

Az ARP platform adathozzáférési és - megosztási modelljei



Pallinger Péter, SZTAKI DSD

pallinger.peter@sztaki.hu

2023. 04. 13.

ELKH | Eötvös Loránd
Research Network

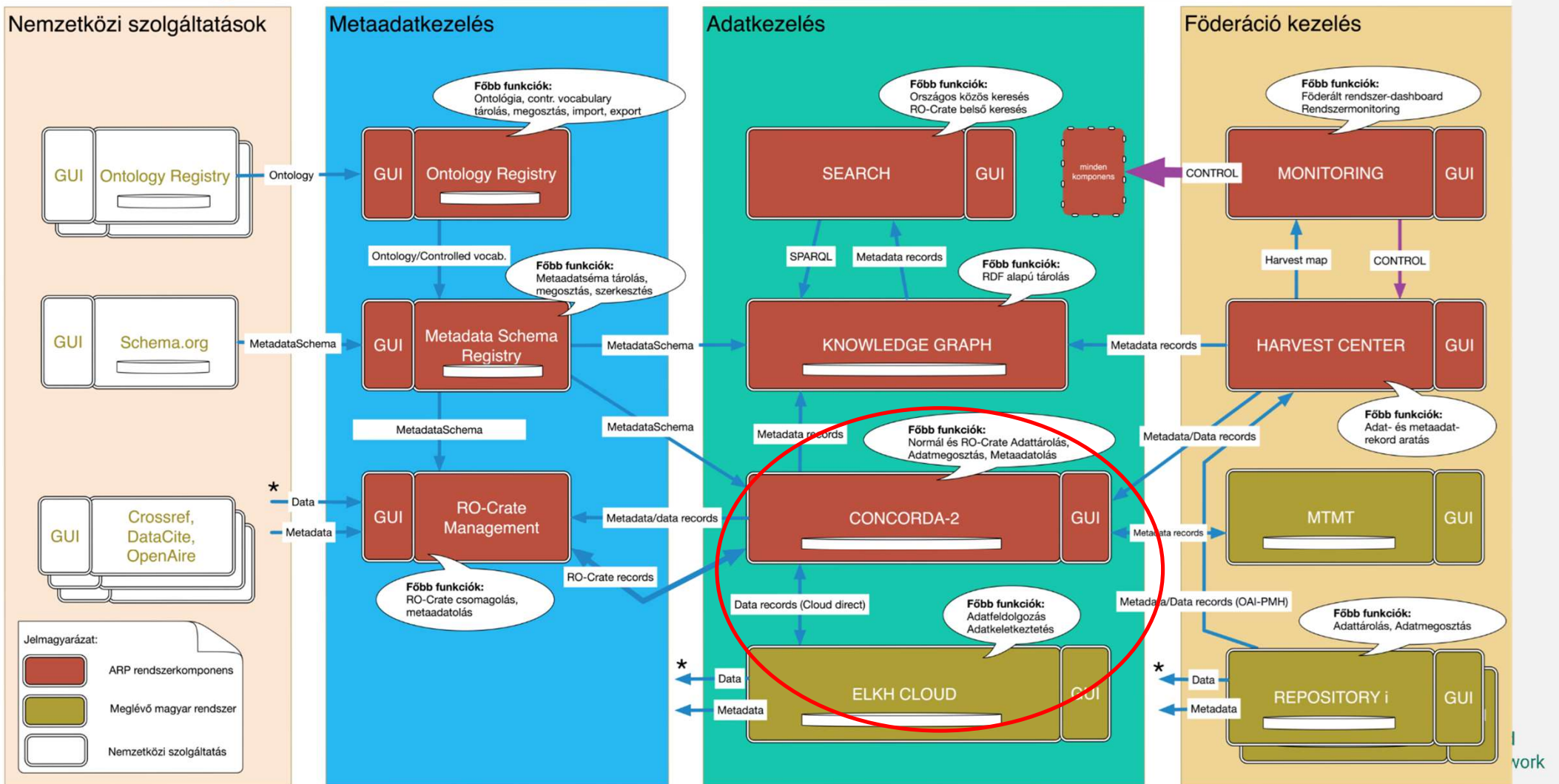
ELKH-ARP projekt

- Cél
 - A teljes ELKH kutatóhálózat számára folyamatos és hosszú távú kutatási repozitóriumi infrastruktúra-szolgáltatás megvalósítása
- Résztvevők
 - Számítástechnikai és Automatizálási Kutatóintézet (SZTAKI)
 - Társadalomtudományi Kutatóközpont (TK)
 - Wigner Fizikai Kutatóközpont (Wigner FK)
- Időtartam
 - 2021.11.01. - 2023.12.31.

