ARP konferencia 2026

Hosszú távon is elérhető? – Adatrepozitóriumok és kutatási adatok: fenntarthatóság, biztonság és innováció

Időpont: 2026. október 8.

Helyszín: Humán Tudományok Kutatóháza és online

1. Szekció: Adatrepozitóriumok, adattárolás, adatmenedzsment

Holl András (MTA KIK): Elmélkedések az adatrepozitóriumokról és a hosszú távú fenntarthatóságról

A kutatási adatok megőrzéséről és elérhetővé tételéről folyó modern hazai diskurzus több, mint tíz éve folyik. Talán érdemes elővenni néhány alapvető kérdést, amelyek egy részére az alkalmazásban már valamilyen válasz született, más részük pedig még továbbra is megoldatlan.

Vizsgáljuk az adatrepozitálás célját, tárgyát, finanszírozását, időtávját, az adatok beágyazottságát, megőrzésének végét, és szót ejtünk az OAIS szabvány értelmezéséről is.

Egyed-Gergely Júlia (ELTE TK): Mérföldkövek a legnagyobb hazai kutatási hálózat kutatásiadat-repozitálási gyakorlatában

A legnagyobb hazai tudományos kutatási hálózat intézményeiben folyamatosan keletkeznek kutatási adatok. A nagy és egyre gyorsabb ütemben növekedő mennyiségű, formailag és tartalmilag rendkívül diverz és legfőképpen rendkívül értékes adatvagyon tárolása, a hosszú távon is megőrzésre érdemes kutatási adatok kiválasztása és azok biztonságos megőrzése feladatot és kihívást jelent minden intézmény számára.

Egyes kutatóhelyek, mint például a Közgazdaság- és Regionális Tudományi Kutatóközpont, a Számítástechnikai és Automatizálási Kutatóintézet vagy a Társadalomtudományi Kutatóközpont, az akkori korhoz és a nemzetközi trendekhez igazodva, már hosszú évekkel ezelőtt módszereket és irányelveket dolgoztak ki, illetve komoly infrastruktúrát fejlesztettek (mint a KRTK Adatbank, a Concorda Adatrepozitórium és a KDK Repozitórium) kutatási adataik megfelelő kezelésének és megőrzésének biztosítására.

2021 őszén az akkori ELKH vezetése támogatását adta egy, a SZTAKI, a TK és a Wigner FK szakmai együttműködésével kidolgozott, közös, interdiszciplináris adatrepozitórium

létrehozására irányuló fejlesztési terv megvalósításához. Ennek eredményeként indult el az Adatrepozitórium Platform (ARP) fejlesztése, amely 2024 tavaszára készült el, és amely azóta a HUN-REN és a hazai felsőoktatási intézmények kutatócsoportjai és kutatói számára biztosít lehetőséget kutatási adataik biztonságos, digitális őrzésére illetve mások számára való elérhetővé tételére. Elindulása óta az ARP a hosszú távú, biztonságos működésre és fenntarthatóságra törekszik, ennek megteremtésén dolgozik technikai, szabályozási, humántámogatási és gazdasági szempontból is.

Az előadás két példán keresztül vizsgálja a repozitóriumi világ működését és fenntarthatósági kérdéseit. Bemutatja, milyen mérföldköveken van túl, hol tart most, és milyen kihívások előtt áll ma a legnagyobb hazai intézményi diszciplináris repozitórium, a KDK Repozitórium, és a legnagyobb hazai interdiszciplináris adatrepozitórium, az ARP Adatrepozitórium. Mik a működés követelményei, a hosszú távú adattárolás biztosítékai és mik a működést nehezítő tényezők? Hogyan lehet megteremteni a szükséges feltételeket, és mit lehet tenni a kihívások kezelése érdekében? A prezentáció ezekre és hasonló kérdésekre keresi a válaszokat.

Kovács László (HUN-REN SZTAKI): ARP, honnan, hová

Az ARP Magyarországon elsőként biztosít teljes körű infrastruktúrát a kutatási adatok repozitálásához. A rendszer nem csupán az adatok tárolását, hosszú távú megőrzését és megosztását támogatja, hanem az adattárolás és a visszakereshetőség alapját képező metaadatsémák kezeléséhez is átfogó szolgáltatásokat nyújt. Ezek magukban foglalják a metaadatsémák regisztrálását, megőrzését és fejlesztésének támogatását.

Az ARP-ben az adatrepozitálás alapegysége a kutatási objektum, amelynek szabványos informatikai reprezentációját az RO-Crate csomagformátum biztosítja.

Az ARP teljes körű informatikai megoldást kínál az RO-Crate formátumú kutatási adatcsomagok kezelésére. A rendszer egyedi, fájlszintű metaadatolást is biztosít, amelynek köszönhetően a tárolt és megosztott kutatási objektumok kereshetővé és átláthatóvá válnak. Nemcsak maguk a kutatási objektumok, hanem az azokat alkotó egyedi fájlok is kereshetők, így a repozitált adatcsomagok részletes tartalma már a letöltés, kibontás és feldolgozás előtt áttekinthető és ellenőrizhető. Ez különösen a nagy méretű kutatási objektumok esetében jelent kiemelkedő előnyt.

Az előadás áttekinti a HUN-REN SZTAKI Elosztott Rendszerek Osztálya által az ARP-ben az elmúlt év során megvalósított fejlesztéseket és szolgáltatásbővítéseket, valamint bemutatja a további fejlesztések irányait. Ezek elsődleges célja az EOSC Hungarian Node kialakításának támogatása, illetve a FAIR alapelvek egyre teljesebb és pontosabb érvényesítése a magyar kutatási adatok kezelésében.

