

ARP kutatási adatrepozitórium platform

Kovács László

SZTAKI DSD Elosztott Rendszerek Osztály

laszlo.kovacs@sztaki.hun-ren.hu

1994 -2024 (30 év)

Mode 2 tudástermelés

27 EU Framework projekt

10 nagy ipari projekt
(T-Mobile, Telekom, RICOH, ...)

Művelt szakmai területek:
magyar web kezdetei
digitális archívumok, repozitóriumok
kollaboráció támogatás, groupware



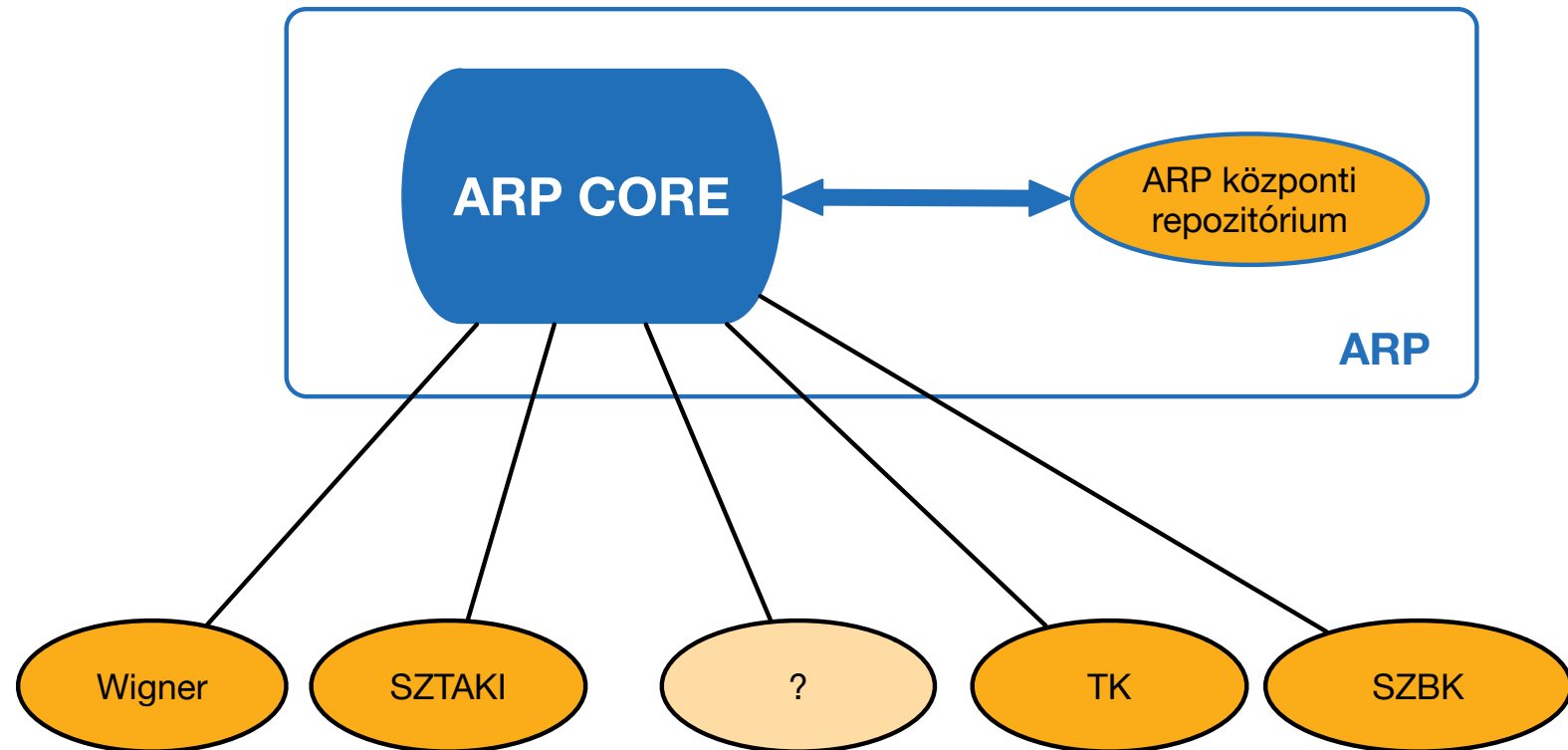
- MIÉRT javasoltuk az ARP projektet:
 - (kutatási) adatok megbízható és hosszú távú digitális tárolása 15-20 éve megoldatlan probléma a HUN-REN-ben (Magyarországon)
 - az adatok (újra) felhasználása nem lehetséges
 - ad-hoc tárolási megoldások, adatvesztés, adatélettartam rövidülés
 - (kutatási) adatok kereshetősége nem optimális
 - F.A.I.R. elvárások (machine actionable = informatikai elvárásrendszer)
 - MI (mesterséges intelligencia) viharos megjelenése
 - “az adat mint érték, adat mint erőforrás”

Hiányzik egy egyszerűen használható értékelési, értékképzési módszertan, mely alapján megbecsülhetővé válna a kutatási adat, adatállomány (pénzben is kifejezhető) közelítő értéke.

