ARP kutatási adatrepozitórium platform

Kovács László
SZTAKI DSD Elosztott Rendszerek Osztály
laszlo.kovacs@sztaki.hun-ren.hu

ARP adatrepozitórium

SZTAKI Elosztott Rendszerek Osztály

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

Az ARP projekt előzménye

- 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 elvárás-rendszer)
  - MI (mesterséges intelligencia) viharos megjelenése
  - “az adat mint érték, adat mint erőforrás”

Értékelési módszertan?

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

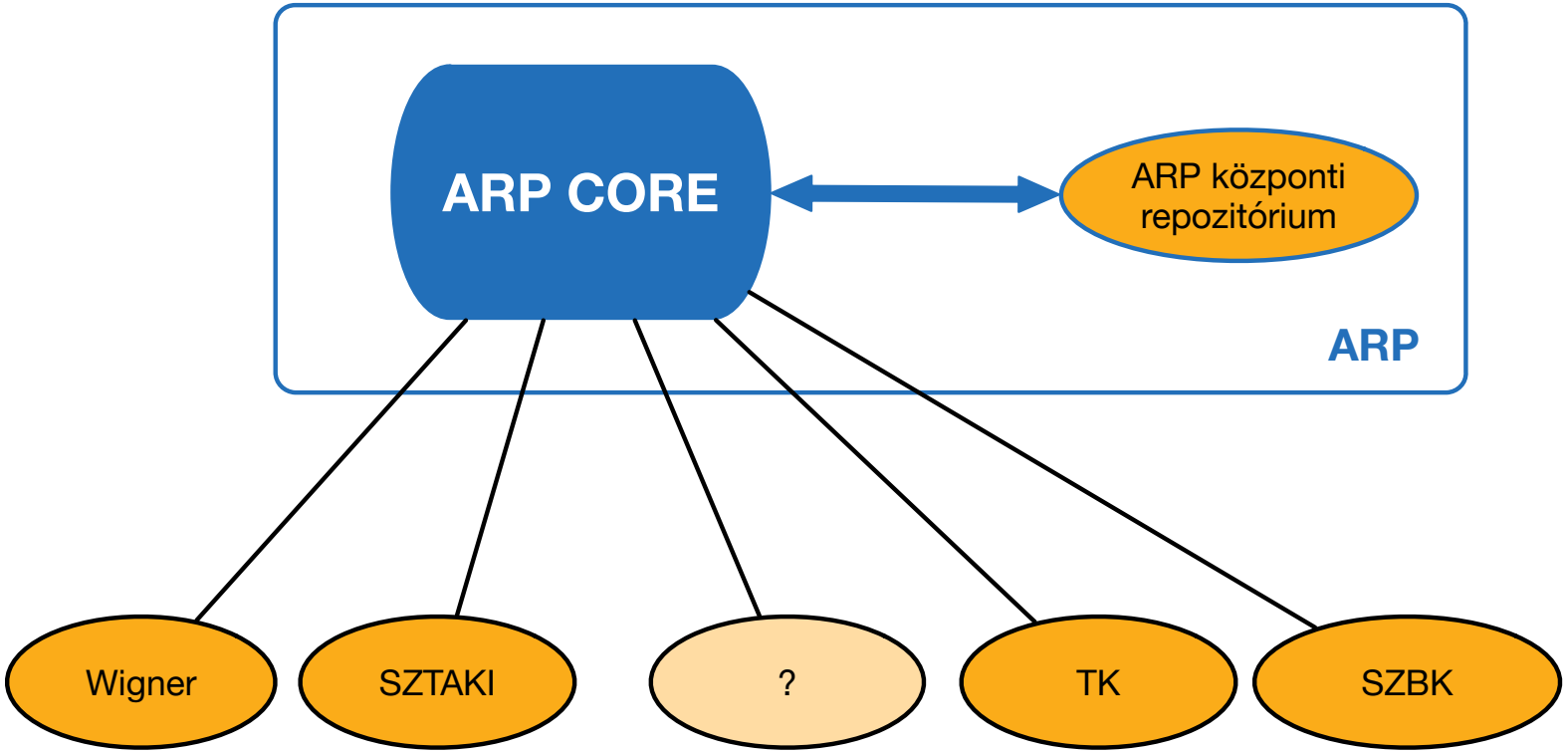
(Kutatók, intézeti vezetők kvázi “vakon repülnek”.)