A CONCORDA adatrepozitórium története

- 2020-ban jött létre
 - SZTAKI és WIGNER intézetek összefogásában
 - Régi MTA/ELKH felhőben
 - mind a SZTAKI, mind a WIGNER telephelyen
 - Valmint a DSD felhőjében
 - Harvard dataverse 4.20 alapokon
 - condorda.sztaki.hu, concorda.hu címeken
- 2021 végén az ARP projekt karolta fel
 - Adattárolás fő elemévé választottuk
 - science-data.hu címre költözött
 - Háttértár kiegészítését fogjuk elvégezni
 - Dataverse verziófrissítések, jelenleg 5.10 az éles
 - Integráció az ARP által fejlesztett szolgáltatásokkal

ARP logikai architektúra

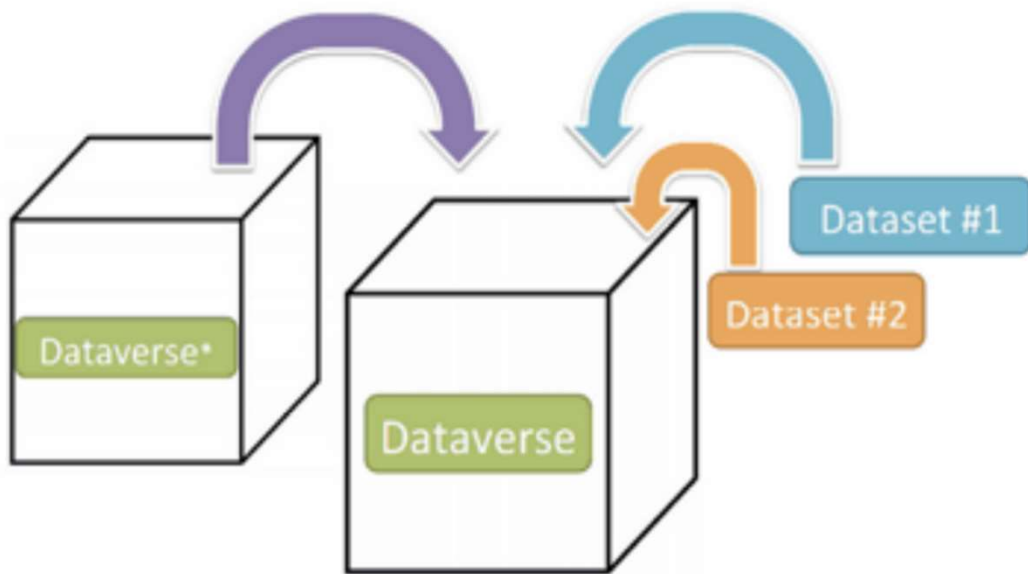


Concorda - Dataverse

- A Dataverse szoftver:
 - Harvard által fejlesztett adatrepozitórium kifejezetten kutatási adatok megosztására
- Adatmodell
 - Tárolók (Dataverse)
 - Tárolókban további tárolók vagy adatcsomagok lehetnek
 - Elsősorban hozzáférés-szabályozás szempontjából van jelentőségük
 - Adatcsomagok (Dataset)
 - Részletes és gazdag metaadatok az adatcsomagokban
 - Adatcsomagok csak fájlokat tartalmazhatnak

A Dataverse adatmodellje

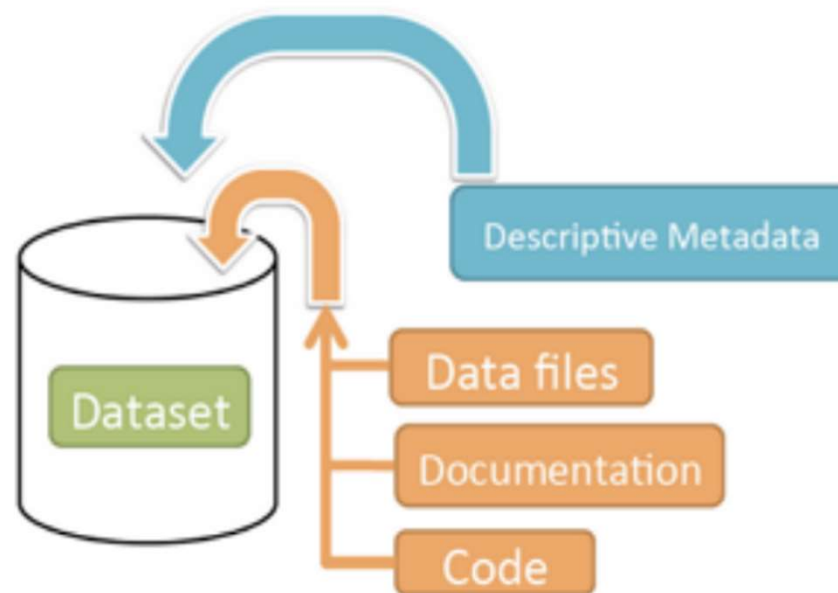
Schematic Diagram of a **Dataverse** in Dataverse 4.0