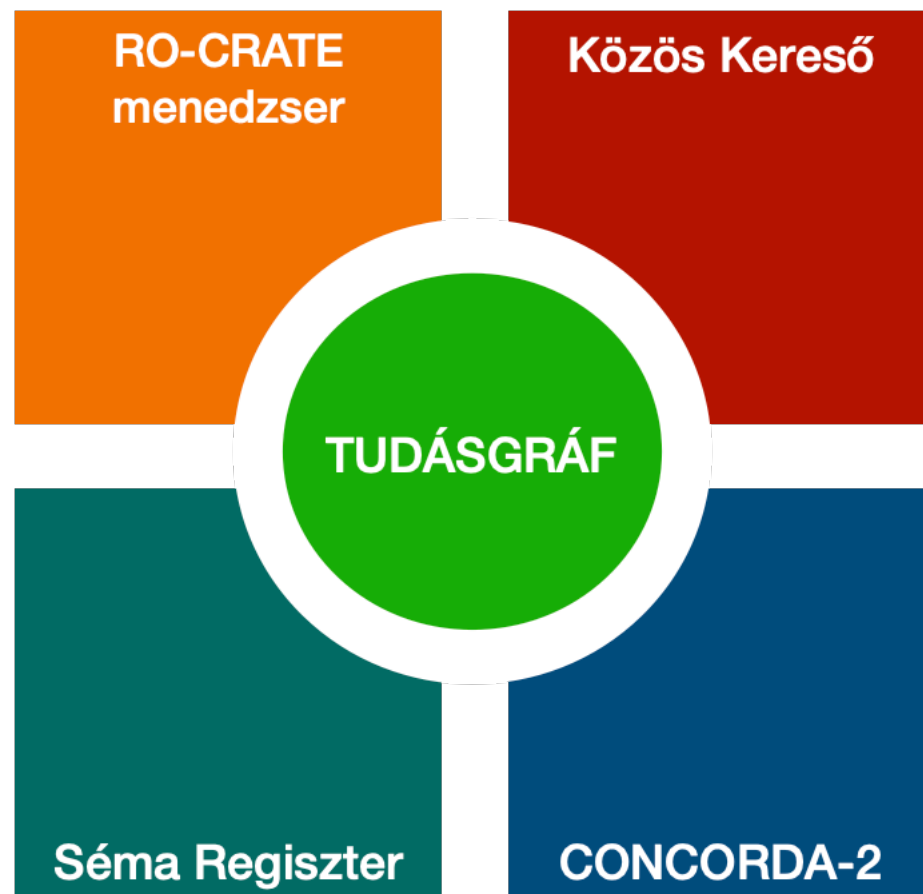
(Kutatók, intézeti vezetők kvázi “vakon repülnek”.)

- **CÉL:** Kutatási adatok megbízható és hosszú távú tárolása
 - egyértelmű azonosítása
 - megőrzése
 - megosztása
 - újrafelhasználásuk elősegítése...
- **HOGYAN:** Föderált (együttműködő) HUN-REN kutatási adatrepozitórium hálózat
 - országos hálózati adat-infrastruktúra az intézeti és ágazati repozitóriumok bevonásával
 - ARP-CORE szolgáltatások kifejlesztése
 - nemzetközi (adat)rendszerekhez való (adat)kapcsolatok kifejlesztése

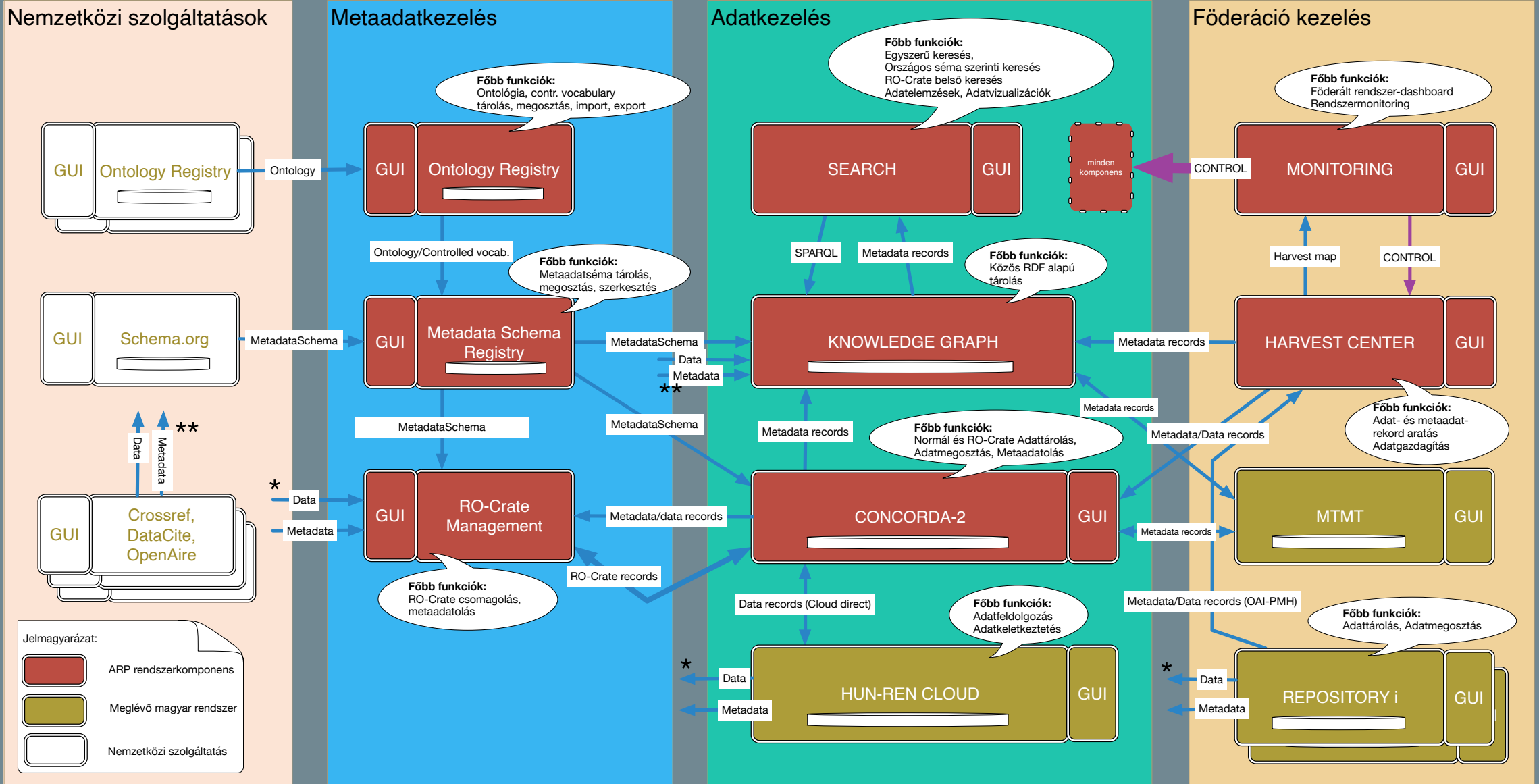
Föderált HUN-REN kutatási adatrepozitórium hálózat