2. Szekció: Specifikus tudományterületek kutatási adatai és kutatásiadat-kezelési jógyakorlatai

Lencsés Ákos (ELTE NYTK): Egy 1941-es egyetemi adatgyűjtés tanulságai

Az előadás bemutatja az egyetemisták számára kiírt 1941-es Csűry Bálint országos népnyelvi szókincsgyűjtő verseny egyik pályázójának adatgyűjtését. A néprajz és a nyelvészet határmezsgyéjén mozgó, 200 fényképes adatlapot tartalmazó pályázati anyag a Nyelvtudományi Kutatóközpontban vészelte át az elmúlt 85 évet. Feldolgozására és szakmai értékelésére azonban napjainkig nem került sor. A beszámoló a pályázati adatlapokkal kapcsolatos kérdéseket és a lehetséges megoldásokat veszi sorba, bemutatva, hogy milyen problémákkal kell szembenézni a digitalizálás, a metaadatolás, a jogi kérdések tisztázása és a megfelelő őrzési hely kiválasztása során.

Kiss László (ELTE TK): A történeti tudásgráfoktól a platformfüggetlen szemantikai referenciaarchitektúráig – egy lehetséges út a FAIR történeti kutatási adatok felé

A digitális bölcsészet és a társadalomtudományok elmúlt két évtizedes fejlődése nyomán nagyszámú történeti adatbázis, tudásgráf és szemantikus információs rendszer jött létre. Ezek jelentős mértékben hozzájárultak a kutatási adatok strukturált kezeléséhez és újrafelhasználhatóságához, ugyanakkor eltérő ontológiai modellekre, adatstruktúrákra és szemantikai megoldásokra épülnek. A prozopográfiai, archontológiai, genealógiai vagy intézménytörténeti adatbázisok, valamint a Wikidata, a FactGrid vagy a CIDOC CRM különböző reprezentációs logikát alkalmaznak, ami jelentősen korlátozza az adatok interoperabilitását és integrált újrafelhasználását. A probléma tehát nem elsősorban újabb tudásgráfok hiánya, hanem egy olyan közös fogalmi keret hiánya, amely lehetővé teszi a különböző kutatási adatmodellek összekapcsolását.

Az előadás amellett érvel, hogy a történeti kutatási adatok FAIR-alapú szervezésének egyik lehetséges iránya egy platformfüggetlen szemantikai referenciaarchitektúra kialakítása. Egy ilyen referenciaarchitektúra nem kívánja kiváltani a meglévő ontológiákat és tudásgráf-platformokat, hanem olyan közös reprezentációs modellt biztosít, amelyre különböző kutatási hagyományok és technológiai környezetek egyaránt leképezhetők. A megközelítés kiindulópontja, hogy a történeti és társadalomtudományi kutatásokban alkalmazott adatmodellek jelentős része ugyanazokat az alapvető fogalmi egységeket – személyeket, intézményeket, eseményeket, dokumentumokat, időt, bizonyítékokat és relációkat – reprezentálja, csupán eltérő modellezési döntésekkel.

A koncepció működését a magyarországi kommunista politikai és államigazgatási elitet (1945–1990) reprezentáló FactGrid-alapú tudásgráf példáján mutatom be. Ebben az

esetben a FactGrid nem végső célként, hanem referenciaimplementációként jelenik meg, amely szemlélteti, miként ültethető át egy absztrakt szemantikai modell működő kutatási infrastruktúrává. Az előadás ismerteti az ontológiai modell kialakítását, a szemantikus adatmodellt, a történeti forrásokból kiinduló adatépítési munkafolyamatot, valamint azt, hogy a perzisztens azonosítók, a szemantikus állítások, az időkezelés, a forráshivatkozások és a SPARQL-alapú lekérdezések miként járulnak hozzá a FAIR kutatási adatok megvalósításához.

Az előadás végkövetkeztetése szerint a történeti kutatási adatok hosszú távú interoperabilitását nem önmagában egy újabb tudásgráf vagy technológiai platform biztosítja, hanem egy olyan platformfüggetlen szemantikai referenciaarchitektúra, amely egységes fogalmi alapot nyújt a különböző kutatási hagyományok és tudásgráf-megoldások számára. Ebben a megközelítésben a tudásgráf már nem a végtermék, hanem a referenciaarchitektúra egyik lehetséges implementációja, amely egyúttal a FAIR kutatási adatkezelési elvek gyakorlati érvényesítésének eszköze is.

Szederkényi Róbert (ELTE HTK): Az ásatási dokumentációtól a repozitóriumig – az adatgondozás intézményi kihívásai

A régészeti adatok a kutatás teljes életciklusa során, egymásra épülő munkafolyamatokban jönnek létre. A terepi megfigyelések, leírások, rajzok, fényképek és térinformatikai állományok együtt alkotják azt a dokumentációs rendszert, amelynek hosszú távú megőrzését, kereshetőségét és újrafelhasználhatóságát az adatgondozás biztosítja. Ehhez elengedhetetlen az állományok eredetének és összefüggéseinek dokumentálása, a megfelelő metaadatolás és a repozitóriumi előkészítés.

