přípravy metodiky, ale bude samostatným výstupem a zahrne několik úrovní správné praxe pro sdílení výzkumných dat získaných z kolaborativního výzkumu se soukromou sférou a výzkumu oborových clustřů zapojených v NRP.

Plánujeme tyto úrovně analýzy:

1. Úroveň potřeb partnerů ve výzkumu a specifik podle oborů s cílem mít co nejvíce obecnou metodiku FAIRifikace citlivých dat. Zohledníme specifika konkrétních repozitářů a spoluprací, např. analýzy pro potřeby sdílení technických dat z kolaborativního výzkumu budou prováděny jinak než pro data sociologická nebo klinická, a i pro technická data se předpokládá jiné zacházení s daty např. pro vesmírný výzkum třeba než pro energetiku.
2. Úroveň citlivosti metadat v repozitářích – v některých případech budou citlivá nejen data, ale i metadata (specifické spolupráce, jejíž náplní může být např. bezpečnostní výzkum).
3. Úroveň typů citlivých dat, včetně technik pro vytváření syntetických dat a zacházení s nimi. Na této úrovni je potřeba se zabývat i metodami anonymizace citlivých dat, na čemž budeme spolupracovat s TKA SOC, a zároveň i spojenými riziky reidentifikace. Případně se dá navázat na pilotní „minizáměr“ anonymizace, který se bude zabývat rešerší dostupných metod, doporučením postupů dostatečně anonymizující citlivá data a praktickým testováním dat ze zdravotnictví.
4. Úroveň potřeby nastavení metodiky v procesu mimo oborový repozitář v případě některých problematik duševního vlastnictví.
5. Úroveň eticko-právních aspektů správy citlivých dat lidského původu (např. sociologický výzkum, biomedicína – genomika, řečová, obrazová, klinická data a další, využívání umělé inteligence) SE_1.

V rámci přípravy metodiky FAIRifikace citlivých dat pro NDI budeme pracovat s výsledky tou dobou probíhající analýzy a budeme vycházet ze zkušeností odborného týmu. V této fázi naší práce bude probíhat nastavení obecných postupů vyvstalých v procesu vytváření oborového repozitáře, např.:

- stanovení workflow pro uchování dat a metadat (podmínky – formální náležitosti, úroveň potřeb dle citlivosti, doba uchování, anonymizační přístupy, nakládání s metadaty, ochrana obchodního tajemství atd.);
- stanovení workflow pro přístup k datům (od otevřeného přístupu až po proces zahrnující specifické řízení přístupu pro nakládání s utajovanými skutečnostmi, prostřední pro zpracování, právní aspekty a smluvní dokumentaci);

- stanovení variant licenčních podmínek při depozici dat;
- stanovení rolí a práv uživatelských skupin v repozitáři v jednotlivých částech workflow.

Budeme se zabývat licenčními podmínkami upravující přístup k citlivým datům v NRP i pro automatizované virtuální systémy (boty), jako jsou např. webové crawly¹⁵⁹, umělá inteligence (artificial intelligence, AI) strojové učení a další, kde hlubší analýza zatím chybí a není jasné, jak se k této problematice postavit. Zaměříme se na přípravu podkladů správy citlivých dat vznikajících ve spolupráci s komerční sférou a aplikovaným výzkumem. V této fázi přípravy Metodiky budeme úzce spolupracovat s týmem zpracovávající téma Compliance v průřezové klíčové aktivitě projektu OS II i s týmem pro danou aktivitu v projektu NRP, využijeme jejich průběžné reporty a dokumenty.

Do metodiky budeme zapracovávat jednotlivá doporučení a expertízy z konkrétních oblastí a repozitářů ohledně zpracování citlivých dat. Společně se správci repozitářů obsahujících citlivá data budeme definovat standardy a procesy správy citlivých dat v jejich oboru. Explicitně se jedná o repozitáře uvedené výše. Ve větší míře pak především v souvislosti s provozováním ČSDA repozitáře¹⁶⁰ a anonymizací citlivých dat z oboru sociologie nebo na příkladech expertíz v automatickém generování metadat a harmonizaci dat v systému Clindata¹⁶¹, spolupráci a součinnosti s dalšími repozitáři z TKA B/H/F – Multiomická lidská data (OmiCZ) a Repozitář pro lidská a animální obrazová data. Ve všech případech bude využita expertíza našeho týmu v oblasti etické problematiky biomedicínského výzkumu v rámci velkých výzkumných infrastruktur a etické problematiky genetického testování a genomických dat. Expertíza naší TKA pro spolupráci se soukromým sektorem bude využita především stran aplikovaného výzkumu se zaměřením na duševní vlastnictví. Ve spolupráci s repozitářem materiálových věd nastavíme procesy napomáhající tvorbě technologií a souvisejícího data managementu v praxi.

Výsledkem výše popsaných spoluprací bude první verze Metodiky, kterou budeme dále iterativně upravovat ve spolupráci s partnery projektu a testovat na poskytnutých datasetech.

V rámci přípravy pilotní verze metodiky budou stanovené postupy testovány na konkrétních datasetech oborových clusterů v rámci minikonsorcia TKA:

- materiálové vědy – technická data, testování bude zaměřeno na ověření metodiky stran kolaborativního výzkumu;
- bio/zdraví/potraviny – velká obrazová data a sekvenační data lidských subjektů, testování bude zaměřeno na eticko-právní část metodiky, protože se jedná o speciální kategorii citlivých dat;

¹⁵⁹ Dle <https://www.elastic.co/what-is/web-crawler> je data crawler volně přeloženo „digitální bot vyhledávače, který používá metadata k objevování a indexování stránek webu. Také se označuje jako pavoučí bot, protože „prolétá“ světový web (odtud „spider“ a „crawler“), aby zjistil, o čem daná stránka je. Poté stránky indexuje a ukládá informace pro budoucí vyhledávání“.

¹⁶⁰ <https://archiv.soc.cas.cz/cz/>.

¹⁶¹ Clindata je komplexní informační systém vyvinutý IMTM. Je navržen pro správu dat v klinických studiích, registrech a dalších klinických nebo vědeckých databázích. Systém se vyznačuje webovou architekturou klient/server, která uživatelům umožňuje přistupovat k němu prostřednictvím webových prohlížečů na počítačích, tabletech nebo chytrých telefonech. Veškerá komunikace mezi klientem a serverem je zabezpečena pomocí šifrování SSL, což zajišťuje bezpečnost dat. Clindata umožňuje sběr klinických a laboratorních dat v souladu s GDPR od pacientů zapojených do různých výzkumných projektů a klinických studií. Podporuje návrh studií, tvorbu elektronických formulářů případových zpráv (eCRF) a správu studijních akcí nebo návštěv. Systém také umožňuje ukládání velkých vědeckých datových souborů, včetně genomických a proteomických dat, připojením k infrastruktuře objektového úložiště IMTM. ClinData navíc usnadňuje FAIRifikaci (naležitelná, přístupná, interoperabilní, opakovaně použitelná) dat exportováním vybraných datových sad do veřejných úložišť. Tento systém bude v rámci TKA B/H/F implementován do NRP jako repozitář pro klinická data.

- sociální vědy – testování bude zaměřeno na ověření technik anonymizace dat při sdílení, které je potřeba nastavit specificky dle požadovaných dat a účelu, kvůli němuž jsou žádána.

Výsledná metodika bude pokrývat doporučené body (tzv. checklist), které by citlivý dataset měl splňovat před přijetím do repozitáře, aby odpovídal požadavkům kladeným na citlivá data a jejich sdílení. Připravíme i návody pro žadatele o datasety obsahující citlivá data, jak o data žádat a jak s nimi mohou nakládat v zabezpečeném prostředí pro zpracování citlivých dat nebo po jejich obdržení z repozitáře. Na konci projektu budou jasně definovaná doporučení a standardy pro zacházení s citlivými daty v NDI v podobě Metodiky pokrývající eticko-právní nakládání s citlivými daty a jejich kvalitu během jejich FAIRifikace. Tato zahrne doporučené postupy od přípravy dat pro nahrání do repozitáře s citlivými daty přes uchování v repozitáři až po podmínky a způsob sdílení citlivých dat různých kategorií. Tyto výstupy budou k dispozici nejen pro potřeby uživatelů NRP, ale i komunitám provozujícím repozitáře mimo NRP, aby byl v případě následné potřeby zajištěn potřebný soulad. Součástí této aktivity budou konzultace a odborné poradenství pro nově vznikající repozitáře, které uchovávají citlivá data. Budeme úzce spolupracovat s compliance týmem NRP a průřezovými aktivitami OS II compliance, kyberbezpečnost, provenance a jazykové modely. V současné době neexistuje souhrnná metodika zabývající se tímto tématem. **SE_2**

Pro systematičtější šíření know-how budou vytvořeny vzdělávací materiály navazující na Metodiku FAIRifikace citlivých dat v NDI. Materiál poskytne nutné znalosti pro FAIRifikaci a správu citlivých dat v NDI jako celku včetně vzorových dokumentů, návazností na oborové repozitáře, Národní metadatový adresář (NMA), právní rámce FAIRifikace citlivých dat, etické aspekty a další.

Obsah metodiky bude prezentován uživatelům NDI a všem, kteří by rádi vytvořili nový repozitář v NRP přes dedikované stránky NDI/NRP.¹⁶² **SE_2.1**

Součástí aktivity zaměřené na vytváření metodiky FAIRifikace citlivých dat jsou zahraniční cesty, které jsou plánované k načerpávání znalostí a zkušeností z jiných zemí, výzkumných infrastruktur a komunit ohledně správy citlivých (výzkumných) dat. Dedikovaní členové týmu, především ti, kteří budou vytvářet metodiky nebo služby/nástroje pro NDI, budou vysláni na zhruba dva několikadenní výjezdy (semináře, konference, kongresy atp.).

ČINNOSTI:

- analýza správné praxe v ČR i zahraničí pro sdílení výzkumných dat získaných z kolaborativního výzkumu se soukromou sférou a výzkumu oborových clustrů zapojených v NRP, vč. eticko-právní analýzy; **SE_1**
- příprava návrhu komplexní Metodiky FAIRifikace citlivých dat v rámci NDI experty realizačního týmu ve spolupráci se správci repozitářů obsahujících citlivá data;
- interní ověření/testování postupů Metodiky (průchod navrženého workflow) na vybraných datasetech;
- finalizace Metodiky na základě interakcí s repozitáři; **SE_2**
- vytvoření vzdělávacích materiálů navazujících na Metodiku FAIRifikace citlivých dat. **SE_2.1.**

¹⁶² Samotná školení budou realizována ve spolupráci se Školícím centrem EOSC CZ, které zajistí odpovídající kapacity, technické zázemí a organizační podporu a nejsou tak zahrnuta do rozpočtu OS II.

KÓDY VÝSTUPŮ PODAKTIVITY¹⁶³

SE_1	Analýza správné praxe pro sdílení citlivých dat z kolaborativního výzkumu
SE_2	Metodika FAIRifikace citlivých dat v NDI
SE_2.1	Vzdělávací materiály navazující na metodiku FAIRifikace citlivých dat

10.8.2. PODAKTIVITA 9.2 – VÝVOJ A PILOTNÍ ZAVEDENÍ SLUŽEB A NÁSTROJŮ PRO ROZVOJ NDI

V aktivitě se zaměříme na vývoj a zavádění nástrojů a služeb, které umožní implementaci bezpečného procesu příjmu a vydání citlivých dat do/z prostředí pro jejich zabezpečené zpracování, např. do/z úložišť a repozitářů v NDI; dále implementaci vhodného rozhraní pro uživatele pro práci s těmito daty v daném zabezpečeném prostředí, např. formou dedikovaných VPN, virtuálních desktopů (VD) pro interaktivní práci, rozhraním pro spouštění dávkových workflows (WF) apod. Účelem je poskytnout uživatelům NDI (výzkumníkům, výzkumným skupinám apod.) bezpečné prostředí, kde mohou zpracovávat citlivá data.

V rámci aktivity využijeme expertízy na Nařízení o evropském prostoru pro zdravotní údaje (EHDS)¹⁶⁴, resp. přípravu implementačních dokumentů tohoto nového nařízení, dále zkušenosti a podklady z GDI apod.¹⁶⁵ Do řešení zahrneme zkušenosti a know-how specialistů ohledně zpracování citlivých dat v řešení SensitiveCloud CERIT-SC¹⁶⁶. Zkušenosti části týmu ohledně správy citlivých (klinických) dat v Clindata platformě budou podstatné pro vývoj nástroje kontrolu kvality (klinických) dat. Vytvoří se také programové rozhraní API, což umožní přístup externích klientů ke QA/QC modulu Clindat.