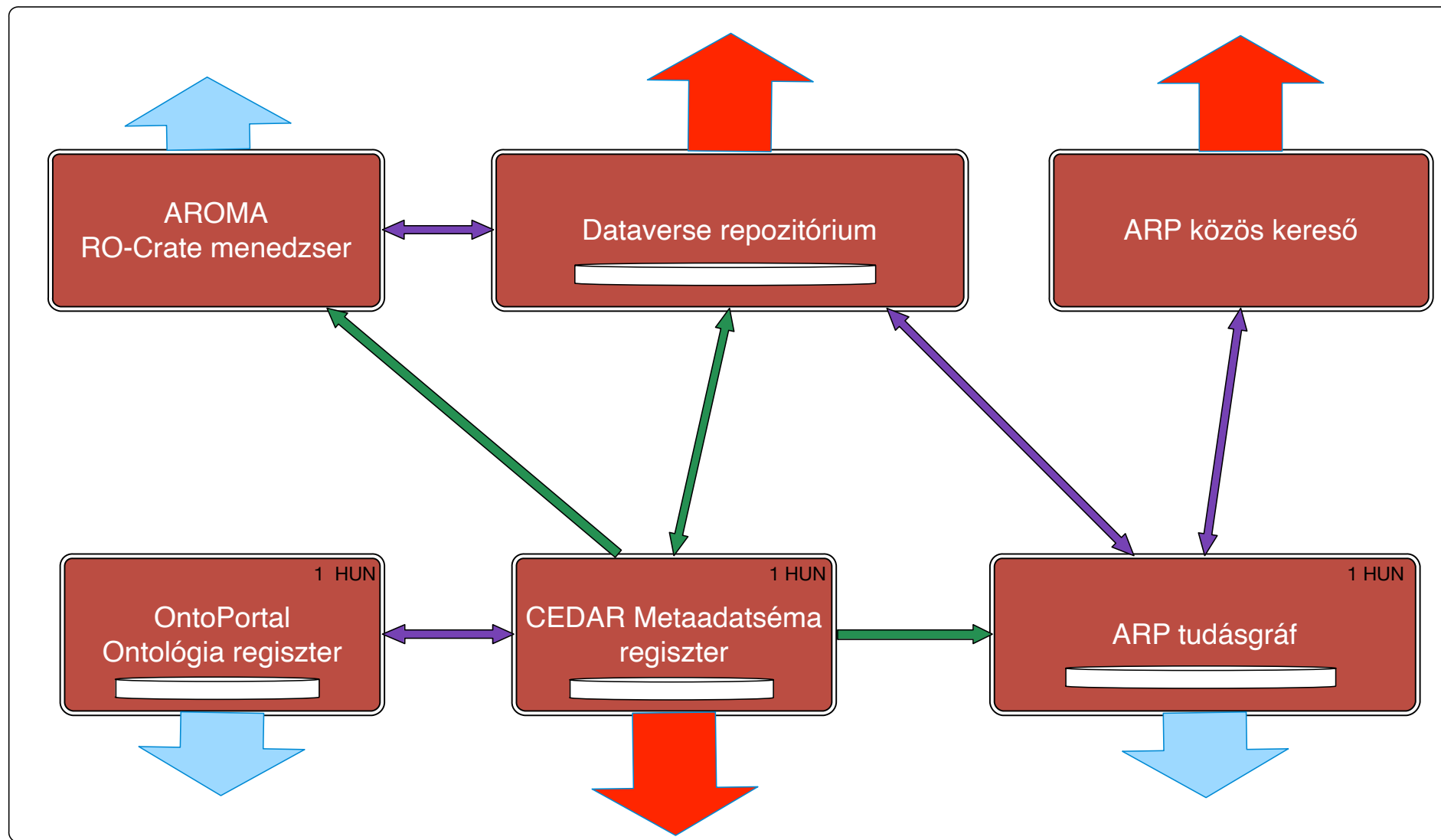


Nagy intézetek vagy ágazatok saját kutatási adatrepozitóriumokkal



- ARP CORE (központi szolgáltatások):
 - központi tároló szolgáltatás (Harvard: Dataverse)
 - metaadatséma regiszter (Stanford: CEDAR)
 - ontológia regiszter (Stanford: OntoPortal)
 - központi tudásgráf (Cornell: VIVO)
 - központi kereső (SZTAKI)
 - RO-Crate menedzser (SZTAKI: AROMA)





ARP AROMA - RO-Crate Manager

Adatcsomag: TTK dataset - 4 class Motor-imagery EEG

Tartalma: #11333ced791965-9e3-4841-8292-5a3661a6728 OR, #a12aa22-235a-490c-b701-2ac27aceab74 OR, #a98622a-7523-4fc1-890c-50a56323e1 OR

Licenz: <http://creativecommons.org/publicdomain/zero/1.0>

Publikálás dátuma: 2023-10-20T14:41:27.155795+02:00

RO-CRATE menedzser

Közös Kereső



ARP Unified Search

Search results: Reversible control of magnetism in FeRh thin films

ARP Schema Registry

Citation Metadata

Name: Citation Metadata, Version: 0.0.1

Title: The main title of the Dataset

Description: Az adatcsomag teljes címe

Séma Regiszter

CONCORDA-2

ARP Adatrepozitórium

CONCORDA-2

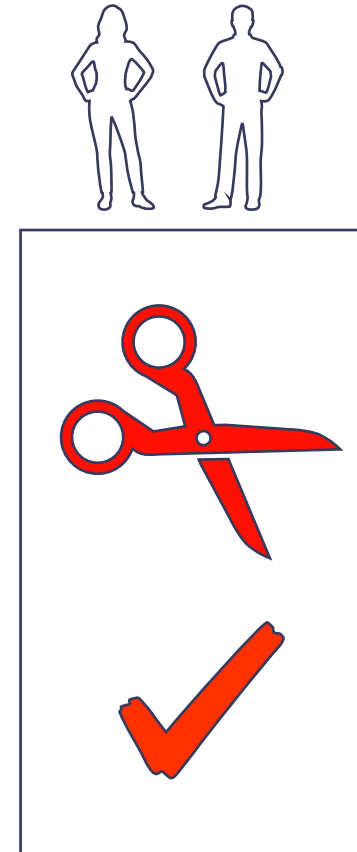
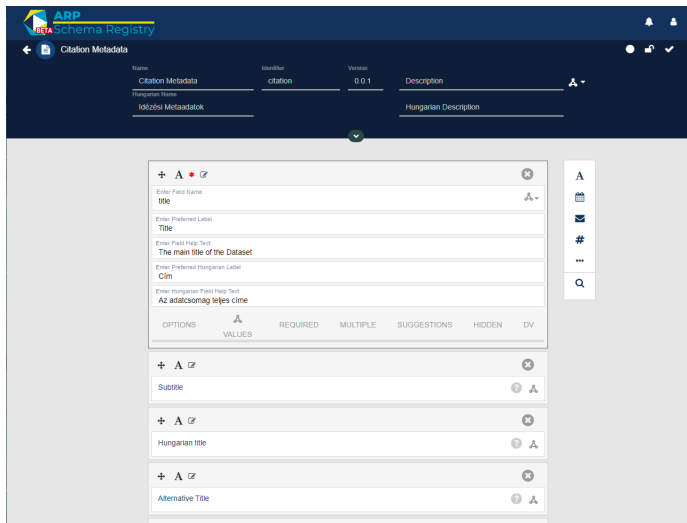
Dataverse of Eötvös Loránd Research Network