Az előadás azt mutatja be, hogy az ELTE HTK Régészeti Kutatóintézetében a hagyományos adattári feladatok miként egészülnek ki a kutatásiadat-kezelés és a repozitóriumi adatgondozás gyakorlatával. A bemutatott tapasztalatok az intézet 2023 óta megvalósuló ARP Nagykövet projektjeire épülnek, amelyek három eltérő keletkezéstörténetű régészeti adatkészlet példáján keresztül mutatják be az adatgondozás intézményi kihívásait.

A három esettanulmány az adatgondozás eltérő intézményi kihívásait mutatja be. Zalavár-Vársziget esetében a primer, jórészt analóg dokumentáció digitalizálása, rendszerezése és metaadatolása állt a középpontban. A Pilismarót-Basaharc projekt egy retrospektív kutatási adatkészlet integrálásának kihívásait szemlélteti, amely korábbi feltárások digitalizált dokumentációját kapcsolja össze az újonnan létrejött digitális feldolgozási állományokkal. Balatonszentgyörgy-Faluvégi dűlő 2. csaknem teljesen digitálisan keletkezett dokumentációja pedig a különböző intézményekben és munkafázisokban létrejött állományok összegyűjtését, a verziók és redundanciák kezelését, valamint a régészeti, természettudományos és térinformatikai adatok

egységes szerkezetbe rendezését igényelte. A három példa jól mutatja, hogy a repozitóriumi elhelyezés az adatéletciklus egyik állomása, nem pedig annak kezdete vagy lezárása.

Az előadás kitér arra is, hogy a repozitóriumi előkészítés során a technikai feladatok mellett szervezeti és jogi kérdéseket is rendezni kell. Az adatkészletek gyakran több intézmény és közreműködő együttműködésében jönnek létre, ezért a közzététel feltételeinek, a felhasználási jogosultságoknak, valamint az embargó vagy a korlátozott hozzáférés alkalmazásának tisztázása a felelős kutatásiadat-kezelés szerves része. Az adatok újrafelhasználhatóságát nem csak a repozitóriumi elhelyezés teremti meg: ehhez a teljes életcikulson át következetes adatgondozásra van szükség.

A hagyományos adattári munka egyre inkább adatgazdászi feladatokkal bővül. Az adattáros közvetítő szerepet tölt be a kutatók, a dokumentációt létrehozó szakmai közösség és a repozitóriumi infrastruktúra között. Feladata nem csupán az állományok megőrzése és összefüggéseik dokumentálása, hanem azok rendezése, metaadatolása és hosszú távú újrafelhasználhatóságuk biztosítása is.

Az előadás végül arra is kitér, hogy a mesterséges intelligencia miként támogathatja az adatgondozási munkafolyamatokat. A gépi látás és a multimodális eljárások támogathatják a képi állományok feltárását és metaadatolását, alkalmazásuk azonban csak dokumentált, szakértői ellenőrzésre épülő munkafolyamatokban tekinthető megbízhatónak. Az MI által generált tömegtartalmak világában ugyanakkor felértékelődik a közgyűjtemények, adattárak, levéltárak szerepe, amelyek nemcsak az adatok megőrzését szolgálják, hanem eredetük, kontextusuk és hitelességük garanciáját is jelentik.

Pálvölgyi-Polyák Eszter (ELTE Egyetemi Könyvtár és Levéltár): (Meta)adatok a jövőnek: Mozaikok az Inter-university Consortium for Political and Social Research (ICPSR) metaadat-kezelési jó gyakorlataiból

A University of Michigan keretein belül működő Inter-university Consortium for Political and Social Research (ICPSR) a világ egyik vezető adatrepozitóriumaként az 1960-as évek óta meghatározó szerepet tölt be a társadalom- és viselkedéstudományi adatok megőrzésében és közreadásában. Jelen előadás célja, hogy az ICPSR-t röviden bemutassa és esettanulmányként használja fel az adatrepozitóriumi metaadat-kezelés jó gyakorlatainak bemutatására. Az ICPSR az elmúlt két évtized folyamán számos olyan projektet valósított meg, amelyek célja többek között a kutatási adatot leíró metaadat fejlesztése, ezáltal a FAIR alapelvek jobb megvalósítása volt. Ezek közé tartoznak például az ICPSR Bibliography of Data-related Literature, amely az elmúlt 25 évben a bizonyítottan adatot felhasználó publikációkat kötötte össze az adatok nyilvános rekordjaival és kiemelt figyelmet szentelt az adathivatkozási jó gyakorlatok

elterjesztésének, vagy az ICPSR Metadata Documentation Project, ami strukturált formában nyilvánosan elérhetővé tette az ICPSR által használt metaadat-elemeket az adatot közzéadó kutatók és a másodlagos felhasználók számára. Emellett megjelenik még a Research Data Ecosystem (RDE) National Science Foundation (NSF) által támogatott adat-infrastrukturális projektje, ami a kutatási adat teljes életciklusára készített új technológiai megoldásokat. Ezek a törekvések mind közjavakként tekintenek a kutatási adatokra és azt a célt szolgálják, hogy az adat alkalmas legyen másodlagos felhasználásra.