Cílem aktivity je zabezpečit kontrolu a zajištění kvality dat v jednotlivých repozitářích založených především na systému Clindata tak, aby dataset splňoval standardy jako například ICD 10¹⁶⁷, LOINC¹⁶⁸, OMOP¹⁶⁹, ISO 8601 pro datum¹⁷⁰, případně dalších, dle potřeby. Současná forma systému Clindata¹⁷¹ je již používána v řadě nemocničních zařízení pro multicentrické klinické studie a lze je považovat za standardní nástroj pro sběr parametrického klinického záznamu. Úpravy systému Clindata, aby mohl sloužit jako NRP repozitář jsou řešena v TKA B/H/F.

Součástí této aktivity jsou také zahraniční cesty, které jsou nezbytné k načerpání zkušeností s nástroji a službami pro práci s citlivými výzkumnými daty na národní a mezinárodní scéně. Cílem cest budou zejména tematické konference, semináře, workshopy, případně krátké stáže v institucích vyvíjejících/provozujících nástroje a služby pro práci s citlivými výzkumnými daty. Výstupy této podaktivity budou i aktivně prezentovány min. na jedné mezinárodní akci.

¹⁶³ Kódy výstupů podaktivity jsou provázány na přílohu č. 3 Harmonogram klíčových aktivit a kap. 11 Hlavní plánované výstupy/produkty Studie proveditelnosti.

¹⁶⁴ Nařízení o evropském prostoru pro zdravotní údaje = European Health Data Space je nové nařízení EU platné od března 2025, více na https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space-regulation-ehds_cs.

¹⁶⁵ Členové týmu jsou zapojeni také v aktivitách VVI BBMRI-ERIC, EATRIS, ECRIN, EBRAINS, Bioimaging, ELIXIR, EOSC a dalších.

¹⁶⁶ Viz <https://www.cerit-sc.cz/infrastructure-services/sensitivecloud>.

¹⁶⁷ <https://icd.who.int/browse10/2019/en>.

¹⁶⁸ <https://loinc.org/>.

¹⁶⁹ <https://ohdsi.github.io/CommonDataModel/>.

¹⁷⁰ ISO 8601-1:2019/Amd 1:2022 - Date and time — Representations for information interchange — Part 1: Basic rules — Amendment 1: Technical corrections.

¹⁷¹ Clindata je komplexní informační systém určený pro správu dat v klinických studiích, registrech a dalších klinických nebo vědeckých databázích.

SW modul pro kontrolu a zajištění kvality dat repozitářů

Kvalita dat při nahrávání do repozitářů NRP není zatím automaticky ošetřena a kontrola zůstává pouze na úrovni ověření kurátorem repozitáře, což je z kapacitních důvodů i z důvodů lidské chybovosti dlouhodobě neudržitelné. Pro potřeby správců a kurátorů dat repozitářů NRP vyvineme QA/QC softwarový modul pro kontrolu kvality nahrávaných dat, s cílem eliminovat chybovost dat a zajistit kontrolu relevantnosti dat v repozitáři. QA/QC modul bude primárně vytvořen pro systém Clindata, protože expertíza QA specialisty a programátorů v týmu podílejícího se na vývoji modulu spočívá ve správě klinických dat. Využijeme tak expertízu z oboru klinické medicíny pro ostatní obory. Repozitáře založené na systému Clindata budou využívat QA/QC modul automaticky. Přístup do tohoto QA/QC modulu pro uživatele repozitáře bude realizován přes programové rozhraní API. Toto rozhraní umožní správcům jiných typů repozitářů implementaci QA/QC modulu do jejich prostředí.

Základní funkce QA/QC modulu

Navrhovaný QA/QC modul bude softwarový systém umožňující validaci nahraného datasetu podle nadefinovaných pravidel. Validační pravidla budou vycházet z existujících harmonizačních standardů (OMOP atd.) nebo ontologií, bude ale možné nadefinovat vlastní pravidla na základně expertního posouzení. Výstupem QA/QC modulu bude výsledný report se seznamem chyb a upozornění v analyzovaném datasetu. QA/QC modul bude využitelný v systému Clindata, ale také v jiných repozitářích díky implementaci programového rozhraní API. Výsledky validace budou nasdíleny kurátorovi repozitáře, do kterého bude modul naimplementován. Konkrétní využití funkcionality a nastavení validačních pravidel modulu bude na správci, resp. kurátorovi dat repozitáře dle oboru.

V rámci vývoje modulu budeme následovat standardní postupy agilního vývoje SW, tj. provedeme analýzu potřeb, rozdělíme vývoj nástroje do jednotlivých částí (tzv. user stories) a na tvorbu základních softwarových částí a programových rozhraní. Navrhujeme testovací scénáře pro ověření funkčnosti modulu, otestujeme i uživatelskou přívětivost grafického rozhraní, provedeme implementaci software iterativně. Základní architektura QA/QC modulu bude tvořena částmi:

- **Vstupní modul** ve formě programového rozhraní API nebo webového rozhraní. Umožní nahrání datasetu do prostředí QA/QC modulu a dále přístup k jednotlivým položkám vložených dat. Data budou importována v textové formě, předpokládáme formát csv, nebo podobný.
- **Analýzátor dat** umožní načtení datasetu do datové struktury QA/QC modulu a vytvoří propojení jednotlivých parametrů s harmonizačním standardem nebo ontologií.
- **Designer validačních pravidel** bude obsluhován přes webové rozhraní, kde uživatel nadefinuje validační pravidla pro jednotlivé parametry a vazby mezi parametry. Znovupoužitelná validační pravidla bude možné vytvořit pouze pro harmonizované parametry.
- Perzistence validačních pravidel bude zajištěna použitím SQL/NoSQL databáze.
- **Validační procedura** umožní spuštění validačních pravidel nad vloženými daty, výsledkem bude sada nálezů, ze kterých se vytvoří report.
- **Reportovací modul** vytvoří výstupní report z validační procedury. Report bude možné získat z webového rozhraní nebo přes programové rozhraní API.
- **Autorizační modul** umožní řízený přístup k webovému rozhraní. Přístup bude řízen na úrovni datasetů a validačních pravidel. Bude propojen s NRP AAI¹⁷².
- **Zabezpečení** bude integrální součástí programového rozhraní API, předpokládáme využití klientských certifikátů pro připojení k API.

¹⁷² NRP AAI = autentizační a autorizační infrastruktura NRP; využívá systém pro správu uživatel nazývaný Perun a slouží k řízení přístupu uživatelů k jednotlivým zdrojům NRP.

- **Logovací a auditovací část.**

Výše uvedená validační procedura bude pracovat s **validačními pravidly, která budou pracovat s daty i metadaty na několika úrovních:**

- Základní validace jednotlivého parametru – kontrola vyplněnosti meta/dat, formátu, rozsahu apod.
- Pokročilá validace jednotlivého parametru dle harmonizačního standardu – kontrola, zdali vyplněné hodnoty odpovídají definici, referenčnímu rozsahu apod.
- Pokročilá validace skupiny parametrů – kontrola vazeb mezi závislými parametry, jedná se o kontroly typu „správná posloupnost událostí“, „hodnota parametru vylučuje nějaký stav“, „podmíněná povinnost parametru“ apod.
- Plausibility – kontrola věrohodnosti dat bude záviset na expertním posouzení a nastavení váhy jednotlivých validačních pravidel.

Validační pravidla budou nastavitelná dle potřeb repozitáře pomocí designeru validačních pravidel. Výstupem validační procedury bude sada nálezů různých závažností, předpokládáme tyto základní úrovně: ok/informace/varování/chyba. Testování a pilotní běh modulu v reálném provozu bude provedeno ve spolupráci s TKA B/H/F a repozitářem klinických dat.

Součástí procesu vývoje QA/QC modulu bude i vytvoření produkčního prostředí pro deployment, zrealizujeme náhodné uživatelské (exploratory) testování, opravíme chyby a zabezpečíme chod **modulu v produkčním prostředí.**

Zabezpečení QA/QC modulu bude řešeno na několika úrovních.

- Zabezpečení přístupů k webovému rozhraní. Každý uživatel se bude muset přihlásit přes centrální autentizační službu napojenou na AAI NRP. Připojení k webovému rozhraní bude zabezpečeno šifrováním SSL a certifikátem.
- Zabezpečení připojení k programovému rozhraní API bude realizováno použitím klientského certifikátu. Toto ale ještě může být přehodnoceno.
- Zabezpečení datasetu. QA/QC modul nebude sloužit k dlouhodobému ukládání validovaných dat. Předpokládáme, že data budou po provedení validace, po nějaké nadefinované době, z dočasného uložště modulu uvolněna. Každý uživatel uvidí v rozhraní modulu pouze svoje datasety, nepředpokládáme sdílení dat mezi uživateli. Z tohoto důvodu nebude třeba řešit zabezpečení datasetů na úrovni jednotlivých záznamů. Do QA/QC modulu bude možné vložit i dataset s citlivými daty, bude zajištěna úroveň zabezpečení pro citlivá data.
- Vytvořená validační pravidla budou ukládána perzistentně a bude je možno sdílet mezi uživateli.

Součástí technické dokumentace QA/QC modulu bude i metodická příručka pro kurátory dat a praktický manuál s vzdělávacím materiálem. Vznikne i detailní popis programového rozhraní API pro vstup a výstup dat z QA/QC modulu umožňující implementaci tohoto modulu v jiných repozitářích určených pro zpracování parametrických dat. Pro QA/QC modul předpokládáme využití následujících **technologií a softwarové architektury:**

- QA/QC modul předpokládáme napsání v programovacím jazyce Java, verze 21 nebo pozdější.
- Jako základní softwarový framework pro backend předpokládáme použití Spring Boot.
- Při vývoji chceme klást důraz na modularitu zdrojového kódu založenou na programových rozhraních. Primárním vstupem a výstupem dat předpokládáme rozhraní REST API. Funkcionalitu

všech modulů předpokládáme ověřit sadou unit testů, které předpokládáme psát paralelně se zdrojovým kódem. Cílem je pokrytí co nejvíce zdrojového kódu testy.

- Testovacím pracovním rámcem (framework) předpokládáme, že bude JUnit.
- Persistenci dat (konfigurace, validační pravidla apod.) předpokládáme zajistit technologií jdbc, Hibernate apod.
- Pro ukládání dat předpokládáme využít databázi SQL.
- Zdrojový kód bude doplněn podrobnými komentáři na úrovni modulů a jednotlivých funkcí.
- Klientská část QA/QC modulu předpokládáme, že bude napsaná v Java scriptu, předpokládáme použití frameworku jako jQuery, Angular nebo React.
- Zdrojový kód předpokládáme uložit v distribuovaném verzovacím nástroji GIT a celý vývoj bude řízen principy průběžné integrace (Continuous Integration, CI), tj. vývoj bude rozdělen do malých celků (User stories/tasks v systému Jira), které budou uvolňovány postupně. Rozpadu celého QA/QC modulu do těchto malých dílků musí předcházet detailnější analýza a není součástí tohoto textu.
- Vznikne vývojové a testovací prostředí, kde bude možné vyvíjené moduly/části kódu a jejich funkce testovat. SE_3

ČINNOSTI:

- analýza vývoje QA/QC modulu a rozpad celku na jednotlivé části (user stories/tasks);
- vývoj softwarového QA/QC modulu po jednotlivých částech;
- pilotování modulu na repozitáři klinických dat (TKA B/H/F, repozitář ClinData);
- automatizované unit testy, kontrola a validace zdrojového kódu;
- tvorba programového rozhraní API pro umožnění přístupu externích klientů;
- vytvoření produkčního prostředí, deployment, náhodné uživatelské (exploratory) testování, oprava chyb, zabezpečení produkčního prostředí, optimalizace chodu;
- tvorba související dokumentace modulu. SE_3

Služby pro rozvoj NDI

Cílem této aktivity je účinné napojení existujících prostředí pro zpracovávání, ukládání, získávání citlivých dat (jako např. SensitiveCloud nebo Clindata) do/z NDI. Tato snaha zahrnuje možnosti technické implementace řízeného procesu příjmu/výdeje dat do/z takového prostředí pro zpracování citlivých dat, implementace vhodné silné ochrany dat při přenosu, inovativních rozhraní pro koncové uživatele pro interaktivní/dávkové zpracování citlivých dat v daném zabezpečeném prostředí. Zavedeme čtyři služby. V rámci této aktivity tedy budou rozvinuty existující nástroje a prostředí do pokročilejší podoby (např. virtuální desktopy), případně standardizované podoby integrující existující řešení (např. REMS).