Az ARP projekt célja

- **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

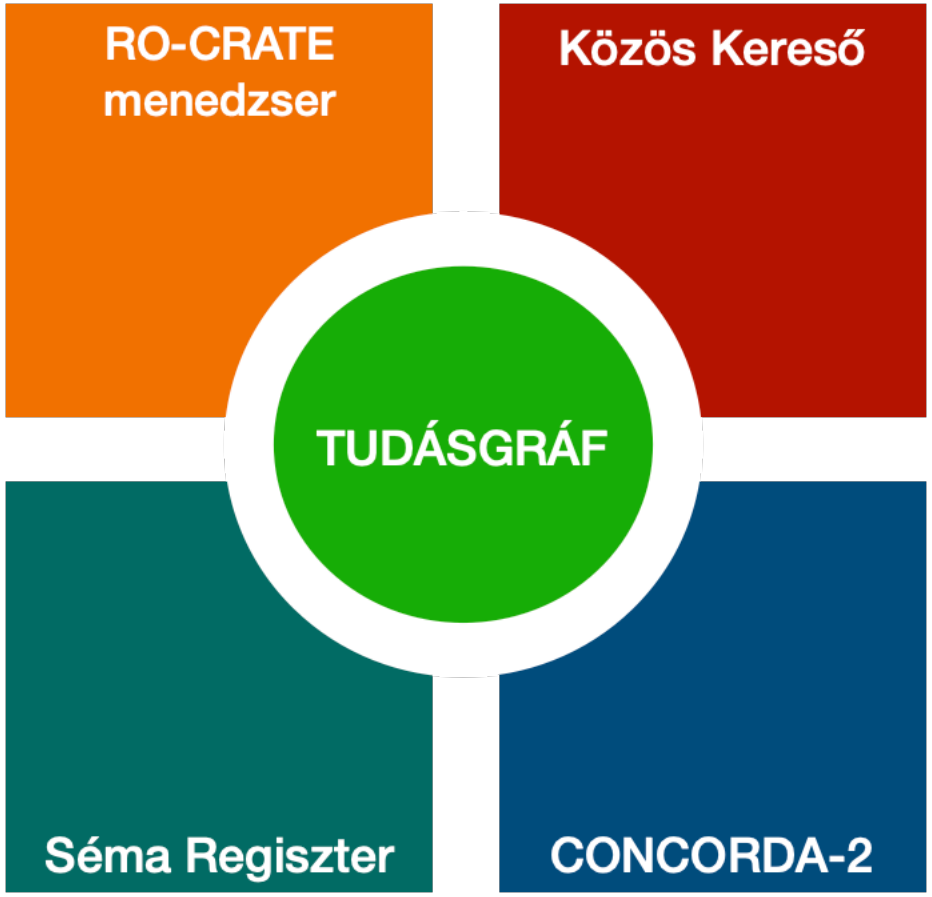
ARP föderáció

Föderált HUN-REN kutatási adatrepozitórium hálózat



Nagy intézetek vagy ágazatok saját kutatási adatrepozitóriumokkal

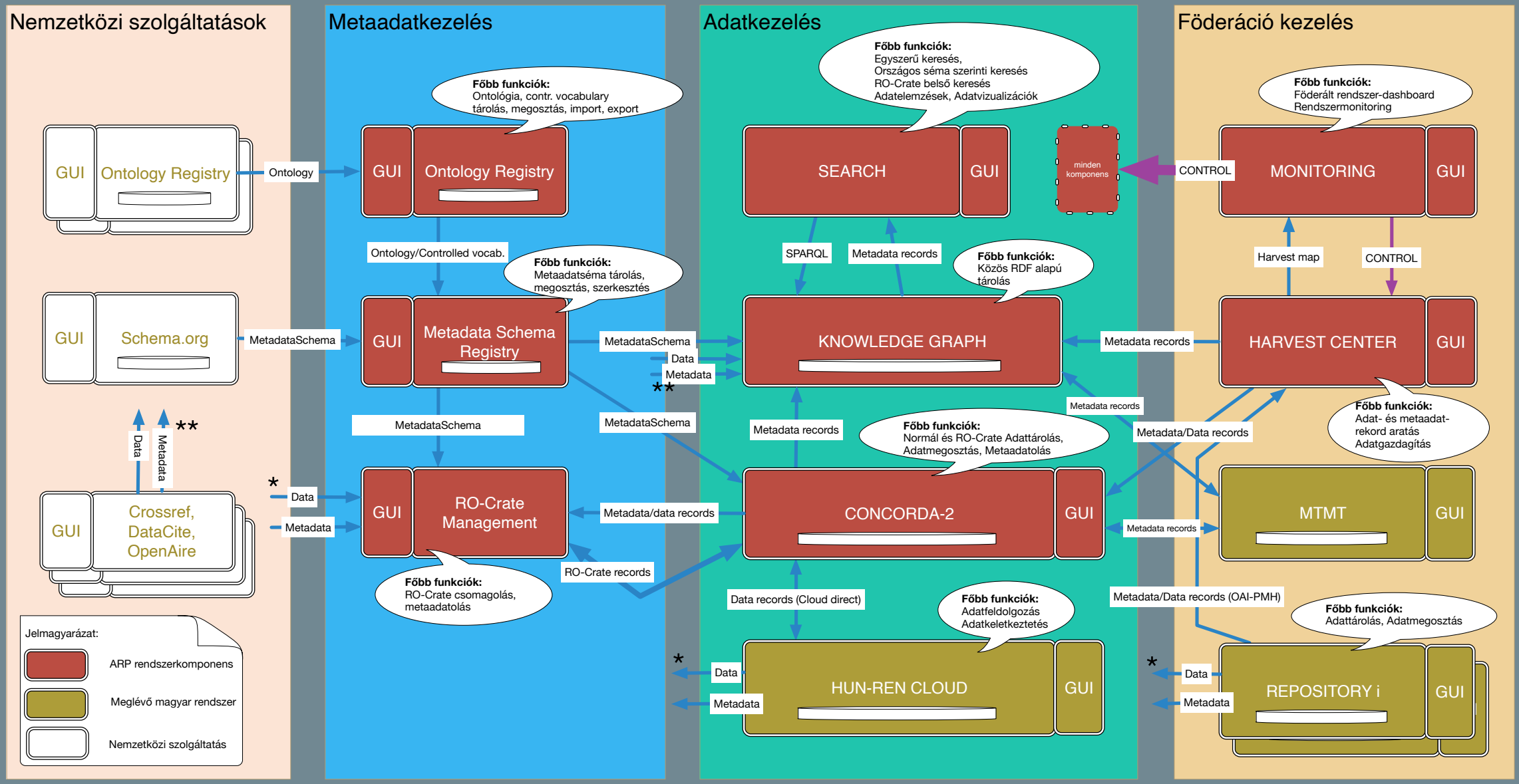
ARP CORE



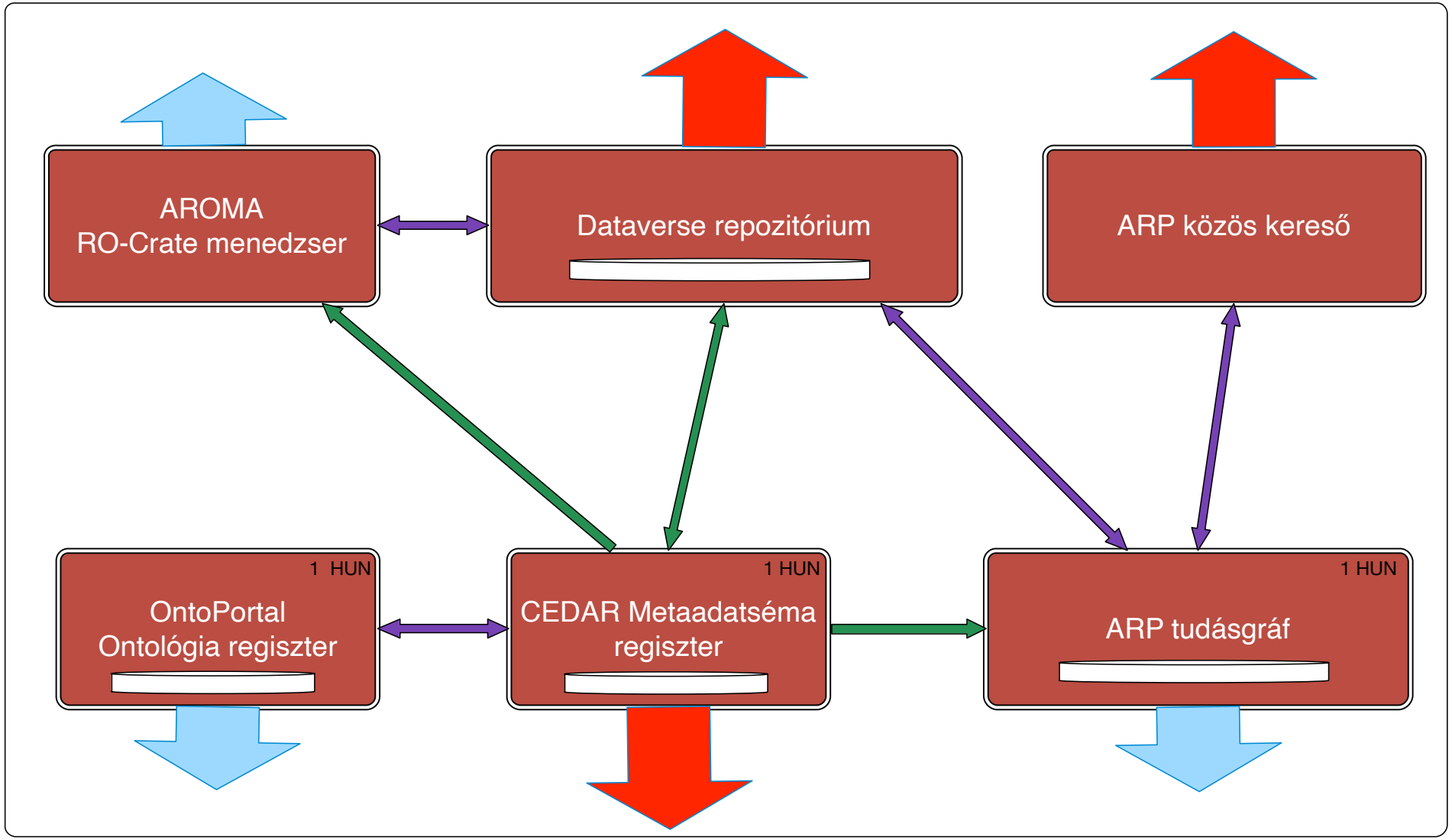
ARP CORE

- 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)