1 - 10 (1489) Esemény

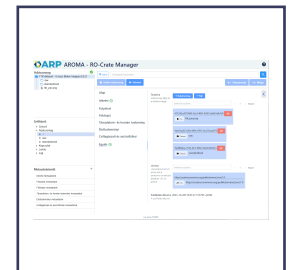
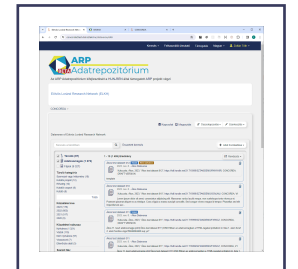
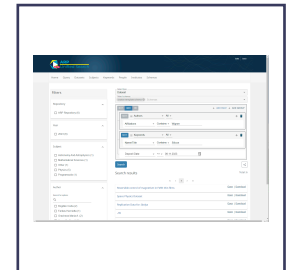
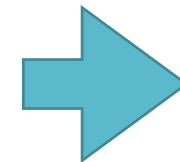
- Dataverse
 - adattárolás, adatmegosztás
- újratervezett metaadatséma kezelés
- ARP Cedar metaadatséma regiszter integrálás
- ARP AROMA RO-Crate menedzser integrálás

The screenshot displays the ARP Adatrepozitórium web interface. The header includes the ARP logo and the text 'Az ARP Adatrepozitórium kifejlesztését a HUN-REN által támogatott ARP projekt végzi'. Below the header, there is a search bar containing 'Eötvös Loránd Research Network (ELKH)'. The main content area shows search results for 'Ákos test dataset' with a list of items including 'Ákos test dataset 012', 'Ákos test dataset 010', 'Ákos test dataset 011', and 'Ákos test dataset 011'. Each item includes a title, date, and a brief description. The interface also features a sidebar with filters for 'Tárolók (87)', 'Adatcsomagok (1 372)', and 'Fájlok (9 327)'. The bottom of the page shows the 'Szerző Név' field.

- közmegegyezés létrehozását támogató szoftverrendszer
- Stanford CEDAR sémaregiszter
- OntoPortal (ontológia regiszter) integrált

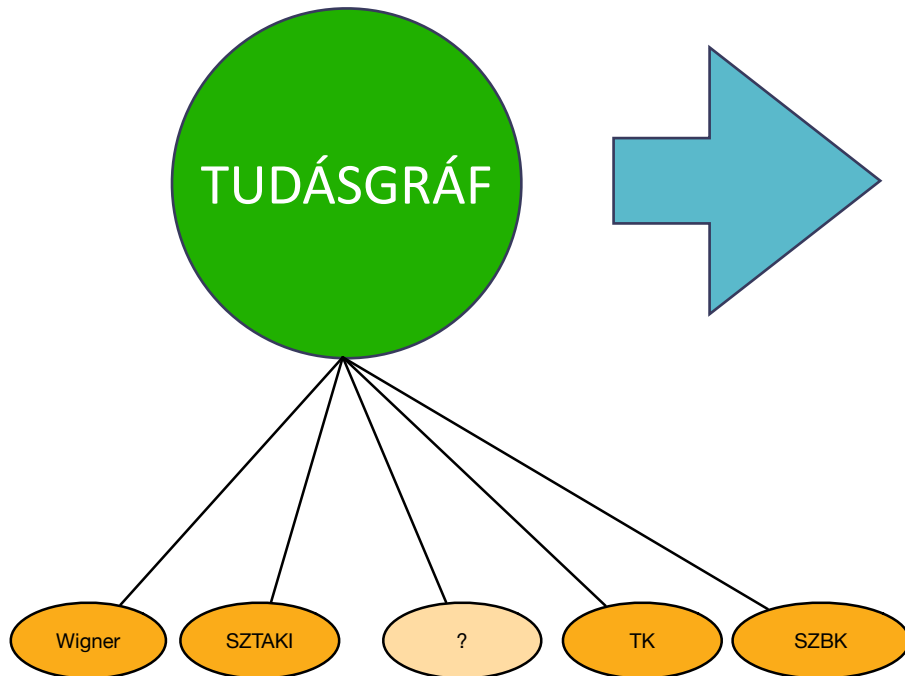


ARP Cedar



ARP CORE

Föderált hálózatban keres
Fájl szinten keres



A screenshot of the ARP Unified Search interface. The page features a navigation menu with links for Home, Query, Datasets, Subjects, Keywords, People, Institutes, and Schemas. The main content area is divided into several sections:

- Filters:** A sidebar on the left with expandable sections for Repository (ARP Repository (6)), Year (2023 (6)), Subject (Astronomy And Astrophysics (1), Mathematical Sciences (1), Other (1), Physics (5), Programozás (1)), and Author (Bogdán Csilla (2), Farkas Henrietta (1), Gracheva Maria A. (2)).
- Select type:** A dropdown menu set to "Dataset".
- Select schemas:** A dropdown menu set to "citation template schema".
- Search Rules:** A section with "NOT AND OR" operators and "ADD RULE" / "ADD GROUP" buttons. It contains two rules:
 - Rule 1: NOT Authors All Affiliations Contains Wigner
 - Rule 2: NOT Keywords All Name/Title Contains Silicon
- Deposit Date:** A date range filter set to "09.11.2023".
- Search:** A blue button to execute the search.
- Search results:** A section showing "Total: 6" results. The first result is "Reversible control of magnetism in FeRh thin films" with "Open" and "Download" links. Other results include "Space Physics Dataset", "Replication Data for: Ibolya", and ".f6".