1/ Služba řízeného zabezpečeného procesu příjmu/výdeje dat do/z prostředí pro zpracování citlivých dat

Tato služba nemá návaznost na konkrétní repozitář, proces může řídit příjem/výdej dat z/do libovolného zdroje, včetně repozitářů NDI. Také podporuje činnosti nezbytné pro FAIRifikaci dat v repozitářích, resp. zpracování citlivých dat. Dále podporuje implementaci služeb systémů typu Trusted Research Environment.

Tato služba bude implementovat zabezpečený řízený proces příjmu/výdeje dat do/z prostředí pro zpracování citlivých dat. Bude tedy např. umožňovat autentizovat poskytovatele/příjemce dat,

umožní definici procesu dat před příjmem/vydáním, např. pro ověření dostatečné kvality před příjmem, dostatečné agregace/anonymizace před vydáním apod. Tímto umožní implementovat bezpečné rozhraní mezi zabezpečeným a potenciálně rizikovým prostředím pro citlivá data.

Implementace bude vycházet z analýzy existujících doporučení, best practices, standardů, požadavků či existujících nástrojů pro příjem/výdeje dat do/z prostředí pro zpracování citlivých dat.

Výstup se stane standardní součástí některého z prostředí pro zpracování citlivých dat a bude v něm nadále udržován. V současnosti předpokládáme implementaci v prostředí SensitiveCloud¹⁷³. Od infrastruktury NDI (např. repozitáře) očekáváme poskytování a udržování standardizovaných rozhraní, která budou pro propojení využitelná.

Průběžně budou sbírány podklady a vytvářena technická zpráva, která vytvoří doporučení k řízenému zabezpečenému procesu příjmu/výdeje dat do/z prostředí pro zpracování citlivých dat. Společně tvoří technickou a metodickou dokumentaci, která umožní implementaci obdobných služeb i v jiných prostředích pro zpracování citlivých dat v kontextu NDI.

ČINNOSTI:

- analýza existujících požadavků, doporučení, nástrojů příjmu/výdeje dat do/ze zabezpečeného prostředí;
- implementace služby příjmu/výdeje dat do/ze zabezpečeného prostředí;
- dále nasazení a provoz služby příjmu/výdeje dat v prostředí SensitiveCloud;
- sběr podkladů a tvorba technické zprávy „Doporučení k řízenému zabezpečenému procesu příjmu/výdeje dat do/z prostředí pro zpracování citlivých dat“;
- finalizace služby a její dokumentace vazba na výstup. SE_4

2/ Rozhraní pro koncové uživatele pro interaktivní/dávkové zpracování citlivých dat v zabezpečeném prostředí

Tato služba umožňuje zpracování citlivých dat např. procesy pro kontrolu/zvýšení jejich kvality. Také podporuje činnosti nezbytné pro FAIRifikaci dat v repozitářích, resp. zpracování citlivých dat. Dále podporuje implementaci služeb systémů typu Trusted Research Environment.

V rámci implementace této služby budou dle analýzy uživatelských potřeb identifikována vhodná rozhraní pro interaktivní zpracování citlivých dat v zabezpečeném prostředí a tato budou implementována. Pro interaktivní zpracování zvažujeme např. implementaci formou izolovaných virtuálních desktopů či řádkových terminálů, u dávkového zpracování dat je zvažováno nasazení vhodného nástroje pro orchestraci procesů (např. Snakemake apod.). V současnosti předpokládáme nasazení v prostředí SensitiveCloud¹⁷⁴.

Průběžně budou sbírány podklady a vytvářena technická zpráva, jež společně utvoří technickou a metodickou dokumentaci, která umožní implementaci obdobných služeb i v jiných prostředích pro zpracování citlivých dat v kontextu NDI. Technická zpráva spolu s technickým řešením služby popsaným výše umožní implementaci obdobných služeb i v jiných prostředích pro zpracování citlivých dat.

ČINNOSTI:

- analýza uživatelských potřeb rozhraní pro zpracování citlivých dat;
- implementace služby uživatelského rozhraní pro interaktivní zpracování citlivých dat;

¹⁷³ Viz <https://www.cerit-sc.cz/infrastructure-services/sensitivecloud>.

¹⁷⁴ Viz <https://www.cerit-sc.cz/infrastructure-services/sensitivecloud>.

- implementace služby uživatelského rozhraní pro dávkové zpracování citlivých dat;
- nasazení a provoz služby uživatelského rozhraní pro interaktivní zpracování citlivých dat;
- nasazení a provoz služby uživatelského rozhraní pro dávkové zpracování citlivých dat;
- sběr podkladů a tvorba technické zprávy „Doporučení k rozhraním pro koncové uživatele pro interaktivní/dávkové zpracování citlivých dat v zabezpečeném prostředí“;
- finalizace služby a její dokumentace SE_5

3/ Služba pro autorizaci dávkového zpracování citlivých dat

Výstupem služby je nastavení a implementace mechanismu, který umožní autorizaci požadavků na dávkové zpracování citlivých dat v prostředí typu Trusted Research Environment. Každý požadavek bude obsahovat informaci o prováděné analýze a její autorizaci prostřednictvím infrastruktury AAI. Tato služba umožní provádět analýzy na citlivých datech bez nutnosti přímého přístupu uživatelů k datům samotným, což je nutný první krok pro řešení problému federovaných výpočtů a konzorciálního sdílení analýz z citlivých dat. Primárním cílem služby je omický repozitář (OmiCZ), který obsahuje citlivá genomická data, ale je kompatibilní a použitelná také pro jakýkoliv jiný repozitář citlivých dat, který potřebuje spouštět analýzy bez přímého přístupu uživatelů k samotným datům. Vznikne i technická zpráva, která bude technickou a metodickou dokumentací výstupu „Rozhraní pro koncové uživatele pro interaktivní/dávkové zpracování citlivých dat v zabezpečeném prostředí“ popsaného výše, jež umožní implementaci obdobných služeb i v jiných prostředích pro zpracování citlivých dat.

Výstup se stane standardní součástí některého z prostředí pro zpracování citlivých dat a bude v něm nadále udržován.

ČINNOSTI:

- sběr požadavků na autorizaci workflow, definice scénářů použití. Mapování typových analýz, návrh způsobu předávání informací o analýze a její autorizaci přes AAI;
- návrh architektury služby, návrh metodiky autorizace dávkového zpracování, integrace s AAI, definice formátů požadavků;
- implementace prototypu služby, testování na pilotních scénářích v prostředí OmiCZ a návaznosti na prostředí pro zpracování citlivých dat;
- finalizace služby, příprava technické a metodické dokumentace (technická zpráva). Předání služby do provozu SE_6

4/ Služba pro vyhodnocení analytických workflow a klasifikaci citlivosti jejich výstupů

Tato služba poskytne systematické vyhodnocení citlivosti výstupů analytických workflow. Služba bude implementována jako poloautomatický systém, který bude kombinovat předem definovaná pravidla, nástroje pro kontrolu struktury výstupů a expertní hodnocení (ve spolupráci s odborníky podílejícími se na vývoji repozitáře OmiCZ – konkrétně s kurátorem výzkumných dat, správcem repozitáře OmiCZ a DevOps specialistou pro analýzu citlivých dat). Součástí výstupu budou také ukázková a šablonová workflow, která budou validována z hlediska citlivosti výstupů a použitelná jako referenční řešení pro konsorcia či výzkumné týmy, které potřebují sdílet jen určité typy dat – např. agregovaná nebo znečitlivěná data. Služba tak poskytne klíčový podklad pro nastavení přístupových pravidel a bezpečné sdílení dat v rámci i mezi výzkumnými infrastrukturami. Primárním cílem služby je omický repozitář (OmiCZ), jenž obsahuje citlivá genomická data, ale je kompatibilní a použitelná také pro jakýkoliv jiný repozitář citlivých dat, který potřebuje spouštět analýzy bez přímého přístupu uživatelů k samotným datům.

Bude vytvořena technická zpráva představující metodickou a technickou dokumentaci ke službě „Služba pro vyhodnocení analytických workflow a klasifikaci citlivosti jejich výstupů“. Zaměřuje se na problematiku stanovení míry citlivosti výstupních dat z analytických procesů, zejména z pohledu souladu s regulacemi GDPR a možnosti jejich bezpečného sdílení. Součástí zprávy je popis poloautomatického systému, který kombinuje předdefinovaná pravidla, kontrolu struktury a typologie výstupů s expertním vstupem. Zpráva rovněž obsahuje sadu ukázkových a šablonových workflow, která byla validována z hlediska citlivosti výstupů a jsou určena jako referenční řešení pro konsorcia a týmy plánující bezpečné sdílení agregovaných nebo anonymizovaných dat. Cílem dokumentace je umožnit přenositelnost metodiky do dalších repozitářů nebo prostředí pro zpracování citlivých dat v rámci NDI a podpořit jednotný přístup ke klasifikaci výstupní datové citlivosti.

ČINNOSTI:

- návrh metodiky hodnocení citlivosti výstupů, analýza požadavků komunit, definice pravidel citlivosti a typových scénářů;
- návrh architektury poloautomatického systému pro hodnocení citlivosti, návrh ukázkových a šablonových workflow;
- implementace systému pro hodnocení citlivosti, validace a testování na reálných scénářích v OmiCZ repozitáře/ prostředí pro zpracování citlivých dat;
- finalizace služby, příprava technické a metodické dokumentace (technická zpráva), zajištění přenositelnosti do dalších prostředí NDI. SE_7

KÓDY VÝSTUPŮ PODAKTIVITY¹⁷⁵

SE_3	SW modul pro kontrolu a zajištění kvality dat repozitářů vč. metodické příručky QA/QC dat v repozitářích
SE_4	Služba řízeného zabezpečeného procesu příjmu/výdeje dat do/z prostředí pro zpracování citlivých dat
SE_5	Rozhraní pro koncové uživatele pro interaktivní/dávkové zpracování citlivých dat v zabezpečeném prostředí
SE_6	Služba pro autorizaci dávkového zpracování citlivých dat
SE_7	Služba pro vyhodnocení analytických workflow a klasifikaci citlivosti jejich výstupů

10.9. PRŮŘEZOVÁ TÉMATA PROJEKTU

10.9.4. PODAKTIVITA 10.4 – ELEKTRONICKÉ LABORATORNÍ DENÍKY PRO NRP¹⁷⁶

Elektronické laboratorní deníky (electronic laboratory notebooks, dále také jako ELN) jsou jedním z několika druhů softwarových nástrojů, které používají vědci ke správě, sdílení a přípravě ke zveřejnění vědeckých dat. Do této kategorie spadají i nástroje označované jako elektronické terénní deníky¹⁷⁷, které se zaměřují na obory, jejich součástí je v práce terénu (např. archeologie, ekologie, terénní biologie).

Kvalitní ELN umožňuje vědcům v uživatelsky příjemném prostředí spárovat získaná data,

¹⁷⁵ Kódy výstupů podaktivity jsou provázány na přílohu č. 3 Harmonogram klíčových aktivit a kap. 11 Hlavní plánované výstupy/produkty Studie proveditelnosti.

¹⁷⁶ V popisu KA jsou plánované aktivity propojeny na realizované výstupy. Výstupy jsou označeny kódem v podobě ELN_x.

¹⁷⁷ Jedná se o identický typ SW nástroje jako jsou Elektronické laboratorní deníky.

např. z experimentu nebo výpočtu, s jejich popisem (metadata), bezpečně je uložit, prohledávat, řídit jejich sdílení s jinými vědci a kontrolovat jejich životní cyklus. V mnohých případech ELN umožňují sběr dat z komplexních měření nebo složitých výzkumných procesů. Výsledkem mohou být plně anotovaná a propojená data.