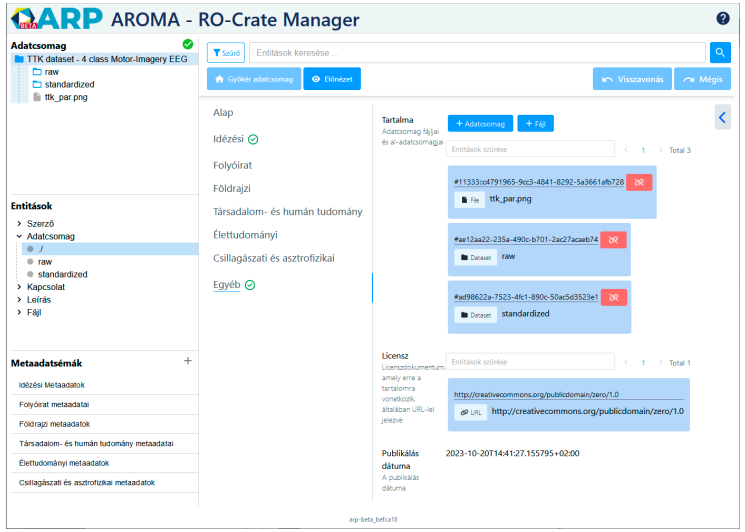
ARP PLATFORM



ARP CORE



ARP CORE



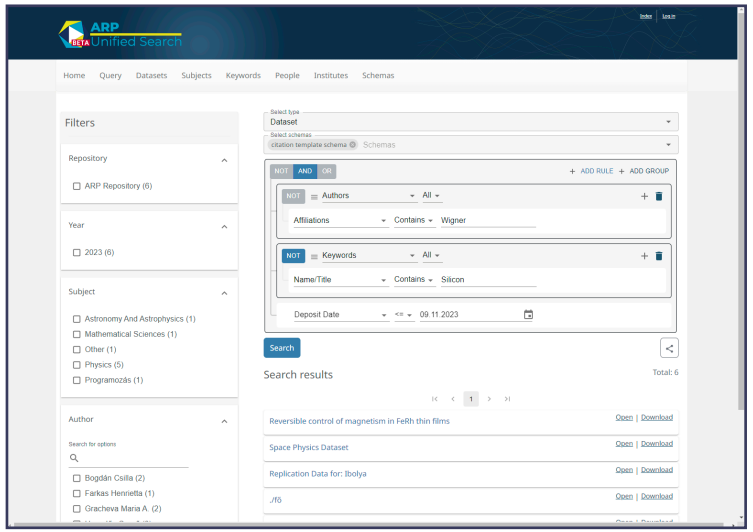
RO-CRATE menedzser

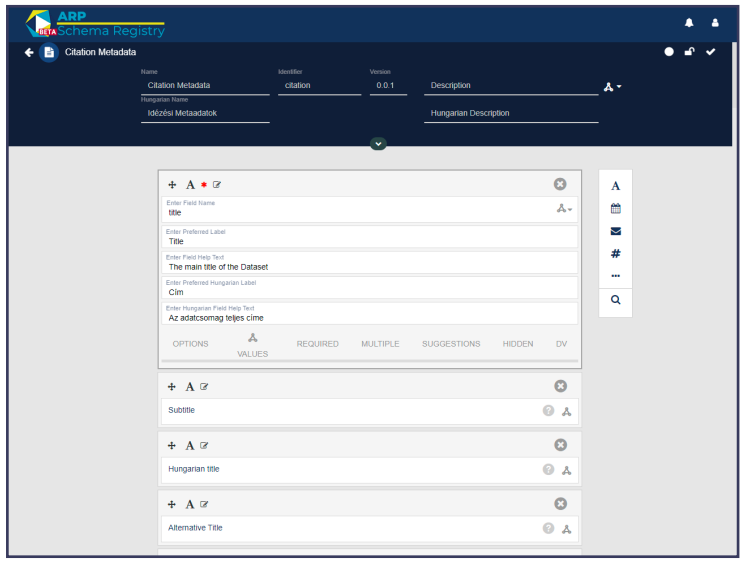
Közös Kereső

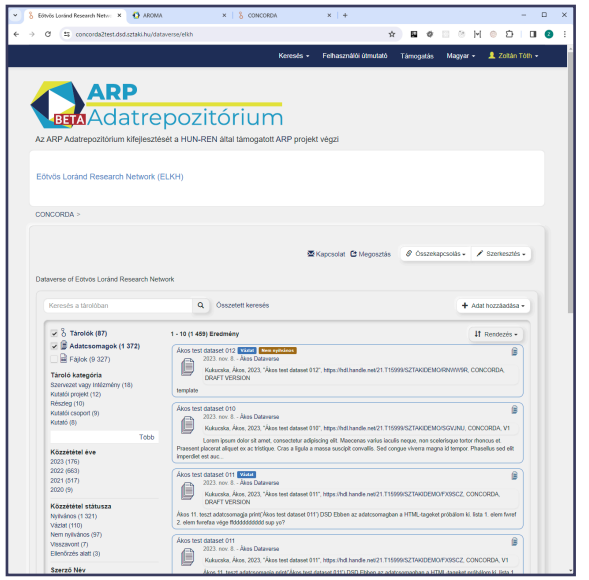
TUDÁSGRÁF

Séma Regiszter

CONCORDA-2