Az előadásban az ICPSR Metadata and Preservation részlegén eltöltött öt éves munkatapasztalatomat is felhasználva mutatom be azokat a metaadathoz kapcsolódó gyakorlatokat, amikre érdemes építeni a társadalomtudományi adatrepozitóriumok fejlesztése során. Célom, hogy bemutassam az egységes dokumentációban és tezaurusz-építésben, a kutatási adatot másodlagosan felhasználó publikációk gyűjtésében és az API- és mesterséges intelligencia-támogatású megoldásokban rejlő potenciált, amelyek egyúttal a FAIR alapelvek gyakorlati megvalósításának kiváló példái is. Az ICPSR egy több száz tagot számláló, hatalmas erőforrásokkal rendelkező intézmény, amely egy, az európaihoz képest merően eltérő jogi környezetben működik, ezért nem minden gyakorlatuk ültethető át teljes mértékben a hazai környezetbe. Az előadásban bemutatott egyes példák viszont hasznos kitekintést nyújthatnak a magyar adatrepozitóriumi közösség számára is.

3. Szekció: Technikai fejlesztések és lehetőségek

Kun László (HUN-REN, Nemzeti Közszolgálati Egyetem, PhD): A digitális adatok menedzsmentjében használt módszertanok, adatkormányzási gyakorlatok bemutatása

Jelenleg a Nemzeti Közszolgálati Egyetem doktorandusz képzésén veszek részt, kutatásom és doktori disszertációm témája a különböző szervezetek -- ezen belül elsősorban nem a profitorientált vállalkozások - által gyűjtött, tárolt és kezelt digitális adatok szervezeten belüli használatának és hasznosításának lehetőségei. A kutatás célja ezen túl olyan megoldások elemzése, amellyel a digitális adatokat kezelő szervezetek működésüket javíthatják, fejleszthetik és felkészülhetnek a digitalizáció által támasztott kihívásokra, gyors változásokra (például a mesterséges intelligencia gyors terjedésére és használatára.).

Általánosan jellemző, hogy a különböző szervezetek működése egyre inkább digitalizálódik, eltolódik a digitális, informatikai eszközökkel és informatikai hálózatokon

megvalósuló megoldások felé. A digitális adatok jelentőségének növekedésével párhuzamosan folyamatosan fejlődnek azok a módszertani megoldások, a digitális adatok felhasználását támogató szervezeti megoldások, adatkormányzási gyakorlatok, amelyek használatával a különböző szervezetek a lehető legjobb módon, hatékonyan képesek használni az általuk gyűjtött, tárolt digitális adatokból keletkező adatbázisokat, adatvagyont. A digitális adatok menedzselését szolgáló megoldások, módszertanok építenek a tudás – és információmenedzsment területére is, amely szélesebb kitekintést, tudásbázist jelent.

Az előadás a szervezeti működés szempontjait támogató módon a digitális adatok kezelésén belül több témakört kíván röviden áttekinteni, szélesebb kontextusba helyezve a kutatási adatok kezelésén kívül általában a digitális adatok kezelésének fontosságát, a digitális adatvagyon felhasználását jelentő megoldásokat. A konferencia előadás részét is képező példák közé sorolható az adatkormányzás (Data Governance) folyamat és módszertan, a digitális adatok teljes életútjának menedzsmentje (Data Lifecycle Management) az adatok gyűjtésétől egészen a megsemmisítésükig, vagy a digitális adatbázisok kapcsolata a mesterséges intelligencia fejlesztésekkel, miért kritikus fontosságúak a digitális adatok a mesterséges intelligencia alkalmazások esetében. Az előadás során gyakorlati példákat is tervezek bemutatni.

Ahogyan a fentiekből látható, a kutatás és az előadás témája nem kizárólag a digitalizált vagy digitális kutatási adatokkal foglalkozna, hanem általában a különböző szervezetekben kezelt digitális adatokkal, emiatt indokoltnak tartom röviden összefoglalni, a konferencia témájához milyen módon kapcsolódik. A HUN-REN Adatrepozitórium Platform és Adatgazdász Hálózat tevékenységéhez több módon is szorosan kapcsolódik az előadás: például a Platform digitális adatokat kezel és a kutatási intézmények, a kutatások jelentős részben, hasonlóan az általános tendencia jellegéhez, egyre nagyobb mértékben támaszkodnak a digitalizált, digitális adatok kezelésére, elemzésére és ezen adatok menedzsmentje, kezelése szintén egyre növekvő fontosságú feladat. Az előadás ezért általában is érdekes lehet a digitális adatok kezelésével foglalkozó kutatóknak, munkatársaknak.

Tóth Zoltán (HUN-REN SZTAKI): RocKIT: RO-Crate alapú kutatási adatcsomagok kezelése a kutatói munkaasztaltól a repozitóriumig

A kutatási adatok hosszú távú értékét nem önmagában a fájlok megőrzése, hanem azok értelmezhető, kereshető és újrafelhasználható leírása biztosítja. A FAIR irányelvek és a FAIR Digital Object szemlélet erre kínálnak gyakorlati keretet: az adatokhoz perzisztens azonosítók, géppel is feldolgozható metaadatok és a felhasználást támogató kontextus kapcsolódik. Az RO-Crate technológia ennek egyik könnyen hordozható megvalósítása, amely egy adatcsomag fájljait, könyvtárait, kapcsolódó személyeit, szervezeteit, szoftvereit és egyéb kontextuális elemeit JSON-LD alapú leírásban kapcsolja össze.

Az előadás a HUN-REN ARP Adatrepozitóriumban használt AROMA komponens és az abból továbbfejlesztett RocKIT alkalmazás összehasonlításán keresztül mutatja be, hogyan tehető a gazdag metaadatolás a kutatói munka természetes részévé. Az AROMA a Harvard Dataverse alapú ARP repozitórium szerveroldali szolgáltatásaként alternatív nézetet és szerkesztőfelületet ad a már feltöltött adatcsomagokhoz, különösen fájl- és könyvtárszintű RO-Crate metaadatok kezelésére. Erőssége, hogy közvetlenül a repozitóriumi környezetben kapcsolja össze az adatcsomag alapvető idézési metaadatait, hierarchikus szerkezetét és a hozzá társított metaadatsémákat.