ELN se zaměřují na práci s *živými daty (hot and warm data)*¹⁷⁸. Mají za cíl zajištění jejich kvality (včetně naplňování FAIR principů) od okamžiku vzniku, přes zaznamenávání procesu jejich zpracování a obohacování o další data nebo metadata, až po verzování v případě opakovaných nebo doplňovaných výstupů.

Hlavní funkcí ELN je pokročilá práce s metadata. Popis dat (metadata) je možné rozvinout do libovolného detailu specifického pro danou vědeckou disciplínu nebo daný projekt. Zároveň je tento proces v ELN zásadně zjednodušen pomocí formulářů a předvyplněných údajů z důvodu zaznamenávání informace, která se opakuje. ELN umožňuje sjednocovat postupy zpracování dat tak, aby byla data a metadata konzistentní napříč různými experimenty a různými pracovníky spolupracujícími na výzkumu, včetně harmonizace napříč institucemi nebo mezinárodními partnery. V některých případech ELN umožňují vizualizaci ukládaných dat nebo doplňování záznamů o schémata, včetně komplexních chemických nebo matematických rovnic pomocí integrovaných nástrojů přímo v ELN. Z pohledu legislativy některá ELN umožňují právně validní identifikaci původu dat (signatures, witnessing, timestamping).

Další důležitou nativní vlastností ELN je zaznamenávání historie dat, jejich vývoje a změn (provenance). To přirozeně navazuje na podaktivitu 10.5, která se provenance zabývá. Synergie téhle vazby spočívá v tom, že podaktivita 10.5 svou metodikou poskytne obecný návod, jak správně používat ELN, aby uživatel respektoval právní rámec provenance. Softwarové nástroje vyvíjené v PKA Provenance zároveň nabízí alternativu managementu provenance v situacích, ve kterých nebudou použity ELN – jsou komplementární. Pro účely NRP/NDI je ELN doporučený nástroj pro konzistentní a uživatelsky příjemný sběr a popis dat, včetně jejich exportu do datových repozitářů.

- Mezinárodně jsou ELN využívány ve vědeckých oblastech, které odpovídajících několika TKA projektu OS II (např. HUMAN, Fyzika, MATECH, ENVIRO nebo Bio/Health/Food) Jsou považovány za standardní nástroj na správu živých vědeckých dat a jejich přípravu pro publikování v důvěryhodných repozitářích FAIR dat. Kvůli kompatibilitě českých vědeckých skupin s mezinárodním prostředím a kvůli jejich konkurenceschopnosti je potřeba mít ELN v portfoliu SW nástrojů NDI podporovaných na národní úrovni.
- V oborech MATECH, Fyzika a Archeologie (součást HUMANITIES) byly ELN identifikovány jako kritické nástroje pro tvorbu FAIR dat a jejich přípravu pro publikování v důvěryhodných repozitářích NRP/NDI. Dá se předpokládat, že u části podoborů clustrů Bio/Health/Food a ENVIRO budou tyto nástroje v nejbližší době také zaváděny pro správu vědeckých dat z důvodu výše uvedené konkurenceschopnosti na mezinárodní úrovni a potřeby zpracovávat komplexní datové sady (viz níže).
- Napříč obory také prudce roste množství vědeckých dat a souvisejících metadat. Rozvíjející se technologie pro zpracování dat, jako například strojové učení, pak umožňují čím dál tím komplexnější schémata práce s daty. To klade rostoucí nároky na správu dat, například z důvodu

¹⁷⁸ Na rozdíl od datových repozitářů ve smyslu projektu NRP. Projekt NRP nenabízí nástroje pro pokročilou správu *živých dat*, které by byly přímo nasaditelné ve vědeckých organizacích. Projekt NRP se zaměřuje na budování infrastruktury. ELN se využívají převážně v technických a přírodovědných disciplínách a některé aspekty ELN (např. metadataové formuláře) jsou oborově specifické.

registrace detailních metadat, provenance nebo standardizace napříč vědeckými disciplínami. ELN je typem nástroje, který dokáže tyhle stoupající nároky plnit a zjednodušovat.

Nasazení ELN v národním měřítku v prostředí NDI si vyžaduje, aby již byly splněny některé cíle projektu NRP (AAI, datové repozitáře, datová úložiště atd.). Implementace této oblasti úzce souvisí s existencí oborově-specifických repozitářů a je nezbytné, aby s nimi byla časově provázaná. V projektu je tedy provázáno takto:

- Pro správnou integraci ELN s catch-all datovým repozitářem NDI je nezbytné, aby měl řešitelský tým k dispozici metadatový model a související publikační workflows. To jsou výstupy projektu OS I. Proto je dokončení výstupu integrace ELN s repozitáři naplánované až po plánovaném dokončení metadatového modelu v OS I.
- Stejně je pro integraci ELN s catch-all repozitářem NDI¹⁷⁹ nezbytné, aby byla v provozu produkční instance samotného repozitáře. Ten specifikuje SW systém, pro který bude nutné vytvořit integrační nástroje v ELN a na vstupu do repozitáře. Proto je dokončení integrace ELN s catch-all repozitářem NDI naplánované až po plánovaném nasazení produkční instance repozitáře.

Na základě důkladného průzkumu různých ELN, jež jsou v současnosti dostupné, byly zvoleny dva, které budou podporované v rámci podaktivity: Kadi4Mat a eLabFTW. Tyto dva deníky nejlíp splňují několik důležitých kritérií zásadních pro dlouhodobé nasazení v NDI:

- Za jejich vývojem stojí stabilní tým zkušených vývojářů. Pravděpodobnost přerušení vývoje v časovém horizontu udržitelnosti projektu je velmi malá.
- Vývoj těchto produktů je finančně podporovaný velkými výzkumnými organizacemi (KIT Německo, CNRS Francie), kde mají taky širokou uživatelskou základnu.
- Oba elektronické laboratorní deníky jsou vyvíjeny jako otevřený software a v budoucnosti mají potenciál podporovat FAIR data.
- Oba zvolené elektronické deníky mají jednoduché intuitivní uživatelské rozhraní, malou režii administrace a žádné jejich specifické vlastnosti je příliš těsně nevážou na kteroukoliv vědeckou oblast.
- Zvolené ELN jsou postavené na standardních softwarových technologiích, u kterých je vysoká pravděpodobnost podpory v časovém horizontu udržitelnosti projektu. Oba deníky mají dobrou strukturu REST API, která umožňuje jejich integraci s jinými systémy v rámci výzkumných organizací, kde budou nasazeny, jako i v NDI.

V rámci podaktivity budou zajištěny SW služby pro integraci podporovaných ELN s repozitáři NRP v rámci NDI. Cílem je, aby vědci měli jednoduchou a uživatelsky příjemnou možnost publikovat záznamy z ELN do datových repozitářů přímo v prostředí ELN bez nutnosti opětovného vyplňování informací nebo ručního exportu/importu dat a metadat.

V rámci podaktivity se zajistí nasazení národních instancí obou podporovaných ELN (Kadi4Mat i eLabFTW), které budou určeny pro organizace, které nebudou mít vlastní instance ELN. Z přímých prostředků PKA ELN zakoupí partner server, na kterém se budou tyhle instance provozovat a zajistí příslušnou softwarovou podporu.

V rámci podaktivity se vypracuje metodika doporučeného nasazení a provozu podporovaných ELN

¹⁷⁹ Viz repo.cz.

v organizacích, které si budou provozovat své vlastní instance. Definuje minimální požadavky na hardware, software, infrastrukturu a lidské zdroje, vypracuje dokumentaci doporučeného postupu nasazení podporovaných ELN s návazností na NDI/NRP.

V rámci podaktivity se zajistí vývoj nových vlastností ELN Kadi4Mat, které dále rozšíří jeho dlouhodobé uplatnění v NDI. Hlavní důraz bude kladený na implementaci komunikace mezi různými instancemi ELN Kadi4Mat, která umožní interaktivně sdílet data mezi různými organizacemi a vědeckými skupinami (včetně mezinárodních spoluprací). Tahle vlastnost ELN Kadi4Mat představuje častý požadavek od budoucích uživatelů tohoto ELN a je o ní v českém výzkumném prostředí velký zájem. Vzhledem k tomu, že ELN Kadi4Mat momentálně přímo nepodporuje model nasazení v prostředí kontejnerové platformy, kterou nabízí NDI, podnikne partner kroky k tomu, aby ke konci projektu bylo možné nasazovat instance Kadi4Mat v prostředí Kubernetes v rámci NDI.

V rámci téhle podaktivity bude realizační tým podporovat dva specifické ELN vyvíjené vědeckou komunitou: Kadi4Mat a eLabFTW.

První zmíněný deník (Kadi4Mat) je velmi obecný a je vhodný pro širokou skupinu vědeckých směrů. Lze jej efektivně rozšiřovat pomocí vlastních modulů, jimiž lze zajistit pokročilou specifickou funkcionalitu, kterou může požadovat konkrétní vědecký tým. Kadi4Mat má vynikající předpoklady pro integraci s vědeckými workflows a pro využití se systémy pro automatický sběr vědeckých dat z měřících zařízení¹⁸⁰.

Druhý podporovaný deník (eLabFTW) je specifitější a svou architekturou víc odpovídá zaznamenávání vývoje vědeckého experimentu v čase. Jeho nasazení je jednodušší než nasazení Kadi4Mat, má specifické funkce relevantní ve vědách blízkých chemii a molekulární biologii.

Společná podpora obou deníků, které mají vzájemně se doplňující vlastnosti, má za cíl pokrýt co nejširší skupinu uživatelů s různorodými nároky.

V rámci podaktivity bude implementovaná služba, která uživateli umožní publikovat data a metadata ze záznamů v ELN Kadi4Mat do catch-all repozitáře NDI a do různých oborových datových repozitářů v NDI. Služba bude implementovaná jak v uživatelském rozhraní, tak na backendu.

Služba se bude skládat ze dvou částí: Jedna bude komponentem Kadi4Mat (plugin), druhá bude nezávislou mikroslužbou mezi ELN a repozitářem.

Funkčnost výstupu bude možné ověřit přístupem do národní instance ELN Kadi4Mat, kde bude jasně viditelná v uživatelském prostředí. Zároveň bude k dispozici veřejný Git repozitář s příslušnými zdrojovými kódy komponentu a historií jeho vývoje.

Zrealizujeme integraci generického catch-all repozitáře `datarepo.eosc.cz` v NDI s národní instancí ELN Kadi4Mat. V rámci jednotlivých TKA bude služba k dispozici a umožněna integrace jednotlivých jejích instancí ELN Kadi4Mat s oborově specifickými repozitáři NRP v NDI. `ELN_1`

Dále bude implementovaná služba, která uživateli umožní publikovat data a metadata ze záznamů v ELN eLabFTW do generického catch-all repozitáře a do různých oborových datových repozitářů a NRP v NDI. Služba bude implementovaná jak v uživatelském rozhraní, tak na backendu. Služba bude využívat standardizované API rozhraní aplikace a repozitářů. To usnadní udržitelnost při minimálních finančních nákladech. Služba se bude skládat ze dvou částí. Jedna bude implementovaná jako komponent v ELN eLabFTW. Ta se outsourcuje u vývojového týmu eLabFTW. Druhá část služby se realizuje jako nezávislá mikroslužba mezi ELN a repozitářem. Funkčnost výstupu bude možné ověřit přístupem do národní

¹⁸⁰ Ta jsou vyvíjena v rámci KA5.3 projektu NRP.

instance ELN eLabFTW, kde bude jasně viditelný v uživatelském prostředí. Zároveň bude k dispozici veřejný Git repozitář s příslušnými zdrojovými kódy komponentu a historií jeho vývoje.

Vývoj té části služby integrace s NRP, která je komponentem eLabFTW, půjdeme zadat vývojovému týmu Delatlot (Francie, NPO) – autorům ELN eLabFTW. Důvodem jsou použité softwarové technologie, na kterých je postavené uživatelské rozhraní eLabFTW. V týmu podaktivity nemáme s těmito technologiemi zkušenosti (PHP, specifická architektura softwarového produktu). Proto bude časově a ekonomicky výhodnější outsourcovat vývoj komponentu autorům eLabFTW. Autoři eLabFTW takovou službu nabízejí a zároveň jsou kompetentní tenhle úkol zvládnout s nižšími náklady, protože detailně znají interní architekturu ELN eLabFTW. Druhá část vývoje služby (samostatná mikroslužba) bude vyvinutá v týmu PKA.