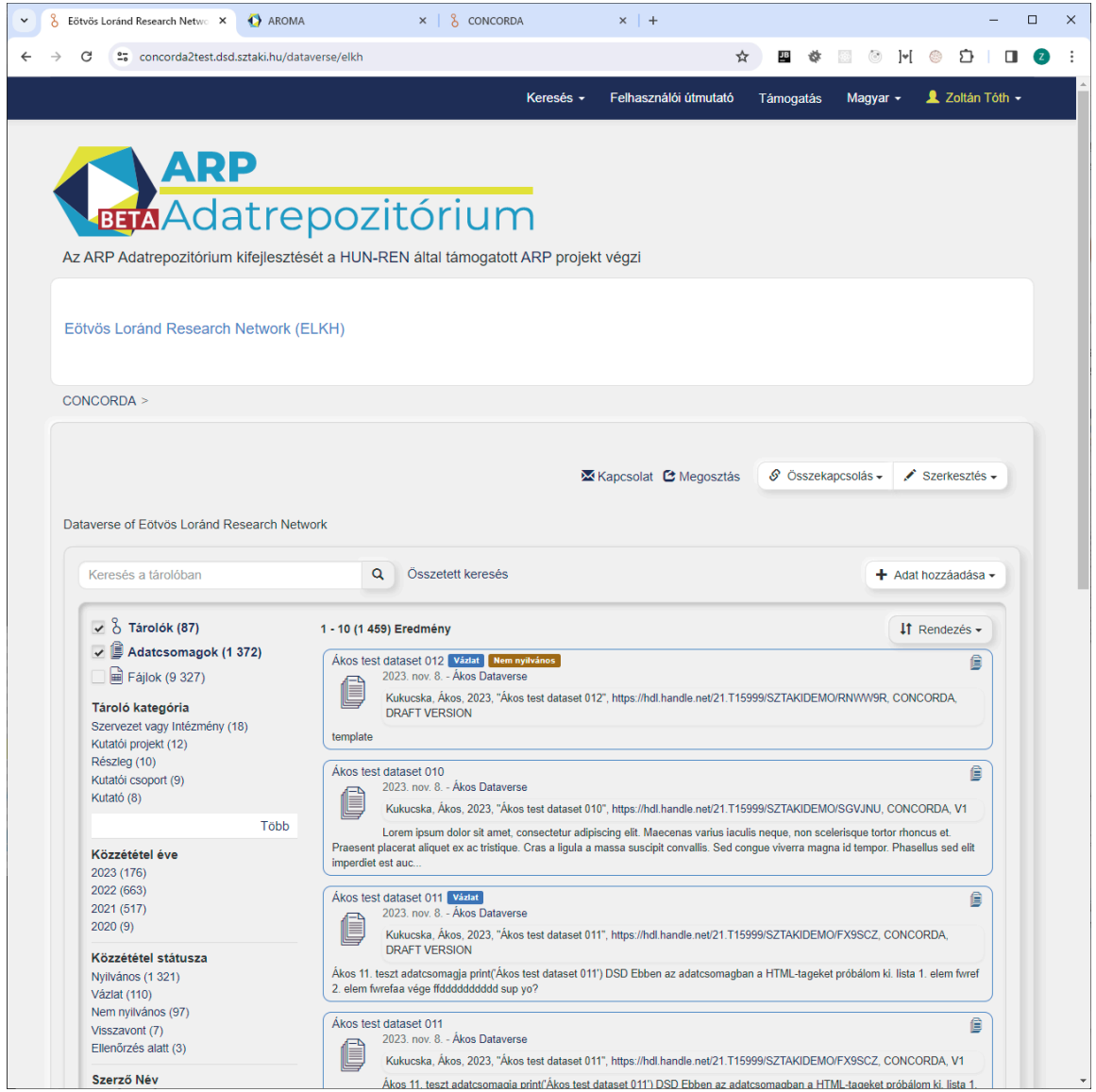






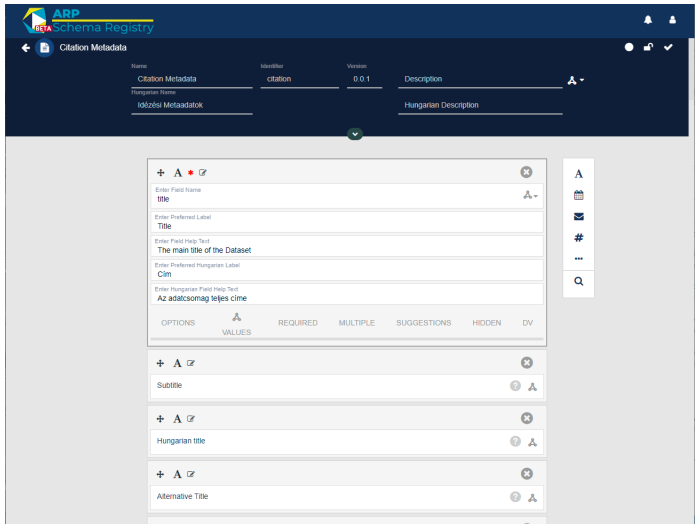
ARP Adatrepozitórium

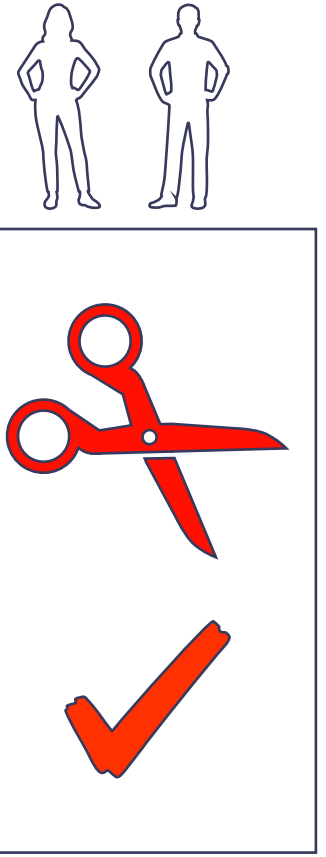
- Dataverse
  - adattárolás, adatmegosztás
- újratervezett metaadatséma kezelés
- ARP Cedar metaadatséma regiszter integrálás
- ARP AROMA RO-Crate menedzser integrálás