- Research Object (RO)
- RO-Crate technológiaalapú csomagképzés
- fájl szintű metaadatolás
- kutatási projekt, kutatási folyamat archiválás

The screenshot displays the ARP AROMA - RO-Crate Manager interface. The main content area shows a dataset named "TTK dataset - 4 class Motor-Imagery EEG" with a tree view on the left containing folders "raw" and "standardized", and a file "ttk_par.png". The "Entitások" (Entities) section is expanded to show the "Adatcsomag" (Dataset) entity, which includes a root entity and sub-entities for "raw" and "standardized". The "Metaadatsémák" (Metadata Schemas) section lists various metadata types such as "Idézési Metaadatok" (Citation Metadata), "Folyóirat metaadatok" (Journal Metadata), "Földrajzi metaadatok" (Geographic Metadata), "Társadalom- és humán tudomány metaadatok" (Social and Human Sciences Metadata), "Élettudományi metaadatok" (Life Sciences Metadata), and "Csillagászati és asztrofizikai metaadatok" (Astronomical and Astrophysical Metadata).

The "Tartalma" (Content) section displays a list of entities, including a file "ttk_par.png" and two datasets: "raw" and "standardized". The "Licenz" (License) section shows the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International license (CC BY-NC-SA 4.0) with the URL "http://creativecommons.org/publicdomain/zero/1.0". The "Publikálás dátuma" (Publication Date) is 2023-10-20T14:41:27.155795+02:00.

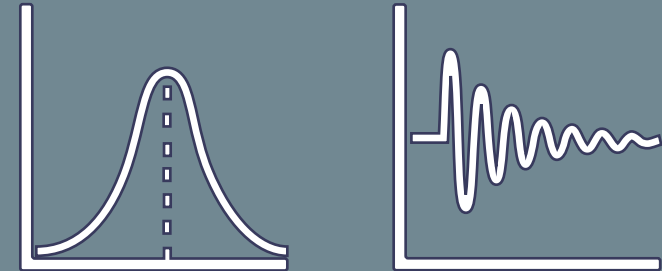
- Cél: A kutatási folyamat és eredményei
 - egyértelmű azonosítása
 - megőrzése
 - megosztása
 - újrafelhasználása ...
- Research Object (RO) - a tradicionális publikáció mellett - helyett
 - digitális leképezés, mely egy csomagban helyezi el a kutatási folyamat és annak körülményei valamint eredményei digitális reprezentánsait (digital resources)
 - csomagképzési technológia
 - digitális entitások
 - digitális entitások egymás közötti relációi

■ Adatok

- kiinduláskori kutatási adatok
- kutatás során keletkezett új adatok

■ Kutatási módszer és módszertan

- fizikai/digitális anyagok/környezetek és tevékenységek/műveletek leírása
- adatfeldolgozás informatikai környezete
- adatkeletkeztetés és adatfeldolgozás módszerei

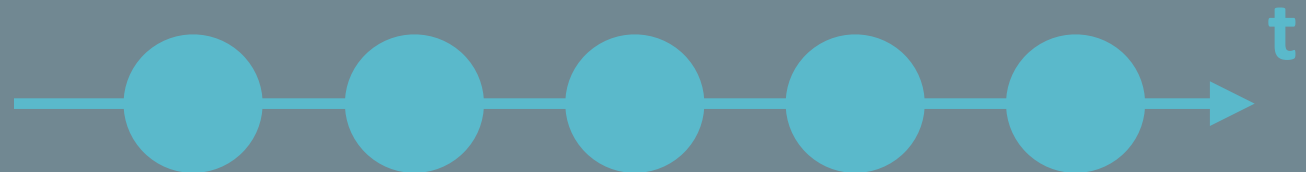
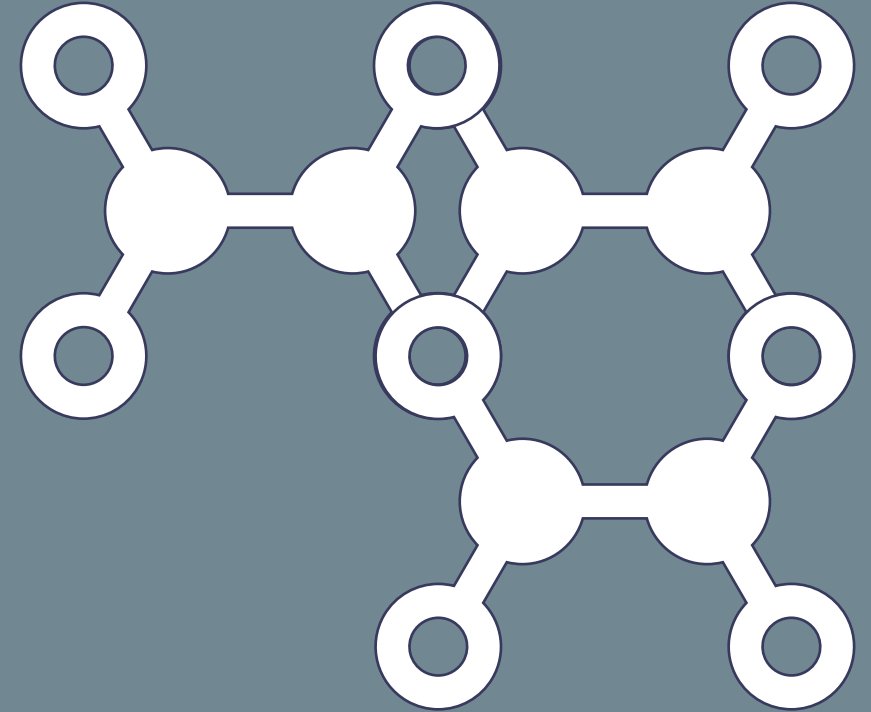


Fizikai tér

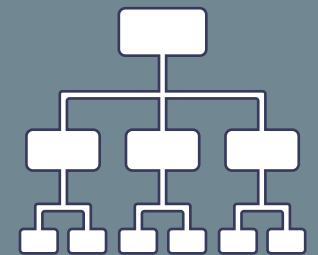
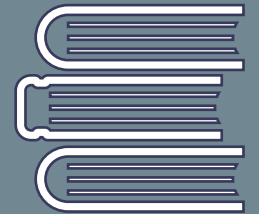
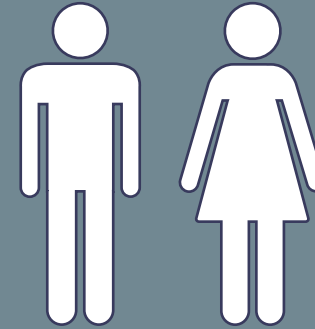