Zrealizujeme integraci catch-all repozitáře v NDI s národní instancí ELN eLabFTW. V rámci jednotlivých TKA bude služba k dispozici – umožněna integrace jednotlivých jejích instancí ELN eLabFTW s oborově specifickými repozitáři NRP v NDI. **ELN_2**

Bude vypracovaná metodika nasazení ELN Kadi4Mat a eLabFTW ve vědeckých organizacích, které si vytvoří a budou administrovat svoji vlastní instanci za použití prostředků NDI. Konkrétně budou specifikované minimální hardwarové a softwarové nároky vzhledem k velikosti organizace a množství dat. Dále budou vytvořeny návody pro integraci ELN s nástroji NDI (AAI, datová úložiště, kontejnerová platforma na poskytování virtuálních služeb) a návody pro škálování nasazení ELN. Součástí výstupu budou také návody pro organizaci šablon záznamů, aby byly kompatibilní s metadatovými modely příslušných repozitářů NRP v NDI a modelová šablona pro záznam metadat pro generický catch-all repozitář (CCMM). Výstup se bude realizovat z úvazku v podaktivitě. **ELN_3**

Bude nasazena jedna instance ELN Kadi4Mat a jedna instance ELN eLabFTW pro organizace, které nemají technické možnosti nebo zájem zřídit si vlastní instance podporovaných ELN. Částečně budou použité části infrastruktury NRP v NDI (AAI, kontejnerová platforma, datová úložiště S3), částečně bude použitý virtualizační server zakoupený z přímých prostředků podaktivity. Důvodem je, že minimálně ELN Kadi4Mat zatím není kompatibilní s kontejnerovou platformou Kubernetes¹⁸¹ dostupnou v rámci NDI. Nasazení a administrace instancí budou financovány z osobních nákladů podaktivity. **ELN_4**

Ve spolupráci s vývojovým týmem na IAM KIT,¹⁸² Karlsruhe, Německo budou implementované nové vlastnosti ELN Kadi4Mat, které jsou relevantní v národním a Evropském vědeckém prostředí. Konkrétně jde o propojení různých instancí ELN Kadi4Mat. To umožní uživatelům po autorizaci interaktivně sdílet data mezi instancemi ve smyslu vzájemného přímého přístupu bez exportu a importu. To je důležité nejen pro velká data, ale i pro záznamy, u kterých je přenos mezi různými instancemi ELN nepraktický a je potřebné tyto úkony zaznamenávat v metadatech záznamů (provenance). Umožní to taky vzdálenou spolupráci různých týmů a různých instancí ELN na zpracování stejných dat. Tento přístup je ve světě ELN velice inovativní, protože tato komplexní vlastnost není typická ani pro jiné ELN.

Dalším prvkem implementovaným v rámci projektu bude systém notifikací, který uživatele interaktivně upozorní na události v ELN a umožní mu reagovat na ně. Práce na tomto výstupu bude zahrnovat také

¹⁸¹ Většina softwarových komponent projektu OS II se bude provozovat na hardwarových prostředcích NDI. Ta zpřístupňuje kontejnerovou platformu Kubernetes – Rancher (<https://www.cerit-sc.cz/infrastructure-services/data-processing/container-platform>). Některé softwarové komponenty plánované pro aktivitu nejsou s touto platformou kompatibilní a bude nezbytné provozovat je mimo kontejnerovou platformu. Proto plánujeme nákup jednoho serveru s hardwarovými parametry vhodnými k provozu několika virtuálních serverů s těmito softwarovými komponenty. Jednorázové náklady na nákup serveru a náklady na jeho provoz jsou za období trvání projektu nižší, než by byly celkové náklady na typické poplatky za pronájem virtuálních serverů u externích poskytovatelů.

¹⁸² Institute of Advanced Materials, Karlsruher Institute of Technology.

přizpůsobení aplikace pro snadnější rutinní nasazení instance ELN v prostředí kontejnerové platformy Kubernetes.¹⁸³ To dále usnadní adopci ELN v dalších výzkumných organizacích na prostředcích NDI.

Tento výstup předpokládá, že vývojář SW absolvuje každoročně krátký pobyt (5 dní) v týmu vývojářů Kadi4Mat v Karlsruhe, Německo. Důvodem je synchronizace aktivit a seznámení se s dalším koncepčním směřováním systému. Kadi se skládá z několika komponent (Kadi4Mat ELN, Kadi Workflows, Kadi FS, Kadi AI atd.), jejichž filozofie a koncept použití se průběžně vyvíjí. Proto je důležité, abychom byli v průběžném kontaktu a tento vývoj sledovali. Umožní to také komunikovat potřeby cílové skupiny¹⁸⁴ směrem k vývojářům a efektivněji je implementovat. Proto se v rámci PKA plánuje každoročně jedna pětidenní služební cesta pro jednu osobu do Karlsruhe, Německo, pro osobní intenzivní spolupráci s vývojovým týmem Kadi4Mat. Po zbylou část roku bude SW vývojář spolupracovat s vývojovým týmem Kadi4Mat vzdáleně. **ELN_5**

ČINNOSTI:

- integrace ELN Kadi4Mat s NRP; **ELN_1**
- integrace ELN eLabFTW s NRP; **ELN_2**
- metodika nasazení ELN Kadi4Mat a eLabFTW; **ELN_3**
- nasazení národních instancí ELN; **ELN_4**
- vývoj nových vlastností ELN Kadi4Mat. **ELN_5**

KÓDY VÝSTUPŮ PODAKTIVITY¹⁸⁵

ELN_1	Integrace ELN Kadi4Mat s NRP
ELN_2	Integrace ELN eLabFTW s NRP
ELN_3	Metodika nasazení ELN Kadi4Mat a eLabFTW (WIKI chat-ready dok)
ELN_4	Nasazení „catch-all“ instancí ELN (1 instance ELN Kadi4Mat, 1 instance ELN eLabFTW)
ELN_5	Inovovaný ELN Kadi4Mat

10.9.5. PODAKTIVITA 10.5 – ZPĚTNÁ DOHLEDATELNOST OBJEKTŮ POMOCÍ PROVENANCE¹⁸⁶

Výstupy vyvinuté v rámci této podaktivity budou umožňovat harmonizovanou práci s důvěryhodnou provenance¹⁸⁷ za účelem podpory zpětné dohledatelnosti (traceability) předchůdců objektů uložených v repozitářích (resp. pocházejících z laboratorních deníků nebo jiného SW) i mimo nich. Provenance bude reprezentována a spravována v souladu se současnými mezinárodními standardy, zejména W3C¹⁸⁸ PROV a CPM¹⁸⁹/ISO¹⁹⁰ 23494, a to za účelem harmonizace a zajištění interoperability s mezinárodními aktivitami a současným vývojem na evropské (např. projekty EvolveBBMRI nebo

¹⁸³ <https://www.cerit-sc.cz/infrastructure-services/data-processing/container-platform>.

¹⁸⁴ Požadavky uživatelů na nové vlastnosti ELN Kadi4Mat se budou sbírat prostřednictvím web nástroje Gitlabu, kde se soustřeďuje vývoj Kadi4Mat. GitLab poskytuje koncovým uživatelům a administrátorům instancí zadávat požadavky na další vývoj. Odkaz na tenhle formulář Gitlabu bude přímo na úvodních stránkách instancí Kadi4Mat.

¹⁸⁵ Kódy výstupů podaktivity jsou provázány na přílohu č. 3 Harmonogram klíčových aktivit a kap. 11 Hlavní plánované výstupy/produkty Studie proveditelnosti.

¹⁸⁶ V popisu KA jsou plánované aktivity propojeny na realizované výstupy. Výstupy jsou označeny kódem v podobě **P_x**.

¹⁸⁷ Informace dokumentující historii popisovaného objektu (např. datasetů) a souvisejících činností, která obsahuje informace o původu nebo zdroji popisovaného objektu, o jakýchkoli změnách, které od jeho vzniku nastaly, a kdo jej měl od jeho vzniku v péči.

¹⁸⁸ World Wide Web Consortium.

¹⁸⁹ Common Provenance Model.

¹⁹⁰ International Organization for Standardization.

UNCAN-Connect) a globální úrovni.

Provenance je nedílnou součástí naplňování Zuvupoužitelnosti (R ve FAIR, „R1.2: *(Meta) data are associated with detailed provenance*“¹⁹¹) a týká se všech datových sad. V aktivitě usilujeme o vývoj a následné pilotní zprovoznění nástrojů a služeb umožňujících zpětnou dohledatelnost objektů (*traceability*) primárně pro podporu ohodnocení kvality (vhodnosti pro použití) dat uložených v repozitářích a podporu interoperability v mezinárodním kontextu. Podaktivita Provenance připraví materiály a SW nástroje pro standardizaci uchování informace o provenance v souladu s ISO normou.¹⁹²

K řešení problematiky Provenance v rámci NRP je nyní možné přistoupit i proto, že podkladový datový model (CPM) pro reprezentaci provenance je již dostatečně zralý pro implementaci v rámci NDI. V současnosti je podkladový datový model stabilní a současně už bylo implementováno několik prototypů pro využití datového modelu a standardu, tj. aktuální vývoj umožňuje problematiku řešit.

K přenášení datových sad a jejich propojování mezi repozitáři je zapotřebí:

- existovala metodika práce s provenance v rámci Národní repozitářové platformy za účelem její harmonizace;
- byl implementován SW, který umožní harmonizovanou práci s provenance (v tomto případě dle ISO 23494) v rámci NRP a zajistí kompatibilitu s mezinárodními aktivitami.

Tyto dva body zajistíme v rámci výstupů podaktivity Provenance a zajistíme tím podporu pro jejich primární uživatele, kterými jsou vývojáři, metodici a řídicí pracovníci repozitářů. Ti mohou integrovat implementovaný SW nebo se řídit metodikou pro správu provenance.

Výstupy Provenance budou využity i v dalších KA projektu, a to zejména v těchto ohledech:

1. jednotný přístup k práci s provenance v rámci NRP, což umožní využití standardních nástrojů pro její generování, ukládání a přístup;
2. zarovnání s mezinárodními standardy pro práci s provenance, s cílem harmonizace s vývojem na evropské nebo globální úrovni;
3. využití současného state-of-the-art v oblasti provenance, což může vytvořit zázemí pro budoucí výzkum v oblasti.

Konkrétně budou výstupy využité minimálně v následujících klíčových aktivitách:

V TKA SENSI, kde bude vznikat metodika práce s metadaty (licence a eticko-právní). Jelikož lze provenance vnímat jako specifickou podmnožinu metadat, tak by tato metodika měla být v souladu s metodikou pro práci s provenance. V TKA DM4AI bude implementován nový repozitář, který využije SW nástroje pro správu dat vytvářených v rámci podaktivity Provenance. V TKA MATECH a B/H/F dojde k využití metodických podkladů podaktivity Provenance s cílem integrovat nový oborový repozitář (DANTE^c, resp. repozitář pro Multiomická Lidská Data) se současnými standardy pro provenance. V PKA oblast ELN bude integrován nový SW, jehož součástí je implementace provenance v souladu s metodikou pro práci s provenance.

V rámci podaktivity navrhne a vyvine architekturu systému pro ukládání a správu provenance. Tento SW bude umožňovat: 1/ ukládání, přístup a verzování provenance dokumentů¹⁹³, 2/ generování

¹⁹¹ <https://www.go-fair.org/fair-principles/r1-2-metadata-associated-detailed-provenance/>.

¹⁹² Různé obory řeší provenance různým způsobem. Např. ELN jsou příkladem, které si provenance uchovávají, ale ne ve standardizované podobě.

¹⁹³ Provenance bude technicky uložena ve standardizovaném strojově zpracovatelném dokumentu (např. PROV-JSON).

metadat pro provenance dokumentů dle uvedených standardů, 3/ uložené provenance dokumenty a související metadata budou za účelem jejich dlouhodobého uchování identifikovatelné pomocí perzistentních identifikátorů. Tyto persistentní identifikátory budou uloženy jako součást metadatového záznamu a po autorizaci se budou resolvovat na obsah uložených provenance dokumentů.