ARP Metaadatséma regiszter

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



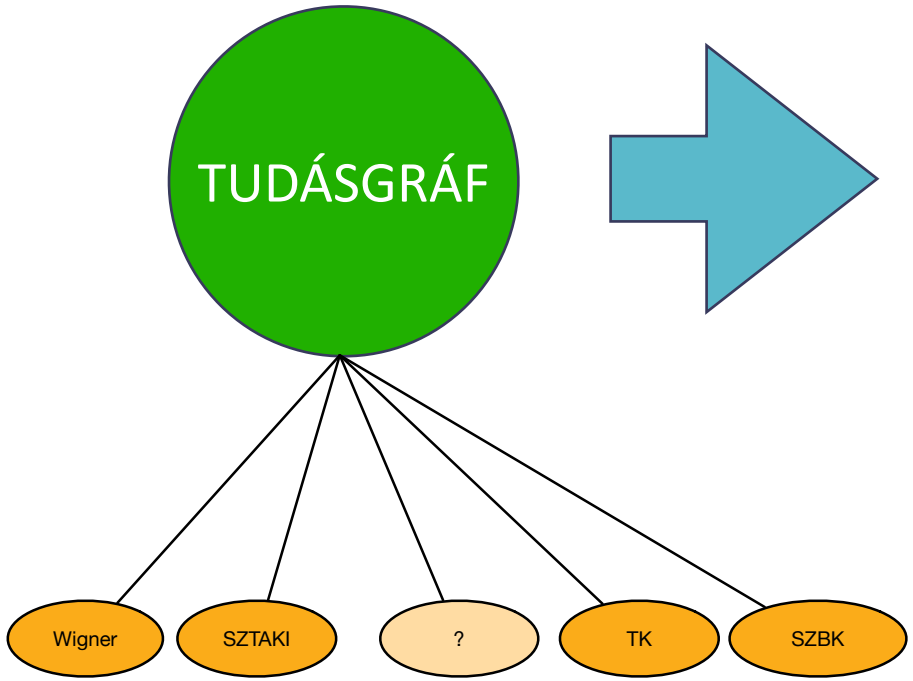


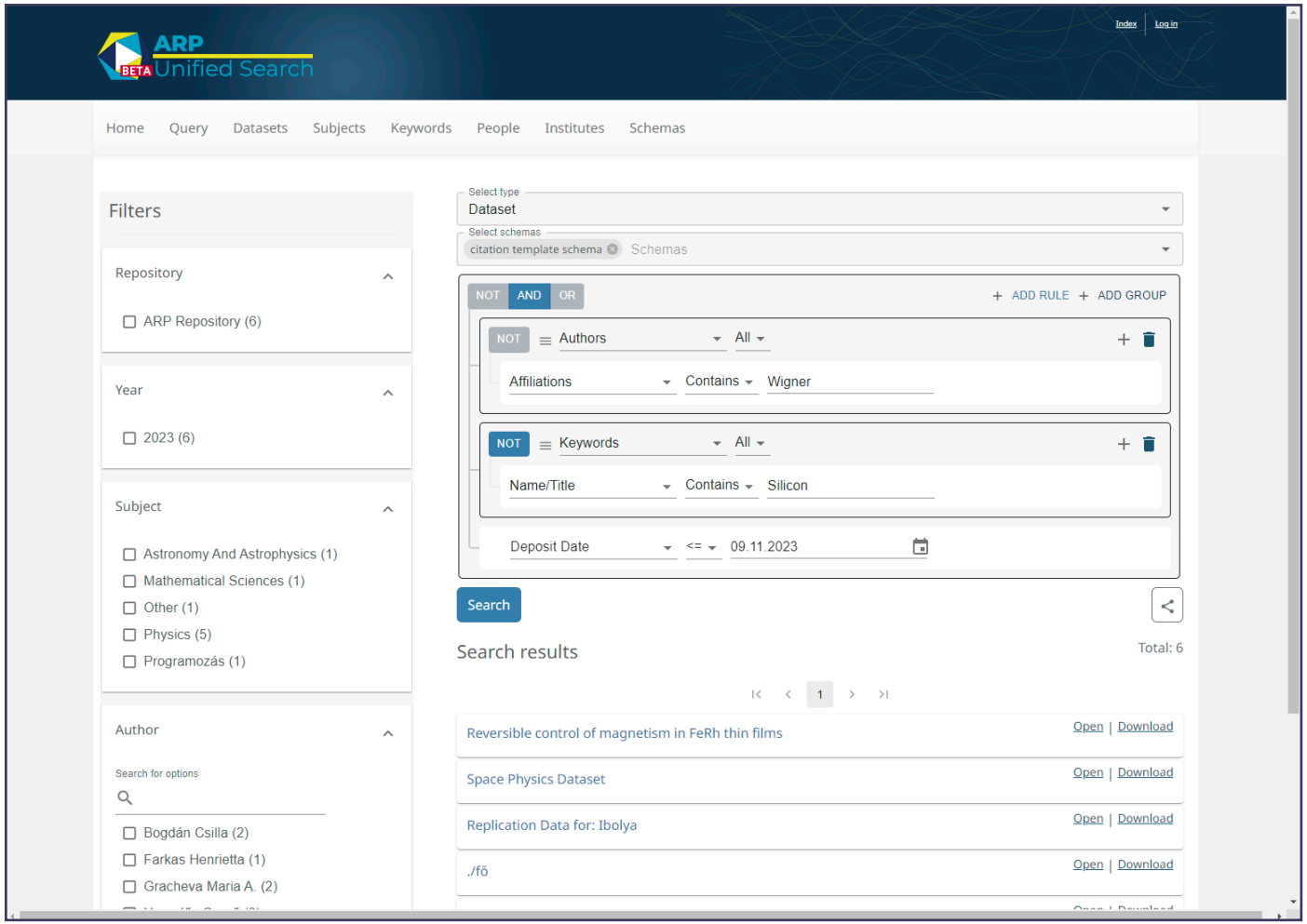
ARP Cedar

ARP CORE

ARP Közös kereső

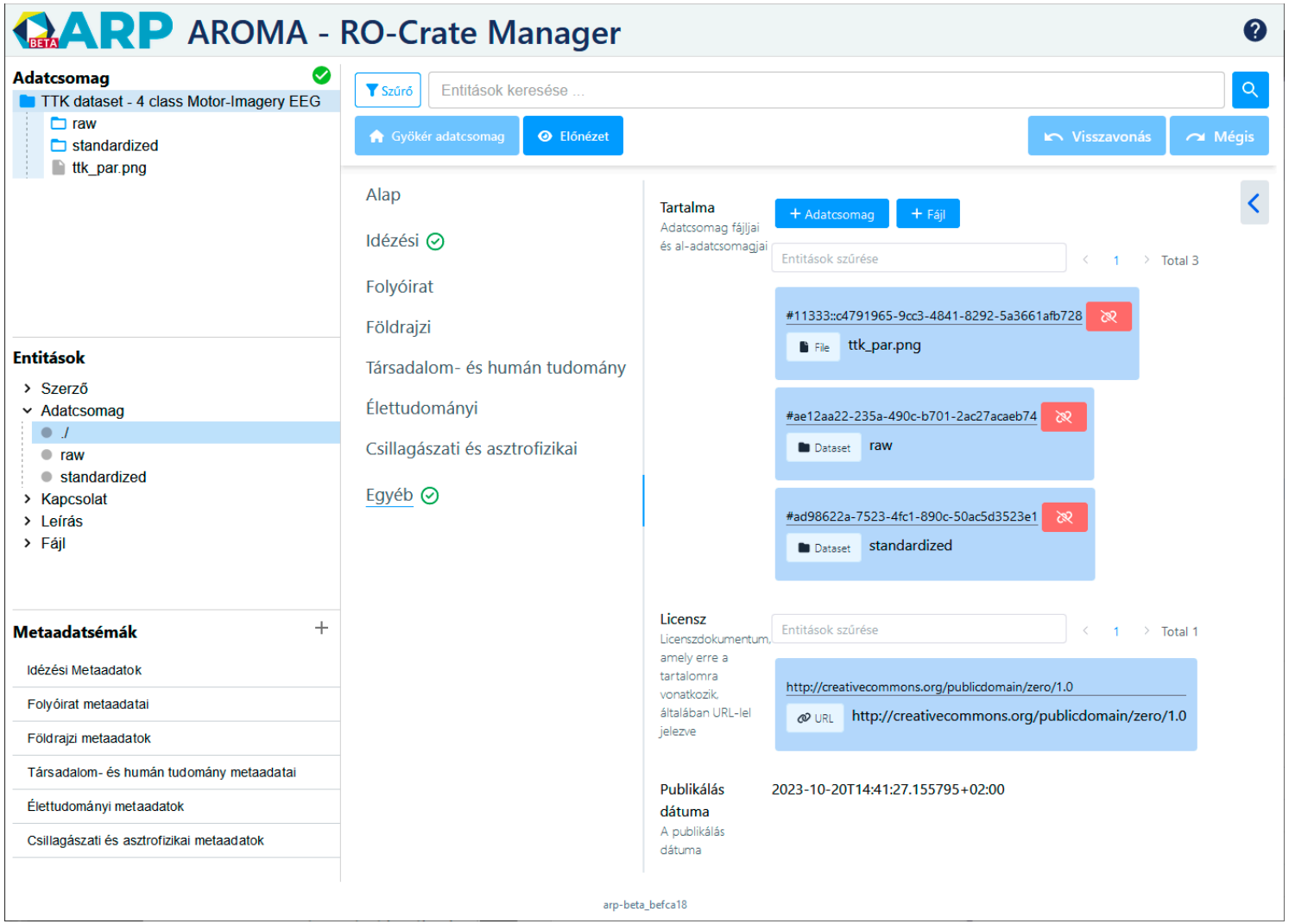
Föderált hálózatban keres
Fájl szinten keres