A RocKIT ezzel szemben a metaadatolás időpontját és helyét hozza közelebb a kutatók mindennapi munkájához. Asztali alkalmazásként a felhasználó saját gépén, a lokális fájlrendszerben kezelt mappákból hoz létre és szerkeszt RO-Crate adatcsomagokat. A kutató már az adat-előkészítés során áttekintheti a fizikai fájlokat és az RO-Crate entitásokat, fájlokat és beágyazott adatcsomagokat kapcsolhat a leíráshoz, schema.org tulajdonságokat és intézményi metaadatsémákat rendelhet az entitásokhoz, valamint több fájl vagy adatcsomag közös jellemzőit tömegesen is szerkesztheti. A RocKIT így nemcsak egy adott repozitóriumi feltöltés előszobája, hanem hordozható adatcsomag-készítő környezet is: ugyanaz a gondosan előkészített RO-Crate több közzétételi célra újrahasznosítható, akár különböző Dataverse alapú rendszerekben, az ARP Adatrepozitóriumban vagy más szolgáltatásokban, például Zenodo-ban. Ez különösen hasznos lehet olyan kutatási projektekben, ahol az adatokat intézményi, tudományterületi és nemzetközi repozitóriumokban is láthatóvá kell tenni.

Bemutatjuk, hogy a RocKIT miként egészíti ki az AROMA repozitóriumi szerepét: hogyan készíti elő és terjeszti ki a repozitálási munkafolyamatot, hogyan segítheti a kutatókat adatcsomagok összeállításában, dokumentálásában, sémakövető metaadatolásában és több repozitóriumot célzó exportjában, valamint hogyan jelenhet meg kontrollált módon az MI-támogatás a metaadatok elkészítésének folyamatában.

Kaján Győző (HUN-REN ÁTK): Robotoljon a robot! – Mesterséges intelligencia az adatgazdászok szolgálatában

Az adatgazdászi munka egyik leginkább idő- és munkaigényes feladata az adatcsomagok részletes metaadatolása. Ennek minősége ugyanakkor alapvetően meghatározza a deponált kutatási adatok megtalálhatóságát, értelmezhetőségét és későbbi újrafelhasználhatóságát, ezért elengedhetetlen a metaadatok pontos és minél teljesebb rögzítése. Az adatcsomagok újrahasznosíthatósága szempontjából szintén fontos a gyakran alkalmazott Excel-táblázatok gépi feldolgozhatóságának javítása, valamint nyíltabb, hosszú távú megőrzésre alkalmas CSV- (comma-separated values) formátumba történő konvertálása. Mindezek mellett szükséges az adatcsomag tartalmának közérthető, „ember által olvasható” dokumentálása is, jellemzően README-fájl formájában.

A generatív mesterséges intelligencia gyors fejlődése lehetőséget kínál e fáradságos és időigényes munkafolyamatok részleges automatizálására. Ennek vizsgálatára három, egymást kiegészítő promptot fejlesztettem. Az első prompt az Excel- és PDF-formátumú táblázatok gépi feldolgozhatóságát javítja: eltávolítja a strukturált adattáblán kívüli címeket és lábjegyzeteket, elvégzi a szükséges karaktercseréket – például a cellákon belüli vesszők pontosvesszőre cserélését –, majd a táblázatokat CSV-formátumba konvertálja. A második prompt az adatcsomag tartalmát és kutatási kontextusát bemutató README-fájlt készíti el Markdown-formátumban. A harmadik prompt a RocKIT RO-Crate-szerkesztő mesterségesintelligencia-alapú felületén támogatja az adatcsomag részletes metaadatolását.

Az első két promptot az OpenCode szoftverkörnyezetben futtattam. A mesterségesintelligencia-modelleket elsősorban a HUN-REN Cloud által biztosított GenAI4Science szolgáltatáson keresztül, API-kulcsos hitelesítéssel értem el. Összehasonlításképpen a ChatGPT és a Claude térítésmentesen hozzáférhető szolgáltatásait is teszteltem. A metaadatoló promptot közvetlenül a RocKIT mesterségesintelligencia-felületén alkalmaztam, amely a ChatGPT Windows 11-en futó alkalmazásához kapcsolódott.

A három prompt segítségével 2026 júliusában hat adatcsomagot dolgoztam fel a HUN-REN Adatrepozitórium Platformba történő feltöltés előtt. A GenAI4Science szolgáltatásban elérhető modellek közül leggyakrabban a Google DeepMind Gemma4:31b, illetve a Qwen3.6:27b modellt használtam. Tapasztalataim szerint a Gemma4:31b rendszerint gyorsabban hajtotta végre a feladatokat, míg a Qwen3.6:27b többféle fájlformátum megnyitására és szerkesztésére volt képes. Egy-egy teszt során a ChatGPT és a Claude is jó minőségű eredményeket szolgáltatott. A ChatGPT ismételt tesztelését ugyanakkor egy szigorú használati időkorlát jelentősen megnehezítette. A RocKIT hatékonyan támogatta az adatcsomagok metaadatolását és a HUN-REN Adatrepozitórium Platformba történő feltöltését.