Implementovaný systém bude dále poskytovat API pro komunikaci se SW, bude nasaditelný ve více instancích a integrovatelný s repozitářovými systémy tak, aby bylo možné vyhledávat předchůdce objektů uložených v repozitářích (tzv. backward traceability) a případně i mimo ně. Samotná integrace systému je obsahem TKA (viz. výše v textu). Systém dále umožní řízení přístupu na úrovni celého prováděného dokumentu a bude umožňovat kontrolu integrity a autenticity uložených provenance dokumentů pomocí hašů a digitálních podpisů. **P_1**

V rámci implementace systému pro ukládání a správu harmonizované provenance bude využita SW knihovna, která bude implementovaná samostatně z důvodu její využitelnosti i v jiných systémech. SW knihovna umožní:

- reprezentaci struktur definovaných v CPM;
- generování provenance dokumentů pomocí šablon (tzv. templating mechanism) v souladu s CPM;
- základní syntaktickou a sémantickou validaci provenance dokumentů, tzn. ověření, že provenance poskytnutá v provenance dokumentu je reprezentována v souladu se standardy;
- vyhledávání předchůdců objektů uložených v repozitářích. **P_2**

Implementovaný SW bude doplněn dokumentem obsahujícím seznam doporučení týkajících se správy harmonizované provenance v NDI tak, aby byla správa provenance v souladu se současnou a dobrou praxí a dostupnými standardy souvisejícími s AI Act. Metodický dokument vzniká za účelem zjednotit integrace SW s repozitářovými systémy a s cílem harmonizace práce s provenance v rámci NDI. **P_3**

V rámci udržení přehledu a mezinárodních kontaktů plánujeme cestu na konferenci (primárně Provenance Week). Vzhledem k potenciálnímu mezinárodnímu uplatnění výstupů PKA provenance plánujeme výstupy na mezinárodní konferenci také prezentovat (primárně Extended Semantic Web Conference nebo Provenance Week).

ČINNOSTI:

- vývoj systému pro ukládání a správu provenance; **P_1**
- vývoj SW knihovny pro práci s provenance dle CPM; **P_2**
- návrh, pilotní ověřování a finální verze metodiky pro provenance; **P_3**
- implementace prototypů SW nástrojů 2027-03;
- návrh metodiky provenance 2027-06.

KÓDY VÝSTUPŮ PODAKTIVITY¹⁹⁴

P_1	SW pro ukládání a správu provenance
P_2	SW knihovna pro práci s provenance dle CPM
P_3	Metodika pro provenance

¹⁹⁴ Kódy výstupů podaktivity jsou provázány na přílohu č. 3 Harmonogram klíčových aktivit a kap. 11 Hlavní plánované výstupy/produkty Studie proveditelnosti.

11. HLAVNÍ PLÁNOVANÉ VÝSTUPY

#ID	Název výstupu	Typ výstupu	Vazba na KA	Vazba na produkt v indikátoru 215 012 / 215 102	Stručný popis výstupu	Způsob doložení výstupu	Vazba výstupu na výstup jiného projektu OP JAK	Datum dosažení výstupu ¹⁹⁵
B_1	Národní omický repozitář Czech Omics Node (OmiCZ)	nový repozitář	KA2	Sada repozitářů pro TKA B/H/F	Repozitář pro bezpečné ukládání a správu citlivých genomických dat, obsahující: metadatový model, integrovanou AAI, rozhraní ProducerUI pro správu dat a metadat, systém pro bezpečné spouštění workflow a Galaxy v TRE, propojení na FEGA, Beacon a ClinData. Vč. dokumentace. VIZ TAKÉ DÍLČÍ VÝSTUPY B_1.1–B_1.6 KA 2.1.	webový odkaz na repozitář, vč. technické dokumentace a uživatelské příručky	ANO – návaznost na výstup/y NRP	XII.28
B_2	Repozitář strukturní a simulační data (BioSimCZ)	nový repozitář	KA2	Sada repozitářů pro TKA B/H/F	Repozitář zaměřený na strukturní simulační data, který bude využívat metadatový model pro simulační strukturní data a rovněž metadatové modely pro vybrané aspekty biomakromolekul, SW nástroje pro rozšíření funkcí repozitáře a dokumentace. VIZ TAKÉ DÍLČÍ VÝSTUPY B_2.1–B_2.5 KA 2.2.	webový odkaz na repozitář, vč. technické dokumentace a uživatelské příručky	ANO – návaznost na výstup/y NRP	XII.28

¹⁹⁵ Nejzazším termínem pro dosažení výstupu je termín ukončení fyzické realizace projektu, je však vhodné nastavit termíny adekvátně vzhledem k harmonogramu klíčových aktivit, tzn. v průběhu realizace projektu – dle charakteru výstupu.

B_3	Repozitář pro lidská i animální obrazová a fyziologická multimodální data (Imaging repozitář)	nový repozitář	KA2	Sada repozitářů pro TKA B/H/F	Repozitář pro lidská i animální obrazová a fyziologická multimodální data sloužící k bezpečnému ukládání, správě a sdílení dat. Vč. metadatového modelu, SW nástrojů pro rozšíření funkcí repozitáře a dokumentace. VIZ TAKÉ DÍLČÍ VÝSTUPY B_3.1–B_3.6 KA 2.3	webový odkaz na repozitář, vč. technické dokumentace a uživatelské příručky	ANO – návaznost na výstup/y NRP	XII.28
B_4	Repozitář pro data chemické biologie	nový repozitář	KA2	Sada repozitářů pro TKA B/H/F	Repozitář umožňující upload vybraných dat chemické biologie, vč. jejich reprezentace pomocí RDF pomocí vybraných ontologií a efektivního prohledávání pomocí jazyka SPARQL. VIZ TAKÉ DÍLČÍ VÝSTUPY B_4.1–B_4.4 KA 2.4.	webový odkaz na repozitář, vč. technické dokumentace a uživatelské příručky	ANO – návaznost na výstup/y NRP	XII.28
B_5	Propojení FAIR repozitáře chemické biologie s vývojem a tréninkem AI modelů pro MS	nástroje na rozvoj NDI	KA2	Nástroje a služby pro rozvoj NDI za TKA B/H/F	Propojené prostředí pro vývoj, trénování a validaci modelů strojového učení (AI) zaměřených především na analýzu hmotnostních spekter.	software uložený na Githubu	ANO – doplňkovost NRP	XII.28
B_6	ClinData Repozitář	nový repozitář	KA2	Sada repozitářů pro TKA B/H/F	Repozitář umožňující bezpečné ukládání, správu a zpracování dat o lidech a jejich napojení na další relevantní repozitáře. VIZ TAKÉ DÍLČÍ VÝSTUPY B_6.1–B_6.4 KA 2.5.	webový odkaz na repozitář, vč. technické dokumentace a uživatelské příručky	ANO – návaznost na výstup/y NRP	XII.28
B_7	Metodika FAIRifikace dat z ukončených regulovaných klinických studií	metodika/standard	KA2	Sada metodik pro TKA B/H/F	Jednotné metodické postupy umožňující dodržení standardů ECRIN a zvýšení kvality a dostupnosti dat a také integraci dat mezi oborově specifickými repozitáři.	odkaz na zveřejněný elektronický dokument	NE	XII.28
B_8	Adaptace ECRIN metadatového modelu pro metodiku	metodika/standard	KA2	Sada metodik pro TKA B/H/F	Metadatový model pro data z ukončených regulovaných klinických studií, vytvořený na základě standardů	odkaz na zveřejněný	NE	XII.28

	FAIRifikace dat z ukončených regulovaných klinických studií				ECRIN a adaptovaný pro potřeby repozitáře pro medicínská data.	elektronický dokument		
B_9	Nástroje pro Repozitář pro biologická zobrazovací data	nástroje na rozvoj NDI	KA2	Nástroje a služby pro rozvoj NDI za TKA B/H/F	Agregovaný soubor oborově specifických nástrojů pro Imaging repozitář.	veřejně dostupná webová stránka zpřístupňující daný nástroj	ANO – doplňkovost NRP	XII.28
M_1	Repozitář DANTE ^c	nový repozitář	KA3	Repozitář TKA MATECH	Repozitář DANTE ^c s 5 komunitami a 5 kolekcemi zahrnující metadatové profily, analýzy, SW rozšiřující funkce repozitáře a metodiky. VIZ TAKÉ DÍLČÍ VÝSTUPY M_1.1–M_1.6 KA 3.1.	webový odkaz na repozitář, vč. technické dokumentace a uživatelské příručky	ANO – návaznost na výstup/y NRP, EOSC CZ	XII.28
M_2	FAIR Implementation profiles pro komunity v repozitáři DANTE ^c	metodika/standard	KA3	Metodika TKA MATECH	Definice standardů pro podobory MATECH se zaměřením na kolekce repozitáře DANTE ^c .	odkaz na zveřejněný elektronický dokument	ANO – návaznost na výstup/y NRP	XII.28
M_3	Integrace ELN Kadi4Mat s repozitářem DANTE ^c	služby pro rozvoj NDI	KA3	Nástroje a služby pro rozvoj NDI za TKA MATECH	Nastavení procesů v ELN Kadi4Mat pro přímý sběr dat a metadat do kolekcí repozitáře DANTE ^c . Úprava repozitáře pro příjem dat z instancí ELN Kadi4Mat.	publikace o přímém odesílání záznamů z ELN do repozitáře	NE	VI.27
M_4	Integrace ELN eLabFTW s repozitářem DANTE ^c	služby pro rozvoj NDI	KA3	Nástroje a služby pro rozvoj NDI za TKA MATECH	Nastavení procesů v ELN eLabFTW pro přímý sběr dat a metadat do kolekcí repozitáře DANTE ^c . Úprava repozitáře pro příjem dat z instancí ELN eLabFTW.	publikace o přímém odesílání záznamů z ELN do repozitáře	NE	XII.28
M_5	SW nástroj pro správu uživatelů a přístrojů a	nástroje na rozvoj NDI	KA3	Nástroje a služby pro	Open source SW nástroj (balíček) pro správu uživatelů a přístrojů a jejich datových výstupů pro výzkumné	odkaz na zdrojové kódy k vyvinutému	NE	XII.28

	jejich datových výstupů ve výzkumných infrastrukturách			rozvoj NDI za TKA MATECH	infrastruktury s napojením na repozitář DANTE ^c a uživatelskou dokumentaci.	open source softwaru ve veřejném git repozitáři		
M_6	Integrace SW nástroje pro správu uživatelů a přístrojů a jejich datových výstupů ve výzkumných infrastrukturách s repozitářem DANTE ^c	služby pro rozvoj NDI	KA3	Nástroje a služby pro rozvoj NDI za TKA MATECH	Nastavení procesů v SW nástroji pro správu dat ve výzkumných infrastrukturách pro přímý sběr dat a metadat do kolekcí repozitáře DANTE ^c .	publikace o přímém odesílání záznamů z ELN do repozitáře	NE	XII.28
AI_1	Repozitář DM4AI	nový repozitář	KA4	Repozitář TKA DM4AI	Repozitář nad infrastrukturou NRP s předpokladem využití datových úložišť NRP umístěných v blízkosti rozsáhlé výpočetní infrastruktury, jako jsou výpočetní clustery e-Infra CZ, poskytující uživateli možnost ukládat natrénované AI modely, pipeline a datové sady a jejich metadata. VIZ TAKÉ DÍLČÍ VÝSTUPY AI_1.1 – AI_1.10 KA 4.1.	webový odkaz na repozitář, vč. technické dokumentace a uživatelské příručky	ANO – návaznost na výstup/y NRP, EOSC CZ, CARDS	XII.28
AI_2	SW definující prostředí pro inferenci generativních AI modelů uložených v datovém repozitáři	nástroje na rozvoj NDI	KA4	Nástroje pro rozvoj NDI za TKA DM4AI	Prostředí pro inferenci generativních AI modelů uložených v datovém repozitáři a interakci s nimi s důrazem na efektivitu inference pro různé typy a velikosti modelů.	nejméně jedna instance repozitářového systému CLARIN-DSpace, která bude mít integrované uživatelské prostředí pro komunikaci s generativními modely	NE	XII.28