Digitális tér

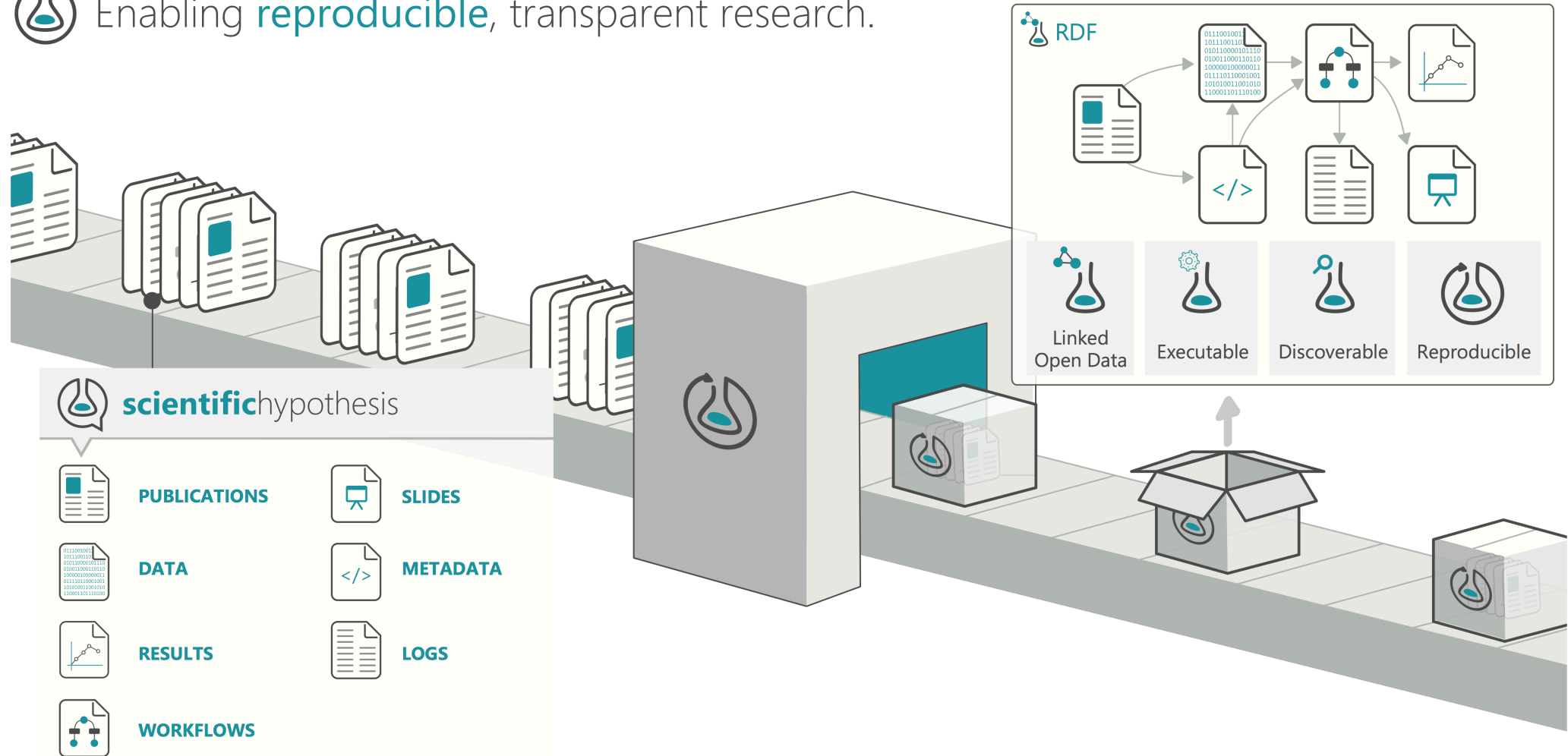
- Tudományos Workflow
 - a fizikai/adatfeldolgozási tevékenység tér-idő dinamikájának leírása
- Történetiség (Provenance)
 - történetiséget jellemző információk
 - a kutatási infrastruktúra használata
 - a használt műszerek, szoftverprogramok, infrastrukturális szolgáltatások beállítási paramétereinek történetisége ...



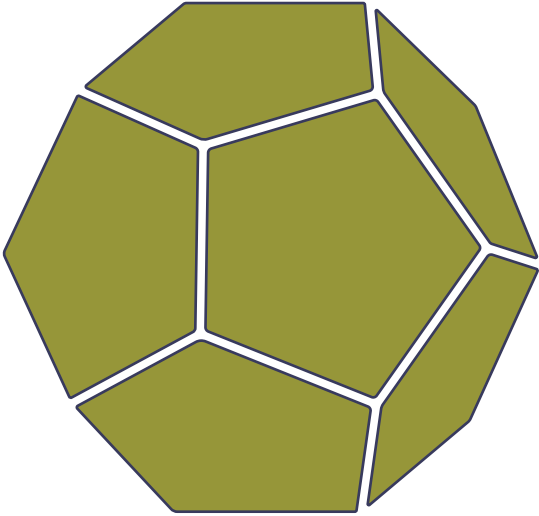
- Kutatók
 - egyértelműen azonosított (pl. ORCID) kutatók és szerepeik, hozzájárulás mértéke a kutatás során
- Eredmények
 - mérési jegyzőkönyvek, labnotes
 - publikációk, szöveges és/vagy multimédia projekt eredmények
 - disszeminációs dokumentumok, ...
- Értelmezések (annotációk)
 - entitások értelmezései (megértés és az egyértelmű szemantikus szintű interpretációik érdekében)
- Relációk
 - entitások relációi



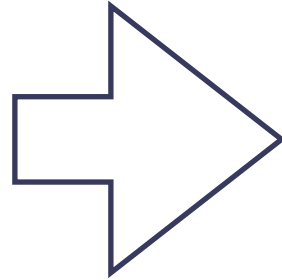
 Enabling **reproducible**, transparent research.