ARP AROMA - RO-Crate Menedzser

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

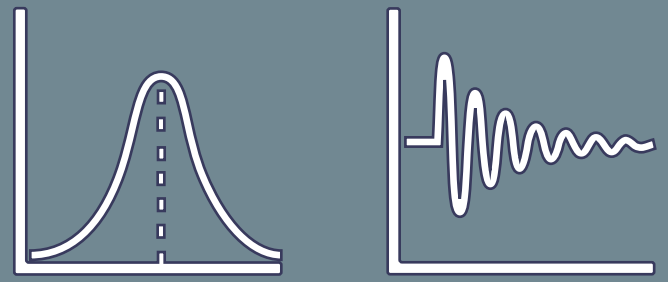


Kutatási Objektum (Research Object)

- 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

Research Object Entitásai

- Adatok
  - kiinduláskori kutatási adatok
  - kutatás során keletkezett új adatok

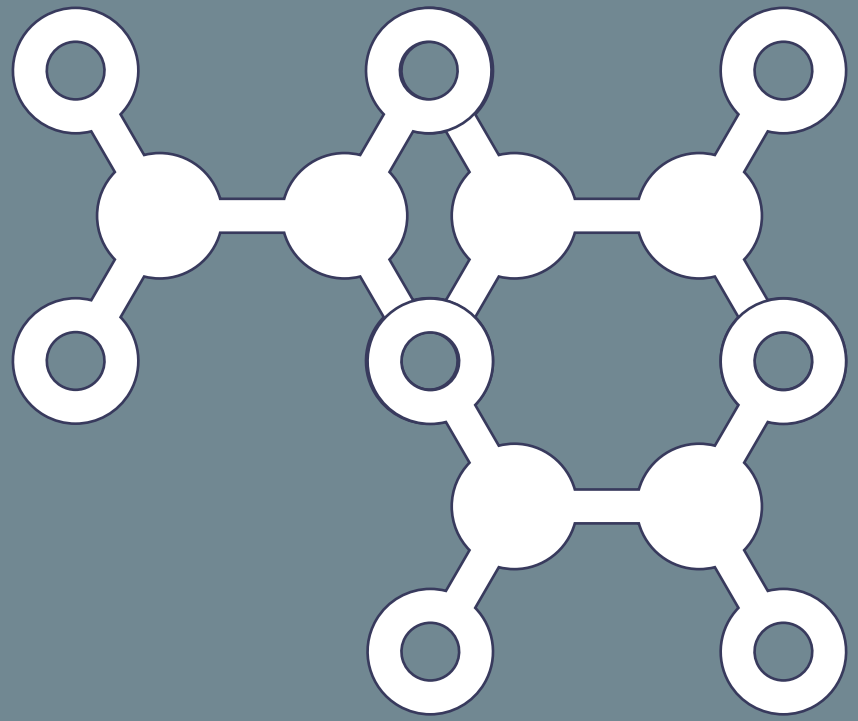


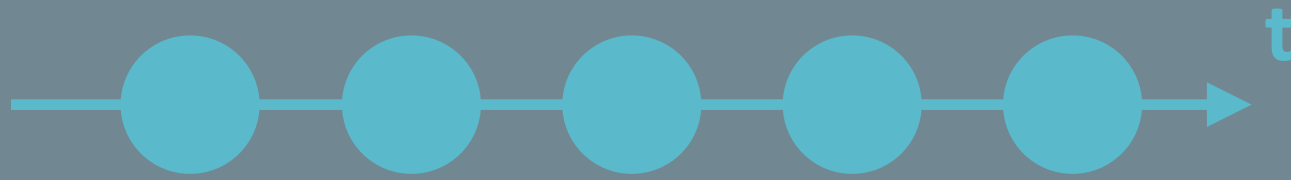
- Kutatási módszer és módszertan
  - fizikai/digitális anyagok/környezetek és tevékenységek/műveletek leírása
  - adatfeldolgozás informatikai környezete
  - adatkeletkeztetés és adatfeldolgozás módszerei

Fizikai tér + Digitális tér

Research Object Entitásai 2

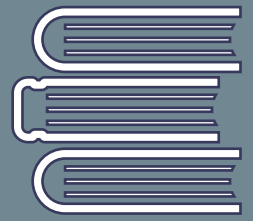
- Tudományos Workflow
  - a fizikai/adatfeldolgozási tevékenység téridő dinamikájának leírása
- Történetiség (Provenance)
  - történetiséget jellemző információk
  - a kutatási infrastruktúra használata
  - a használt műszerek, szoftverprogramok, infrastrukturális szolgáltatások beállítási paramétereinek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