S_1	Napojení repozitáře ČSDA do NRP a jeho rozvoj a provoz	inovovaný repozitář	KA5	Sada repozitářů pro TKA SOC	Integrace rozšířeného systému Dataverse do NRP, vč. integrace datových kolekcí. Obsahuje metadatové modely, kolekce dat a zahrnuje analýzy, metodiky. VIZ TAKÉ DÍLČÍ VÝSTUPY S_1.1–S_1.10 KA 5.1.	webový odkaz na repozitář, vč. identifikace inovovaných prvků	ANO – návaznost na výstup/y NRP	XII.28
S_2	Nový repozitář citlivých dat pro sociální vědy	nový repozitář	KA5	Sada repozitářů pro TKA SOC	Vybudování nového repozitáře pro citlivá data ze sociálněvědních výzkumů s využitím prostředků NRP, vč. metodiky a pilotní kolekce citlivých dat. VIZ TAKÉ DÍLČÍ VÝSTUPY S_2.1–S_2.2 KA 5.2.	webový odkaz na repozitář, vč. technické dokumentace a uživatelské příručky	ANO – návaznost na výstup/y NRP	XII.28
S_3	Inovovaný repozitář DataHub	inovovaný repozitář	KA5	Sada repozitářů pro TKA SOC	Inovace DataHub a Mapového a datového centra prostřednictvím zavedení napojení na NRP/NMA. VIZ TAKÉ DÍLČÍ VÝSTUPY S_3.1–S_3.3 KA 5.3.	webový odkaz na repozitář, vč. identifikace inovovaných prvků	ANO – návaznost na výstup/y NRP	XII.28
S_4	Nástroj pro vizualizaci dat	nástroje na rozvoj NDI	KA5	Nástroje pro rozvoj NDI za TKA SOC	Vizualizační softwarový webový nástroj sociálněvědních dat z výběrových šetření zobrazující vývoj postojů, názorů a přesvědčení cílové populace v čase.	veřejně dostupná webová stránka zpřístupňující daný nástroj	Ne	XII.28
S_5	Platforma Atlasobyvatelstva.cz pro specializované mapy	služby pro rozvoj NDI	KA5	Sada služeb pro rozvoj NDI za TKA SOC	Integrace platformy Atlasobyvatelstva.cz pro specializované mapy do datových služeb DataHub.	platforma dostupná jako součást DataHub na Webu	NE	XII.28
S_6	Demografická databáze integrovaná do datových služeb DataHub	služby pro rozvoj NDI	KA5	Sada služeb pro rozvoj NDI za TKA SOC	Databáze zaměřená na podrobná demografická data České republiky na různých geografických úrovních a časových osách.	databáze dostupná jako součást DataHub na Webu	Ne	XII.28

S_7	Integrace externích zdrojů dat	služby pro rozvoj NDI	KA5	Sada služeb pro rozvoj NDI za TKA SOC	Provedení landscape analýzy zaměřené na významné producenty dat a datových zdrojů, budou vybrány datové zdroje, jejichž data nelze nebo není vhodné umístit přímo do repozitářové platformy, a jejich zpřístupnění.	odkaz na webovou stránku	NE	XII.28
S_8	Analýza sdílení a praktického využívání dat z mezinárodních výzkumných infrastruktur	analýza	KA5	NE	Vypracování analýzy, která se zaměří na způsoby sdílení a praktického využívání dat z mezinárodních výzkumných infrastruktur jako ISSP, EVS, SHARE ERIC, ESS ERIC a GGP.	odkaz na zveřejněný elektronický dokument	NE	XII.28
S_9	Strategie komunikace a šíření výstupů projektu uvnitř odborné komunity v oborech sociálních věd	komunikační strategie	KA5	NE	Zacílená komunikační strategie ke zvýšení povědomí o datových službách a ke zlepšení kultury sdílení a managementu dat v českých sociálních vědách prostřednictvím celé řady diseminačních aktivit.	odkaz na zveřejněný elektronický dokument	NE	XII.28
S_10	Informační centrum k problematice nových forem dat a využití AI ve výzkumu do datových služeb	služby pro rozvoj NDI	KA5	Sada služeb pro rozvoj NDI za TKA SOC	Sady informačních materiálů, zdrojů a postupů, které budou zařazeny do struktury služeb ČSDA.	odkaz na zveřejněný elektronický dokument	Ne	XII.28
F_1	Vytvoření a provoz repozitáře pro obor Fyzika	nový repozitář	KA6	Repozitář TKA FYZIKA	Repozitář clusteru FYZIKA s 5 komunitami, zahrnující metadatové profily. VIZ TAKÉ DÍLČÍ VÝSTUPY F_1.1–F_1.6 KA 6.1.	webový odkaz na repozitář, vč. identifikace inovovaných prvků	ANO – návaznost na výstup/y NRP	XII.28
F_2	Nástroj pro automatizovanou tvorbu metadat	nástroje na rozvoj NDI	KA6	Sada nástrojů pro rozvoj NDI za TKA FYZIKA	Nástroj pro automatizovanou tvorbu metadat z měření fyzikálních veličin	veřejně dostupná webová stránka	NE	XII.28

					a jejich propojení s Elektronickým laboratorním deníkem (ELN).	zpřístupňující daný nástroj		
F_3	Nástroj pro vizuální zobrazení datových souborů	nástroje na rozvoj NDI	KA6	Sada nástrojů pro rozvoj NDI za TKA FYZIKA	Nástroj pro vizuální zobrazení datových souborů měření fyzikálních veličin přímo v internetovém prohlížeči.	veřejně dostupná webová stránka zpřístupňující daný nástroj	ANO – návaznost na výstup/y NRP	XII.28
H_1	Repozitář LINDAT/CLARIAH-CZ – zapojení jazykových korpusů infrastruktury ČNK jako samostatné kolekce repozitáře LINDAT/CLARIAH-CZ	inovovaný repozitář	KA7	Sada repozitářů pro TKA HUMA	Rozšíření repozitáře LINDAT/CLARIAH-CZ o novou komunitu Český národní korpus, jejím obsahem budou metadata všech veřejně dostupných korpusů ČNK. VIZ TAKÉ DÍLČÍ VÝSTUPY H_1.1–H_1.9 KA 7.1.	webový odkaz na repozitář, vč. identifikace inovovaných prvků	ANO – návaznost na výstup/y NRP	XII.28
H_2	Repozitář Digitalia MUNI ARTS	inovovaný repozitář	KA7	Sada repozitářů pro TKA HUMA	Repozitář transformující výzkumné kolekce dat v jednotlivých platformách na AIP balíčky ve stanoveném formátu dat a struktuře metadat. VIZ TAKÉ DÍLČÍ VÝSTUPY H_2.1–H_2.4 KA 7.2.	webový odkaz na repozitář, vč. identifikace inovovaných prvků	ANO – návaznost na výstup/y NRP	X.26
H_3	Repozitář ArchaeoVault	inovovaný repozitář	KA7	Sada repozitářů pro TKA HUMA	Repozitář ArchaeoVault transformuje výzkumné kolekce dat v jednotlivých platformách na AIP balíčky ve stanoveném formátu dat a struktuře metadat a uloží je do úložiště NRP. Vč. analýzy best practice a metodiky zpracování dat. VIZ TAKÉ DÍLČÍ VÝSTUPY H_3.1–H_3.6 KA 7.3.	webový odkaz na repozitář, vč. identifikace inovovaných prvků	ANO – návaznost na výstup/y NRP	V.28
H_4	Repozitář pro bibliografická data	nový repozitář	KA7	Sada repozitářů pro TKA HUMA	Nový repozitář pro sběr, ukládání a zpřístupňování bibliografických dat, vč. metadatových modelů, workflow a metodik pro práci s daty. VIZ TAKÉ DÍLČÍ VÝSTUPY H_4.1–H_4.9 KA 7.4.	webový odkaz na repozitář, vč. technické dokumentace a uživatelské příručky	ANO – návaznost na výstup/y NRP	XII.28

H_5	Soubor nadstavbových nástrojů pro rozvoj repozitářů TKA HUMA	nástroje na rozvoj NDI	KA7	Sada nástrojů pro rozvoj NDI TKA za HUMA	Soubor nadstavbových nástrojů pro zpracování a vizualizaci dat poskytující vyšší komfort pro práci s obsahem konkrétních repozitářů. VIZ TAKÉ DÍLČÍ VÝSTUPY H_5.1–H_5.4 KA 7.5.	veřejně dostupná webová stránka zpřístupňující daný nástroj	ANO – návaznost na výstup/y NRP	XII.28
E_1	Inovovaný repozitář GENASIS	inovovaný repozitář	KA8	Sada repozitářů pro TKA ENVIRO	Inovovaný repozitář pro prezentaci dat o chemické expozici matric životního prostředí a lidských tkání umožňující integraci různých dat, vč. metadatového modelu, setů metadat a FAIRifikovaných datasetů. VIZ TAKÉ DÍLČÍ VÝSTUPY E_1.1–E_1.4 KA 8.1.	webový odkaz na repozitář, vč. identifikace inovovaných prvků	ANO – návaznost na výstup/y NRP	XII.28
E_2	Repozitář pro ukládání dat z necílených hmotnostně spektroskopických analýz pro účely hodnocení lidského expozomu	nový repozitář	KA8	Sada repozitářů pro TKA ENVIRO	Nový repozitář pro primární hmotnostně spektroskopická data z necílené metabolomiky a expozomiky pro lidské vzorky, vč. metadatového modelu, setů metadat a FAIRifikovaných datasetů. VIZ TAKÉ DÍLČÍ VÝSTUPY E_2.1–E_2.3 KA 8.2.	webový odkaz na repozitář, vč. technické dokumentace a uživatelské příručky	ANO – návaznost na výstup/y NRP	XII.28
E_3	Repozitář pro ukládání toxikologických a ekotoxikologických dat	nový repozitář	KA8	Sada repozitářů pro TKA ENVIRO	Repozitář pro ukládání toxikologických a ekotoxikologických dat, vč. analýzy potřeb, vč. metadatového modelu a FAIRifikovaných datasetů. VIZ TAKÉ DÍLČÍ VÝSTUPY E_3.1–E_3.4 KA 8.3.	webový odkaz na repozitář, vč. technické dokumentace a uživatelské příručky	ANO – návaznost na výstup/y NRP	XII.28
E_4	Nové úložiště umožňující propojení geokódovaných dat z různých domén	nový repozitář	KA8	Sada repozitářů pro TKA ENVIRO	Nové úložiště umožňující propojení geokódovaných dat z různých oblastí za účelem studia zdravotních rizik. Vč. analýzy potřeb, na jejímž základě bude vytvořen design a metadatový	webový odkaz na repozitář, vč. technické dokumentace a uživatelské příručky	ANO – návaznost na výstup/y NRP	XII.28

					model tohoto repozitáře. VIZ TAKÉ DÍLČÍ VÝSTUPY E_4.1–E_4.2 KA 8.4.			
E_5	Nový repozitář pro referenční obrazová data živých rostlin a rostlinných společenstev	nový repozitář	KA8	Sada repozitářů pro TKA ENVIRO	Nový repozitář pro referenční obrazová data živých rostlin a rostlinných společenstev, vč. metadatového modelu, setů metadat, FAIRifikovaných datasetů a metodiky pro uživatele. VIZ TAKÉ DÍLČÍ VÝSTUPY E_5.1–E_5.4 KA 8.5.	webový odkaz na repozitář, vč. technické dokumentace a uživatelské příručky	ANO – návaznost na výstup/y NRP	XII.28
E_6	Nový repozitář pro genetický biomonitoring a genetická data volně žijících organismů	nový repozitář	KA8	Sada repozitářů pro TKA ENVIRO	Nový repozitář pro genetický biomonitoring a genetická data volně žijících organismů, vč. metadatového modelu, setů metadat a FAIRifikovaných datasetů a metodiky pro uživatele. VIZ TAKÉ DÍLČÍ VÝSTUPY E_6.1–E_6.4 KA 8.6.	webový odkaz na repozitář, vč. technické dokumentace a uživatelské příručky	ANO – návaznost na výstup/y NRP	XII.28
E_7	Nový repozitář pro zoologické sbírky	nový repozitář	KA8	Sada repozitářů pro TKA ENVIRO	Nový repozitář pro zoologické sbírky. Rozšíření tematického clusteru Repozitáře pro biodiverzitní data o novou sekci pro negenetické zoologické sbírky (taxidermické a lihové preparáty, kostry) a jeho propojení s aktuálně připravovanou sekcí pro herbářové sbírky. VIZ TAKÉ DÍLČÍ VÝSTUPY E_7.1–E_7.4 KA 8.7.	webový odkaz na repozitář, vč. technické dokumentace a uživatelské příručky	ANO – návaznost na výstup/y NRP	XII.28
E_8	Nové bioinformatické nástroje na analýzu hmotnostně spektroskopických dat z necílených metabolomických a expozomických analýz	nástroje na rozvoj NDI	KA8	Sada nástrojů pro rozvoj NDI za TKA ENVIRO	Standardizované a dokumentované nástroje pro vytěžování hmotnostně spektroskopických dat s vysokým rozlišením ke kvalitativní i kvantitativní analýze širokého spektra exogenních i endogenních látek.	veřejně dostupná webová stránka zpřístupňující daný nástroj	ANO – návaznost na výstup/y NRP	XII.28