Az eredmények alapján a generatív mesterséges intelligencia jelentősen támogathatja az adatgazdászok munkáját, és számottevően csökkentheti a rutinszerű, időigényes feladatokból eredő terhelést. A kidolgozott promptok ugyanakkor még nem teszik lehetővé a munkafolyamatok teljes automatizálását. Továbbra is szükség van az adatcsomag szakértői kurálására, a fájlok tartalmának és a fájlnevek jelentésének megismerésére, valamint a promptok adott adatcsomaghoz történő hozzáigazítására. Elengedhetetlen továbbá az elkészült eredmények alapos emberi ellenőrzése, mivel a mesterségesintelligencia-modellek hallucinációi, téves következtetései és technikai hibái jelenleg még nem zárhatók ki teljes bizonyossággal.

Vereckei András (ELTE KRTK): Hogyan alakítják át a nagy nyelvi modellek és a rájuk épülő eszközök a kutatást és a kutatási adatok repozitálását?

A nagy nyelvi modellek (LLM-ek) alapvetően alakítják át a kutatás minden szakaszát: a szakirodalom-feldolgozástól a chat-alapú munkán át a kódírásig, a metaadatok előállításáig és egészen az adatok repozitálásáig. Előadásom azt mutatja be, hogyan változtatja meg a nagy nyelvi modellek használata a kutatók napi munkáját, és milyen lehetőségeket, illetve kihívásokat rejt ez az adatrepozitóriumok és a kutatási adatok hosszú távú fenntarthatóságára és biztonságára nézve.

Az elmúlt években az LLM-ek a mindennapi kutatómunkában is használhatóvá váltak, 2025-ben pedig megjelentek az ügynökalapú (agent) kódolóeszközök és a skill-ek, amelyek a modelleket szervezett, feladatra szabott munkafolyamatokba ágyazzák

Előadásom középpontjában négy kérdés áll.

Első: milyen alapvető fogalmak jellemzik az LLM-ek és az ügynökök munkáját, és hogyan használom őket a kutatói munkámban? A nagy nyelvi modellek szöveges bemenetek (prompt) alapján állítanak elő szöveget, az ügynökök (agent) pedig a modellt eszközökkel és feladatra szabott munkafolyamatokkal ruházzák fel. Az ügynökök nem csupán válaszolnak, hanem futtatnak, fájlokat írnak és tesztelnek. Az előadás során ezt a VS Code Copilot és a Hermes Agent segítségével, OpenRouteren elérhető modellekkel mutatnám meg.

Második: mi a repozitálás gyakorlata a mindennapi munkám során? Ezt a saját eszközünk, a Bead (github.com/bead-project) nyílt forráskódú adatproveniencia-eszköz segítségével mutatnám be. A Bead a kívánt tartalmakat önálló, kriptográfiai hash-sel ellátott zip-csomagokként (Bead-ekként) menti el, a Bead-eket pedig lokális boxokba tárolja.

Egy zip-fájlba tömörítve mentésre kerülnek az adatbemenetek, a kódok és az általuk létrehozott kimenetek. Így bármikor visszakövethetővé válik, hogy melyik kód melyik input segítségével milyen outputot állított elő. A Bead-ek összekapcsolhatóak: az egyik Bead outputja a következő inputja lehet. Az így létrehozott "Bead chain"-eket grafikusan is lehet ábrázolni.

A Bead nem váltja ki az ARP-ot, de megoldást kínál az adatok lokális verziókövetésére: a megosztott repozitálás és közzététel a kutatói közösség számára az ARP-ban történik.

Harmadik: hogyan segítik az LLM-ek és az ügynökök az adatrepozitálást a mindennapi munkában? Hogyan könnyítik meg a metaadat-sémák generálását, a README és a Makefile készítését, a kísérő dokumentáció összeállítását, valamint a feldolgozó kódok unit-tesztelését. Az előadásban bemutatnám, hogy jelenleg hol tart ez a technológia és hol vannak a korlátai.

Negyedik: mi lehet a repozitálás jövője? Hogyan lehet alkalmazkodni a folyamatosan változó technológiai és szoftveres kihívásokhoz?
Várhatóan az egyedi ügynökökön túlmutató, egymással együttműködő ügynökrendszerek (swarm és fleet), a munkafolyamatokat összefogó agent harness-ek, valamint az ezekkel járó kiberbiztonsági kihívások alakítják majd a jövőt. Különösen érdekes lehet a védett, kutatószobai infrastruktúrák és az érzékeny adminisztratív mikroadatok esetében a lokális GPU-n futtatott LLM-ek alkalmazása. Az LLM-ek és az ügynökök új támadási felületet is jelenthetnek, ezért a biztonságos repozitálásnak a technológiai innovációval párhuzamosan kell fejlődnie.

Összegezve az előadásom gyakorlati esettanulmány arról, hogyan segíti a Bead az adatfeldolgozás közben létrehozott helyi, reprodukálható adatcsomagok és a hozzájuk tartozó proveniencia-metaadatok előkészítését, és egyben stratégiai helyzetelemzés arról, hogyan maradhatnak az adatrepozitóriumok hosszú távon is elérhetőek, biztonságosak és fenntarthatóak a gyorsan változó mesterséges intelligencia által formált környezetben.