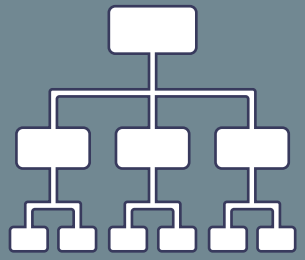




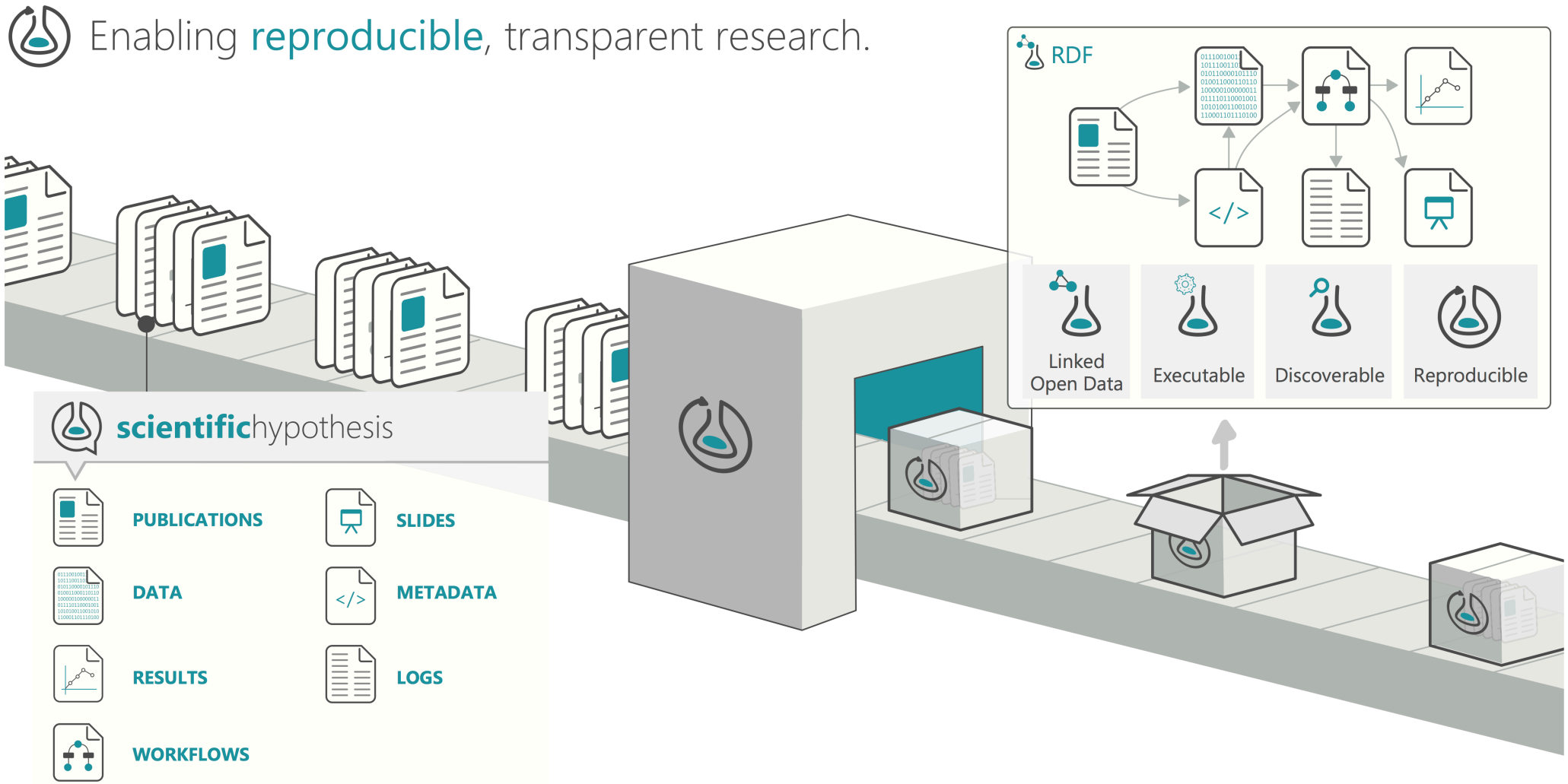
Research Object Entitásai 3

- 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



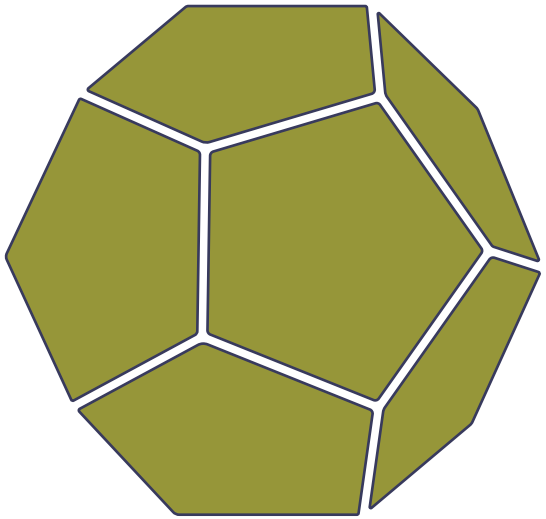


Research Object



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

RO-Crate használata



RO-Crate

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

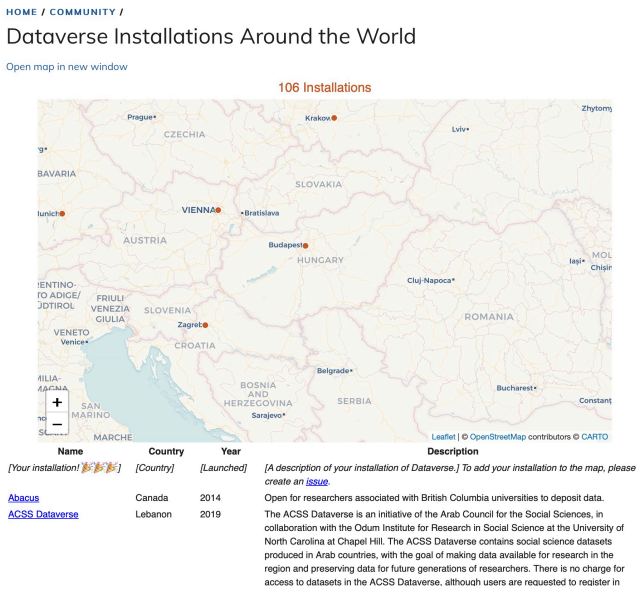
	dokumentum (publikáció)	kutatási adatcsomag
keresés	metaadat (pl. DC) alapján	specializált adatséma alapú metaadat alapján
megjelenítés	PDF (generikus) megjelenítő	csak szoftverrel (benne implicit értelmezés)
értelmezés	olvasás (emberi értelem)	csak szoftverrel (+ MI)