E_9	AI nástroje na analýzu obrazu biologických objektů a automatické čtení herbářových sched	nástroje na rozvoj NDI	KA8	Sada nástrojů pro rozvoj NDI za TKA ENVIRO	Nasazení a modifikace AI modelu pro obrazovou analýzu herbářových položek (např. HESPI https://rbturnbull.github.io/hespi/), integrace LLM modelů pro automatizaci přepisu metadat z herbářových sched do workflow repozitáře.	veřejně dostupná webová stránka zpřístupňující daný nástroj	ANO – návaznost na výstup/y NRP	XII.28
SE_1	Analýza správné praxe pro sdílení citlivých dat z kolaborativního výzkumu	analýza	KA 9	NE	Dokument shrnující správnou praxi stran sdílení citlivých dat z kolaborativního výzkumu.	odkaz na zveřejněný elektronický dokument	NE	III.27
SE_2	Metodika FAIRifikace citlivých dat v NDI	metodika/standard	KA 9	Metodika pro TKA SENSÍ	Souhrnná metodika pro komplexní správu citlivých dat v NDI.	odkaz na zveřejněný elektronický dokument	ANO – návaznost na výstup/y NRP	VI.28
SE_3	Softwarový modul pro kontrolu a zajištění kvality dat repozitářů	nástroje na rozvoj NDI	KA 9	Nástroje pro rozvoj NDI za TKA SENSÍ	Softwarový modul pro kontrolu a zajištění kvality dat repozitářů NRP vyvíjen v eCRF systému ClinData.	veřejně dostupná webová stránka zpřístupňující daný nástroj, včetně dokumentace	ANO – doplňkovost NRP	XII.28
SE_4	Služba řízeného zabezpečeného procesu příjmu/výdeje dat do/z prostředí pro zpracování citlivých dat	Služby pro rozvoj NDI	KA 9	Sada služeb pro rozvoj NDI za TKA SENSÍ	Implementace zabezpečeného řízeného procesu příjmu/výdeje dat do/z prostředí pro zpracování citlivých dat.	technická implementace služby dostupná pro výzkumníky v ČR, včetně dokumentace	ANO – návaznost na výstup/y NRP	XII.28
SE_5	Rozhraní pro koncové uživatele pro interaktivní/dávkové zpracování citlivých dat	Služby pro rozvoj NDI	KA 9	Sada služeb pro rozvoj NDI za TKA SENSÍ	Identifikace vhodných rozhraní pro interaktivní a dávkové zpracování citlivých dat v zabezpečeném prostředí a tato budou implementována.	technická implementace služby dostupná pro výzkumníky	ANO – návaznost na výstup/y NRP	XII.28

	v zabezpečeném prostředí					v ČR, včetně dokumentace		
SE_6	Služba pro autorizaci dávkového zpracování citlivých dat	Služby pro rozvoj NDI	KA 9	Sada služeb pro rozvoj NDI za TKA SENSÍ	Nastavení a implementace mechanismu, který umožní autorizaci požadavků na dávkové zpracování citlivých dat v prostředí typu TRE.	technická implementace služby dostupná pro výzkumníky v ČR, včetně dokumentace	NE	XII.28
SE_7	Služba pro vyhodnocení analytických workflow a klasifikaci citlivosti jejich výstupů	Služby pro rozvoj NDI	KA 9	Sada služeb pro rozvoj NDI za TKA SENSÍ	Služba poskytne systematické vyhodnocení citlivosti výstupů analytických workflow.	technická implementace služby dostupná pro výzkumníky v ČR, včetně dokumentace	ANO – doplňkovost NRP	XII.28
LLM_1	Technická zpráva: sběr požadavků a návrh architektury pro využití velkých jazykových modelů	technická zpráva	KA 10	NE	Technická zpráva se závěry analýzy požadavků kladených na technologii virtuálního asistenta z hlediska společenských, právních, psychologických a etických aspektů a vytvořený návrh architektury a procesů pro strojové zpracování uživatelských vstupů.	odkaz na zveřejněný elektronický dokument	NE	II.26
LLM_2	Výsledná verze systému virtuálního asistenta pro nasazení v helpdesk prostředí	nástroje na rozvoj NDI	KA 10	Sada nástrojů pro rozvoj NDI za PKA	Systém s technickou dokumentací a dokumentací pro obsluhu, který bude pilotně nasazen a provozován pro uživatele-výzkumníky a L2/L3 pracovníky helpdesku.	odkaz na zdrojové kódy k vyvinutému open-source softwaru ve veřejném git repozitáři	NE	XII.28
KB_1	Požadavky na systém bezpečnostního monitoringu	technická zpráva	KA 10	NE	Technická zpráva popisující závěry analýzy požadavků kladených na systém bezpečnostního monitoringu a vytvořený	odkaz na zveřejněný	NE	II.26

					návrh architektury včetně popisu jednotlivých komponent systému.	elektronický dokument		
KB_2	Doporučené postupy pro prostředí Open Science pro efektivní obranu proti kybernetickým útokům	technická zpráva	KA 10	Sada metodik za PKA	Doporučené postupy pro zabezpečení specifického prostředí NRP. Soubor osvědčených metod a strategií pro efektivní obranu proti kybernetickým útokům.	odkaz na zveřejněný elektronický dokument	NE	XII.26
KB_3	Sada nástrojů proaktivního bezpečnostního monitoringu	nástroje na rozvoj NDI	KA 10	Sada nástrojů pro rozvoj NDI za PKA	Sada nástrojů proaktivního bezpečnostního monitoringu.	odkaz na zdrojové kódy k vyvinutému open source softwaru ve veřejném git repozitáři	ANO – doplňkovost NRP	XII.27
KB_4	Nástroje detekce bezpečnostních hrozeb a podpory reaktivní bezpečnosti	nástroje na rozvoj NDI	KA 10	Sada nástrojů pro rozvoj NDI za PKA	Sada nástrojů pro detekci bezpečnostních hrozeb a poskytování informací pro podporu reaktivní bezpečnosti.	odkaz na zdrojové kódy k vyvinutému open source softwaru ve veřejném git repozitáři	ANO – doplňkovost NRP	XII.28
KB_5	Dokumentace vyvinutého ekosystému a zpráva o pilotním nasazení	technická zpráva	KA 10	Sada metodik za PKA	Dokumentace vyvinutého ekosystému poskytující podrobný přehled o architektuře, funkcionalitách a integraci jednotlivých vyvinutých částí systému.	odkaz na zveřejněný elektronický dokument	ANO – doplňkovost NRP	XII.28
C_1	Sada doporučení a metodika zejména v oblasti umělé inteligence a zpracování dat	technická zpráva	KA 10	Sada metodik za PKA	Dokument zaměřený na problematiku využití LLM z pohledu práva, shrne stav a postup realizace technologických prostředků pro bezpečnost a uživatelskou podporu s využitím metod umělé inteligence.	odkaz na zveřejněný elektronický dokument	ANO – doplňkovost NRP	XII.28

ELN_1	Integrace ELN eLabFTW s repozitáři NRP	nástroje na rozvoj NDI	KA 10	Sada nástrojů pro rozvoj NDI za PKA	Implementovaný nástroj, který uživateli umožní publikovat data a metadata ze záznamů v ELN Kadi4Mat do různých oborových datových repozitářů a do „catch-all“ repozitáře NRP v NDI.	technická implementace služby dostupná pro výzkumníky v ČR, včetně dokumentace	ANO – návaznost na výstup/y NRP	VI.27
ELN_2	Integrace ELN Kadi4Mat s repozitáři NRP	nástroje na rozvoj NDI	KA 10	Sada nástrojů pro rozvoj NDI za PKA	Implementovaný nástroj, který uživateli umožní publikovat data a metadata ze záznamů v ELN eLabFTW do různých oborových datových repozitářů a do „catch-all“ repozitáře NRP v NDI.	technická implementace služby dostupná pro výzkumníky v ČR, včetně dokumentace	ANO – návaznost na výstup/y NRP	XII.28
ELN_3	Metodika nasazení ELN Kadi4Mat a eLabFTW	metodika/standard	KA 10	Sada metodik za PKA	Metodika nasazení ELN Kadi4Mat a eLabFTW ve vědeckých organizacích, které si vytvoří a budou administrovat svoji vlastní instanci za použití prostředků NDI.	odkaz na zveřejněný elektronický dokument	ANO – návaznost na výstup/y NRP	XII.27
ELN_4	Nasazení „catch-all“ instancí ELN	nástroje na rozvoj NDI	KA 10	Sada nástrojů pro rozvoj NDI za PKA	1 instance ELN Kadi4Mat, 1 instance ELN eLabFTW.	technická implementace služby dostupná pro výzkumníky v ČR, včetně dokumentace	ANO – návaznost na výstup/y NRP	XII.26
ELN_5	Inovovaný ELN Kadi4Mat	nástroje na rozvoj NDI	KA 10	Sada nástrojů pro rozvoj NDI za PKA	Propojení různých instancí ELN Kadi4Mat, které umožní uživatelům po autorizaci interaktivně sdílet data mezi instancemi ve smyslu vzájemného přímého přístupu bez exportu a importu.	Zdrojový kód veřejně přístupný spolu s historií jeho vývoje	NE	XII.28
P_1	SW pro ukládání a správu provenance	nástroje na rozvoj NDI	KA 10	Sada nástrojů pro rozvoj NDI za PKA	SW, který umožní kontrolu integrity a autenticity uložených provenance	veřejně dostupná webová stránka	ANO – doplňkovost NRP	VI.28

					dokumentů pomocí hašů a digitálních podpisů.	zpřístupňující daný nástroj		
P_2	SW knihovna pro práci s provenance dle CPM	nástroje na rozvoj NDI	KA 10	Sada nástrojů pro rozvoj NDI za PKA	SW knihovna umožní reprezentaci struktur definovaných v CPM; generování provenance dokumentů pomocí šablon atd.	veřejně dostupná webová stránka zpřístupňující daný nástroj	ANO – doplňkovost NRP	VI.28
P_3	Metodika pro provenance	metodika/standard	KA 10	Sada metodik za PKA	Dokument obsahující seznam doporučení týkajících se správy harmonizované provenance v NDI tak, aby byla zpráva provenance v souladu se současnými best practice a dostupnými standardy souvisejícími s AI Act.	odkaz na zveřejněný elektronický dokument	ANO – doplňkovost NRP	XII.28
K_2	Sada vzdělávacích materiálů OS II	vzdělávací materiály	KA 10	Vzdělávací materiály OS II	Sada vzdělávacích materiálů obsahuje: AI_1.11 (KA 4) Vzdělávací materiály/ tutoriály; S_10.1 (KA 5) Materiály pro vzdělávání datového managementu a metodologie sociálněvědního výzkumu; S_10.2 (KA 5) Vytvoření guidelines materiálů pro vzdělávání datového managementu a metodologie sociálněvědního výzkum; F_4.1 (KA 6) Série tří interaktivních e-learningových kurzů FYZIKA; F_4.2 (KA 6) Vzdělávací materiály pro čtyři workshopy FYZIKA; H_1.9 (KA 7) FAIR data při nasazení a provozu instance repozitáře DSpace v rámci EOSC CZ; H_4.10 (KA 7) Vzdělávací materiály pro práci s repozitářem pro bibliografická data;	odkaz na zveřejněný elektronický dokument	ANO – podklad pro školicí akce pořádané Školícím centrem EOSC CZ	XII.28

					E_1.4 (KA 8) Standardní operační postupy a tutoriály k tvorbě a validaci FAIRifikovaných datasetů; SE_2.1 (KA 9) Vzdělávací materiály navazující na metodiku FAIRifikace citlivých dat.			
K_1	Konference OS II	Konferen-ce	KA 10	NE	4 odborné oborové konference, které poskytnou prostor pro sdílení výstupů projektu a networking pro odbornou komunitu.	Program konference, prezentace	ANO – kooperace se Školicím centrem EOSC CZ	XII.28