4. Szekció: Nemzetközi körkép

Pompor Zoltán (Pro-M Zrt): A bővülő EOSC nyújtotta lehetőségek

Az előadás az EOSC Gravity projektet mutatja be, amely meghatározó szerepet játszik az EOSC öderáció megszilárdításában és a 2027 utáni fenntartható, egységes európai nyílt tudományos infrastruktúra irányítási és finanszírozási modelljének előkészítésében. A projekt fő pillérei közé tartozik az európai koordináció és az INFRAEOSC projektek közötti szinergiák erősítése, az EOSC éves rendezvényeinek lebonyolítása, EOSC Szövetség (EOSC Federation) továbbfejlesztése, az EOSC-csomópontok (EOSC Nodes) létrehozásának ösztönzése által. Továbbá az EOSC Gravity kiemelt hangsúlyt helyez a közösségépítésre, a képzési és tudásátadási tevékenységekre, az EOSC Academy működtetésére, képzéskínálatának fejlesztésére.

A prezentáció kiemelten foglalkozik az EOSC Föderáció bővülésével és az új EOSC Node-ok létrejöttével, amelyeket a projekt nyílt pályázati felhívásokkal is támogat. EOSC Föderációja olyan elosztott architektúrára épül, amelyben az egyes, egymástól függetlenül működő szervezetek és kutatási infrastruktúrák az úgynevezett EOSC Node-okon keresztül kapcsolódnak egymáshoz. Az EOSC Node-ok szabványos interfészeket, közös azonosítási és jogosultságkezelési (AAI) mechanizmusokat, interoperabilitási követelményeket, valamint egységes metaadat- és szolgáltatáskezelési megoldásokat biztosítanak annak érdekében, hogy a kutatók az adatokhoz, számítási kapacitásokhoz és kutatási szolgáltatásokhoz egységes módon férhessenek hozzá.

Az EOSC Föderáció működése decentralizált megközelítést követ: az egyes csomópontok megőrzik saját működési autonómiájukat, miközben a közösen meghatározott műszaki, szemantikai, szervezeti és jogi szabályoknak megfelelően biztosítják a szolgáltatások összekapcsolását. Ez a modell lehetővé teszi az EOSC fokozatos bővítését, a nemzeti, regionális és tematikus kutatási infrastruktúrák integrálását, valamint a FAIR (Findable, Accessible, Interoperable, Reusable) alapelvek európai szintű érvényesítését.
Az előadás célja, hogy részletesen bemutassa a bővülési folyamatban rejlő konkrét lehetőségeket, különös tekintettel a magyar kutatási infrastruktúrák, intézmények és felhasználók bekapcsolódási pontjaira.

Fazekas-Paragh Judit (DEENK): Nemzetközi kapcsolódások az EOSC és OpenAIRE ökoszisztémában: infrastruktúrák és kompetenciák a FAIR kutatási adatok szolgálatában

A kutatási adatok megfelelő kezelése, megosztása és újrahasznosíthatósága az Open Science egyik meghatározó pillérévé vált. A FAIR alapelvek ugyanakkor csak akkor érvényesülhetnek a gyakorlatban, ha megfelelő technológiai infrastruktúrák, intézményi szolgáltatások és kutatástámogató szakemberek együttesen állnak a kutatók rendelkezésére. Az European Open Science Cloud (EOSC) ennek az európai szintű, föderált kutatásiadat-ökoszisztémának a kialakítását célozza, amelyhez az OpenAIRE infrastruktúrája és szolgáltatásai több ponton kapcsolódnak.

Az OpenAIRE a kutatási adatok teljes életciklusát támogató szolgáltatási környezetet épített ki. Az ARGOS az adatkezelési tervek (DMP) elkészítését és géppel feldolgozhatóvá tételét segíti; a Zenodo lehetőséget biztosít kutatási adatok és más kutatási eredmények megőrzésére, azonosítására és megosztására, az OpenAIRE Graph pedig a publikációk, kutatási adatok, szoftverek, projektek, finanszírozók és intézmények közötti kapcsolatok feltárásával járul hozzá a kutatási eredmények nemzetközi láthatóságához és újrahasznosíthatóságához. Az OpenAIRE interoperabilitási irányelvei és szolgáltatásai emellett lehetővé teszik, hogy a nemzeti és intézményi infrastruktúrák bekapcsolódjanak az európai Open Science és EOSC ökoszisztémába.

A technológiai infrastruktúra azonban önmagában nem elegendő a FAIR kutatási adatkezelés megvalósításához. Egyre nagyobb jelentősége van azoknak a szakembereknek, adatgazdászoknak, data stewardoknak és kutatástámogató könyvtárosoknak, akik összekötő szerepet töltenek be a kutatók, az intézményi infrastruktúrák, a szabályozási környezet és a nemzetközi szolgáltatások között. Képzésük ezért nem korlátozódhat egyes eszközök használatának elsajátítására. A kompetenciafejlesztésnek ki kell terjednie a kutatási adatok életciklusára, a FAIR alapelvekre, a metaadatokra és tartós azonosítókra, a repozitóriumok használatára, a licencekre, az érzékeny adatok kezelésére, valamint a reprodukálhatóság és a kutatás integritás kérdéseire is.

Az előadás azt mutatja be, hogyan kapcsolhatók össze az OpenAIRE és az EOSC infrastruktúrái a kutatási adatok kezelésének mindennapi gyakorlatával, és hogyan fordíthatók le ezek a lehetőségek a magyar kutatók számára ténylegesen használható szolgáltatásokká. Kiemelt figyelmet kap az adatgazdászok képzése és az a kompetenciaalapú megközelítés, amelyben a data steward már nem pusztán az adatkezelés technikai támogatója, hanem a FAIR és felelős kutatási gyakorlat intézményi közvetítője, valamint a helyi és az európai kutatásiadat-ökoszisztéma közötti egyik legfontosabb kapcsolódási pont.