Forrás: <https://www.researchobject.org/images/research-objects-illustration.png>



RO-Crate



- saját felhasználás (elsődleges)
- kollaborációs (közösségi) felhasználás
 - adatmegosztás
 - adatpublikálás ...

	dokumentum (publikáció)	kutatási adatsomag
keresés	metaadat (pl. DC) alapján	specializált adatséma alapú metaadat alapján
megjelenítés	PDF (generikus) megjelenítő	csak szoftverrel (benne implicit értelmezés)
értelmezés	olvasás (emberi értelem)	csak szoftverrel (+ MI)

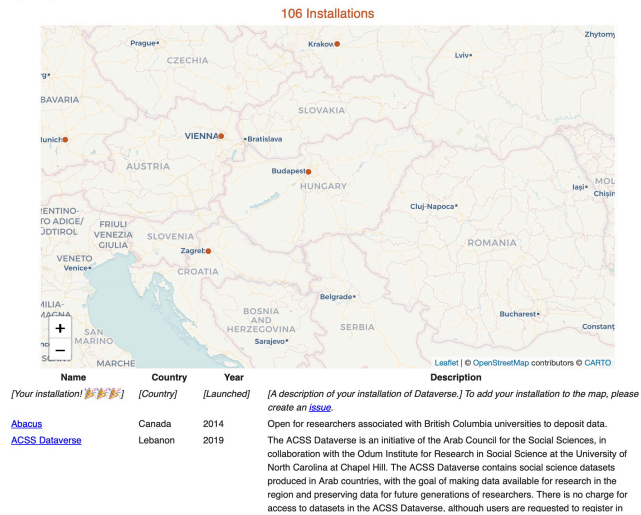
- Élő fejlesztői kapcsolatok:
 - Harvard Dataverse community
 - Stanford CEDAR community
 - Stanford OntoPortal community
 - Queensland University (Australia) (RO-Crate community)
- Nemzetközi kapcsolatok
 - RDA (Research Data Alliance), RDA Hungarian Node
 - ARP nemzetközi regiszterekben regisztrálva: FairSharing, Re3data, OpenDOAR, OpenAIRE
 - EOSC (European Open Science Cloud)
 - FAIR-IMPACT (Enabling FAIR Signposting and RO-Crate for content/metadata discovery and consumption”) EU projekt

ARP Dataverse (106.)

HOME / COMMUNITY /

Dataverse Installations Around the World

Open map in new window



HOME / COMMUNITY /

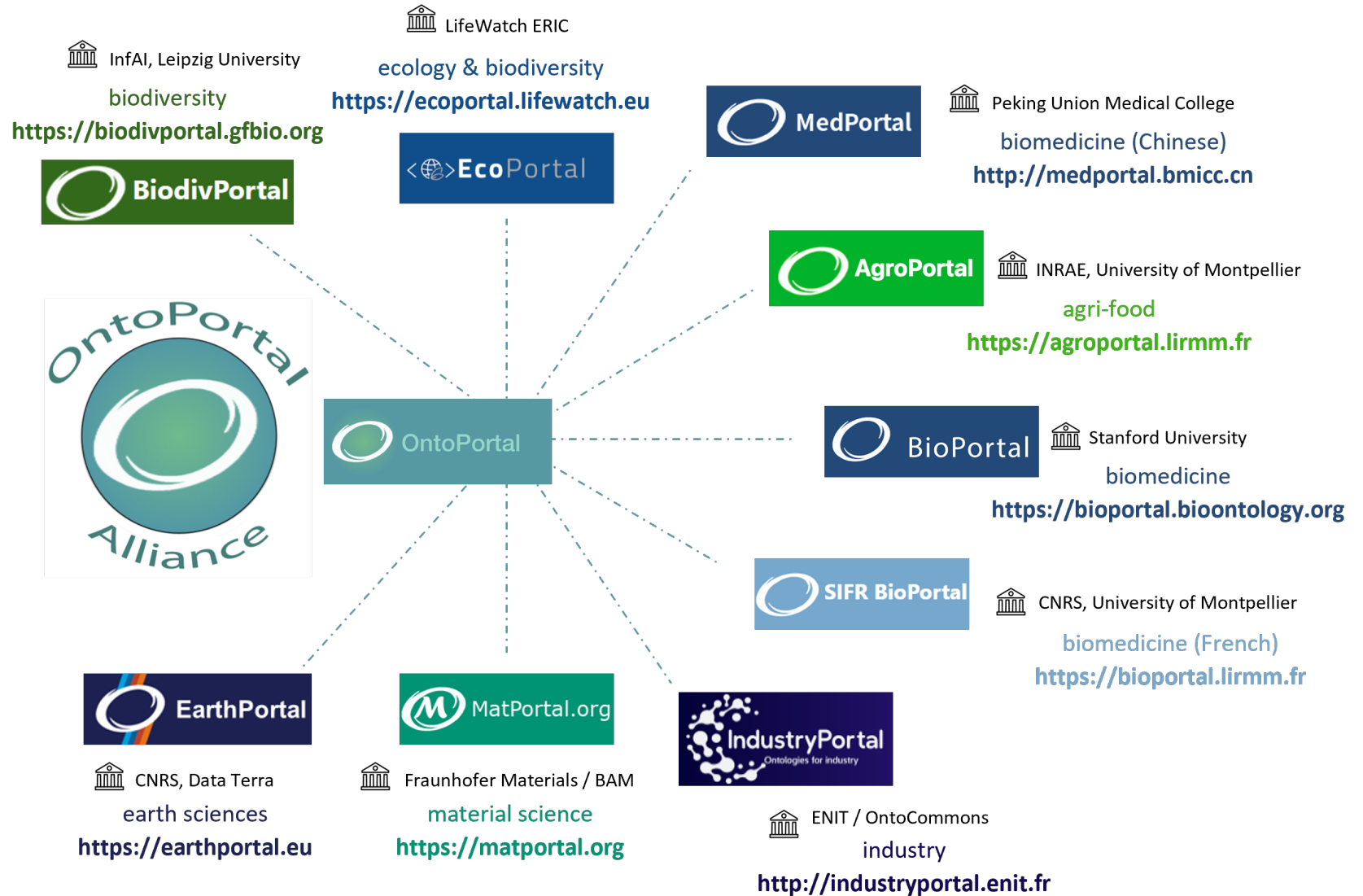
Dataverse Installations Around the World

Open map in new window

105 Installations



Name	Country	Year	Description
[Your installation! 🇧🇪 🇩🇪 🇫🇷]	[Country]	[Launched]	[A description of your installation of Dataverse.] To add your installation to the map, please create an issue .
Abacus	Canada	2014	Open for researchers associated with British Columbia universities to deposit data.
ACSS Dataverse	Lebanon	2019	The ACSS Dataverse is an initiative of the Arab Council for the Social Sciences, in collaboration with the Odum Institute for Research in Social Science at the University of North Carolina at Chapel Hill.



Federated and open multi-disciplinary environment

The European Open Science Cloud

2015: the European Commission proposed creating a European Open Science Cloud (EOSC)

The aim was to **federate existing research data infrastructures** in Europe and realise a web of FAIR data and related services for science, making research data interoperable and machine actionable following the FAIR guiding principles.