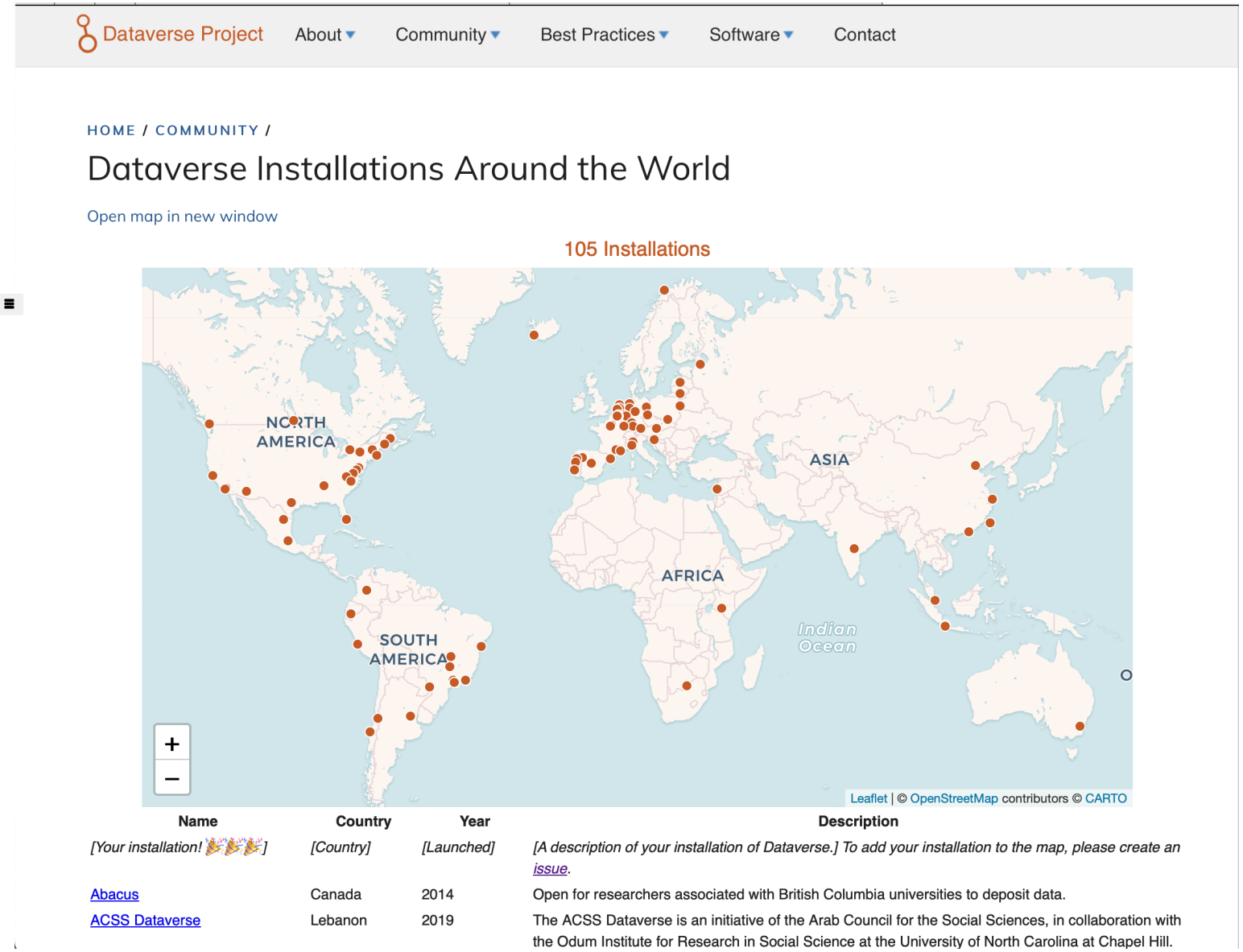
ARP nemzetközi beágyazódása

- 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

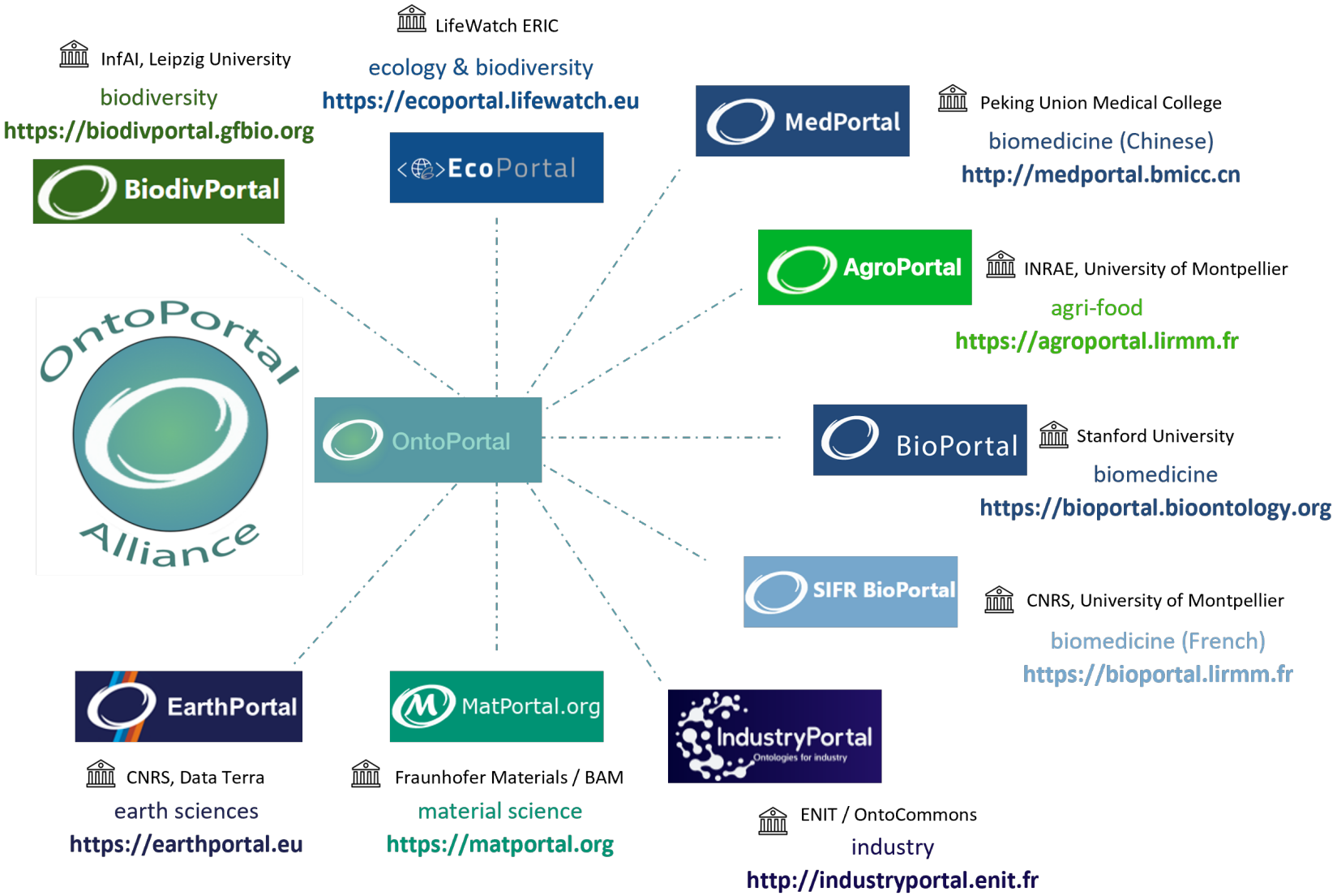
Dataverse community

ARP Dataverse (106.)





OntoPortal Alliance



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 - ARP CORE szolgáltatások összehasonlítása

	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 - ARP CORE szolgáltatások összehasonlítása 2

	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)	-	-

Következtetések

- Az ARP nem egy repozitórium hanem **infrastruktúra**
  - együttműködő szolgáltatások összehangolt rendszere
  - jövőálló, a megtérülés 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

Következtetések 2

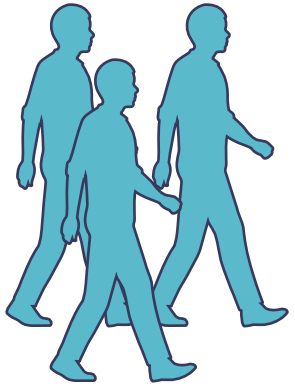
- Az ARP tölthetné be a magyar EOSC Csomópont (National Node) szerepet
  - komplex adatinfrastruktúra, alapvető föderációs szolgáltatásokkal
  - Magyarországon jelenleg az egyetlen ilyen
  - “belépőjegy” az EOSC föderációba
- Az ARP egy élő és folyamatosan fejlesztendő informatikai rendszer szoftverfunkciók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

ARP 2024-ben és azon túl

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” (koncepció vázlat) 2023. augusztus 25.
  - projekt csak mint átmeneti megoldás

ARP 2. fázis? és/vagy EOSC integráció?



ARP 2. fázis

kutatási munkafolyamat + ARP integrálás
automatikus metaadat keletkeztetés

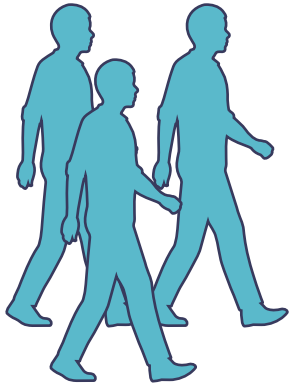
ARP

EOSC

ARP EOSC föderálás

?

ARP 2. fázis és EOSC integráció



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

ARP adatrepozitórium