Sipos Gergely (EGI Foundation): Az EOSC hiányzó láncszeme: Adatrepozitóriumok és alkalmazások interoperabilitása az EOSC Data Commons segítségével

While the volume of preserved research data and analysis software has grown significantly over the last decade, much of this valuable asset remains siloed in national or thematic repositories, in heterogeneous data and metadata formats that make dataset integration difficult. As a consequence, valuable scientific insights remain unexplored as researchers should spend increasing amounts of time with data discovery, interpretation, formatting and transformation. EOSC Data Commons (EDC) is building the next-generation federated research data infrastructure to eliminate cross-repository and cross-format data fragmentation.

At the centre of the EDC technical architecture there are two complementary services, designed to bridge disconnected systems:

1. EOSC Matchmaker: a discovery layer enabling researchers to discover datasets, software tools, and execution services across multiple scientific domains and EOSC Nodes. It operates alongside a catalogue of analytics tools that can be deployed close to the data, significantly reducing data movement and improving processing efficiency.

2. EOSC Data Player: An execution platform integration service that provides harmonised APIs, shared metadata specifications, and common mechanisms for authentication, authorisation, and provenance tracking. This ensures that researchers can combine data and software tools from different providers without manually reconciling metadata formats or security protocols.

To achieve scalable interoperability, the project uses a range of advanced semantic technologies, including knowledge graphs for rich metadata representation and standardised interfaces to support cross-infrastructure integration.
EOSC Data Commons adopts community standards such as RO-Crate for packaging research artefacts and their metadata in a machine-actionable, highly interoperable manner.
This technical stack is continuously validated through real-world use cases spanning diverse scientific domains. For data repositories, the technical frameworks established

by EOSC Data Commons offer a highly practical blueprint. Testing EDC's technological ecosystem through its open call, national repositories, such as ARP, can integrate with the broader European Open Science Cloud and expand their reach into new scientific communities, while maintaining their operational autonomy and data security.

This presentation outlines the core technological solutions, semantic standards, and interoperability frameworks developed by EDC to enable seamless, cross-disciplinary data reuse. The talk aims to generate interest in the Hungarian ARP community to join the growing network of EDC repositories.

Faragó-Szilvási Tibor (MTA KIK): Az EU kutatásbiztonsági ajánlásai és ezek lehetséges hatása az adatkezelési tervek tartalmi bővülésére.

Az Európai Unió kutatás- és innovációs politikájában az elmúlt években egyre nagyobb hangsúlyt kapott a kutatásbiztonság kérdése. Ez a folyamat azonban nem a nyílt tudomány és a nemzetközi kutatási együttműködések visszaszorítását célozza. A 2024-es Council Recommendation on enhancing research security egyértelműen rögzíti: „Openness, international cooperation, and academic freedom are at the core of world-class research and innovation". Ugyanakkor a növekvő geopolitikai kockázatok miatt szükségessé vált a nemzetközi együttműködések „újra kiegyensúlyozása", valamint olyan „megfelelő és arányos védelmi intézkedések" alkalmazása, amelyek a kutatási együttműködéseket egyszerre teszik nyitottá és biztonságossá.

A kutatásbiztonsági megközelítés nem általánosan a kutatás minden területére vonatkozik, hanem elsősorban bizonyos, kockázatokkal érintett kutatási területekre és különösen a nemzetközi együttműködésekre fókuszál. Az akadémiai szabadság továbbra is magában foglalja a kutatási partnerek szabad megválasztását, ugyanakkor e szabadság „akadémiai felelősséggel" kell párosulnia. Ennek megfelelően a cél nem a nemzetközi együttműködés korlátozása, hanem a kockázatok arányos azonosítása és kezelése.

Ez a szemlélet a következő Horizon Europe ciklusban várhatóan még hangsúlyosabbá válik. A 2021–2027-es programban a FAIR, az Open Science és a felelős adatmegosztás meghatározó keretet adnak a kutatási adatok kezelésének. A 2028–2034-es ciklus (FP 10) jelenlegi jogalkotási folyamatában ehhez egyre explicitebben társul a kutatásbiztonság, a kockázatok azonosítása, értékelése és mérséklése, valamint a „dual-use" és a kapcsolódó „safeguards" kérdésköre. A Tanács 2026. június 26-i álláspontja már kifejezetten egy kockázatalapú és arányos keretrendszert irányoz elő.

Mindez új kérdéseket vet fel a kutatási adatok kezelésében is. Az „as open as possible, as closed as necessary" elv továbbra is érvényes, de a kutatásbiztonsági megközelítés bizonyos kutatási területeken új dimenzióval egészíti ki annak gyakorlati alkalmazását: a

hozzáférés és adatmegosztás mellett megjelenik a kutatási kockázatok előzetes értékelésének szükségessége.

Az előadás végén felvetődnek kérdések:

1) A jelenlegi adatkezelési terv (DMP) sablonok és adatkezelési folyamatok mennyiben támogatják ezt a megközelítést?

2) Szükség lehet-e a közeljövőben egy kutatásbiztonsági kockázatelemzési komponens beépítése a DMP-kbe, különösen nemzetközi együttműködések esetén?

3) Ez a változás hogyan bővítheti az adatgazdászok jelenlegi szerepét: a FAIR és Open Science támogatása mellett az adatgazdászok milyen mértékig tudják a jövőben a kockázatalapú és felelős adatmegosztást támogatni?