In the initial phase of development until 2020, the Commission invested around **€320 million to start prototyping the EOSC**.

Phase: initial phase of implementation (2018-2020)
phase of implementation (2021-2030)

Budget: 1 billion EUR (2023-2030)

EOSC tripartite governance:
European Commission
EOSC Steering Board
EOSC Association

	EOSC-CORE szolgáltatások	ARP-CORE szolgáltatások	ARP 2. fázis (terv)
jelenlegi szolgáltatási állapot	november 1-től béta üzemelés	október 13-tól béta üzemelés	
adattárolás	-	Dataverse (1,3 petabájt)	?
PID kezelés	PID Graph PID Meta Resolver (PIDMR)	- (az ARP tudásgráfban implicit PID kezelés) - (DOI, ARK, handle kezelés beépítve, default resolverek használata)	-
szolgáltatáskatalógus	Service Catalogue	ARP tudásgráfban manuálisan kezelve	Adatszolgáltatások kereshető listája az ARP portálon tervezett (áttekintés)
metaadatséma regiszter	Metadata Schema and Crosswalk Registry (MSCR)	CEDAR metaadatséma regiszter -	crosswalk regiszter funkció kifejlesztése tervezett

	EOSC-CORE szolgáltatások	ARP-CORE szolgáltatások	ARP 2. fázis (terv)
autentikáció	community alapú föderált (interoperability framework, rules of participation, support services)	EDUID	?
metaadat element kezelés (szemantika)	EOSC Data Type Registry (DTR)	OntoPortal	
föderáció elvárások	Compliance Assessment Toolkit (CAT)	ARP tudásgráfban manuálisan kezelve (FAIR, PID, GDPR, felhasználói licenszek)	szoftver céleszközök tervezettek ezekre a célokra
kutatási projekt azonosítás	Research Activity Identifier Service (RAiD)	-	-

- Az ARP nem egy repozitórium hanem **infrastruktúra**
 - együttműködő szolgáltatások összehangolt rendszere
 - jövőálló, a megterülés hosszabb távon értelmezett
 - az országban egyedülálló szolgáltatások jelennek meg (ontológia regiszter, metaadatséma regiszter, RO-Crate, központi tudásgráf,...)
 - szinkronban van az európai trendekkel és technológia színvonallal
- EOSC = “system of system”
 - meglévő és működő adat-infrastruktúrákat föderál,
 - egyedi szolgáltatások csak a szolgáltatáskatalógusban jelenhetnek meg

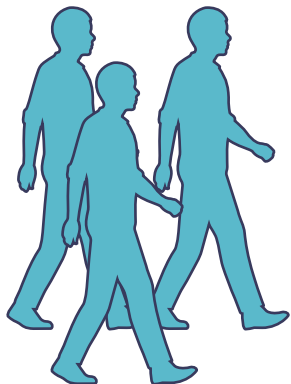
- Az ARP tölthetné be a magyar EOSC Csomópont (National Node) szerepet
 - komplex adatinfrastruktúra, alapvető föderációs szolgáltatásokkal
 - Magyarországon jelenleg az egyetlen ilyen
 - “belépőjegy” az EOSC föderációba
- Az ARP egy élő és folyamatosan fejlesztendő informatikai rendszer szoftverfunkciók és adattartalom szempontokból
 - az ARP hiányzó szolgáltatásai miatt (ARP projekt 2. fázis)
 - az EOSC informatikai (föderációs), rendszerszervezési és szervezeti elvárásai miatt

projekt

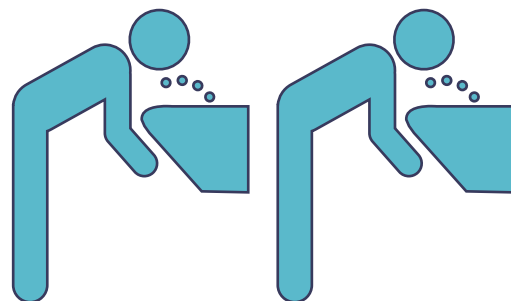


önálló memóriaintézmény

- Működtetés szervezeti formája?
 - a “bizalom” mint archívum termék
 - EOSC föderációs elvárás: az infrastruktúra legyen független jogi személyiség
 - Kovács László: “ARP Adatrepozitórium, mint független ELKH memóriaintézmény” (konceptió vázlat) 2023. augusztus 25.
 - projekt csak mint átmeneti megoldás



ARP 2. fázis



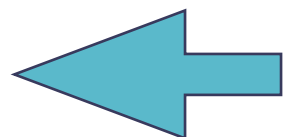
ARP



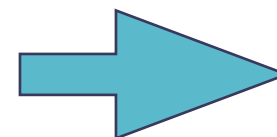
EOSC

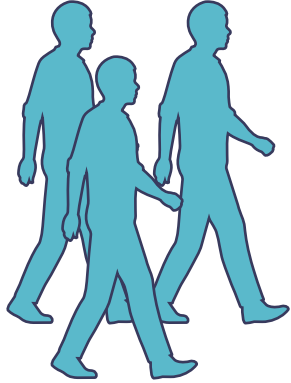
kutatósi munkafolyamat + ARP integrálás
automatikus metaadat keletkeztetés

ARP EOSC föderálás

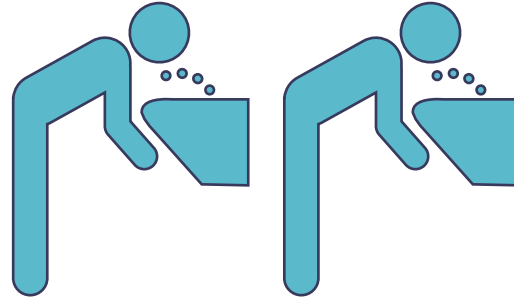


?





ARP 2. fázis

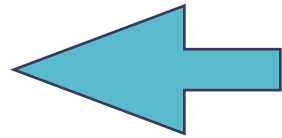


ARP

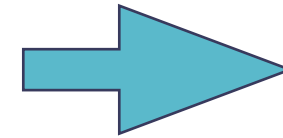


EOSC

HUN-REN



mindkettő irány



kormányzat

ARP kutatási adatrepozitórium platform

Dr. Kovács László

SZTAKI DSD Elosztott Rendszerek Osztály

laszlo.kovacs@sztaki.hun-ren.hu