Container for your **Datasets** and/or **Dataverses***

* Dataverses can now contain other Dataverses (this replaces Collections & Subnetworks)

Schematic Diagram of a **Dataset** in Dataverse 4.0



Container for your data, documentation, and code.

Fájlok és adatcsomagok feltöltése helyi tárolókból

- Mit értünk helyi tároló alatt
 - Intézményekben (kutatóintézet, egyetem) elsődleges adattárolásra használt szerver
 - Elsősorban Synology NAS-t célozzuk meg
 - Ilyet vettünk, több kutatóintézetben ilyet használnak
 - De bármilyen linux alapú szerveren működnek a lenti módszerek
- Megoldások helyi tárolókból feltöltésre
 - Manuális feltöltés weben
 - Közvetlen API elérés: curl, java/python/ruby/R/stb. könyvtárak
 - Dataverse Mass Uploader
 - DVUploader
 - RDM Integration Dashboard
- Helyi tároló helyett vagy mellett
 - Köztes tároló az ELKH felhőben

Manuális feltöltés webes felületen

- Egyszerű
- Szinkron
 - A feltöltő asztali számítógépe / laptopja bekapcsolva kell maradjon
 - 1 TB feltöltése minimum 3 óra
 - de inkább 30 óra
- Megszakadt feltöltés esetén manuálisan kell ellenőrizni, hogy mi lett már feltöltve
- Ha az adat helyi tárolóban van tárolva, akkor további hibalehetőségek lépnek fel

8

2023.
04. 13.



CONCORDA
Concentrated Cooperation on Research Data

A CONCORDA új verziójának kifejlesztését az ELKH támogatta ARP projekt végzi

CONCORDA > Eötvös Loránd Research Network (ELKH) > Wigner Research Centre for Physics > Finite-size scaling of the photon-blockade breakdown dissipative quantum phase transition >

Files

All file types are supported for upload and download in their original format. If you are uploading Excel, CSV, TSV, RData, Stata, or SPSS files, see the guides for tabular support and limitations.

Upload with HTTP via your browser ^

Select files or drag and drop into the upload widget. Maximum of 1,000 files per upload. File upload limit is 100.0 GB per file.

+ Select Files to Add

Drag and drop files here.

Done



SZTAKI

ELKH

Eötvös Loránd
Research Network

Dataverse Mass Uploader

- Python script
 - Létező adatcsomagokba tölt fel fájlokat
 - pyDataverse könyvtárat használja, de lehet 2 GiB fölötti fájlokat is feltölteni, ha van curl telepítve
 - `python3 uploadFilesToExistingDataset.py -u https://science-data.hu -d HANDLE_OR_DOI -k $APIKEY FILE1 FILE2`
- Szkriptelhető, csomagolható
 - Pl. több adatcsomagba feltöltéshez
- Elérhető
 - <https://github.com/dsd-sztaki-hu/DataverseMassUploader>
- Leendő fejlesztések
 - Fájlok felülírása
 - Adatcsomag- és fájl-metaadat feltöltés
 - RO-Crate támogatás

DVUploader

- Java command line alkalmazás
 - Létező adatcsomagokba tölt fel fájlokat
 - java8+ kompatibilis
 - részletesen konfigurálható, rengeteg opcióval
 - rate limiting
 - konfigurálható várakozási idők
 - reguláris kifejezések
 - stb.
 - `java -jar DVUploader-*.jar -server=https://science-data.hu -did=<Dataset DOI or handle> -key=<API Key> <file or directory list>`
- Szkriptelhető, csomagolható
 - Pl. több adatcsomagba feltöltéshez
- Nincs metaadat-feltöltés
 - és nem is terveznek
- Elérhető
 - <https://github.com/GlobalDataverseCommunityConsortium/dataverse-uploader>

RDM Integration Dashboard

- Hasonló céllal készült mint a korábbi kettő
 - <https://github.com/libis/rdm-integration>
- Szép grafikus interfésszel
 - Elsősorban szinkronizálásra kitalálva
- Sok storage backend támogatott
 - Lokális
 - Github, Gitlab, IRODS
- Go nyelven
 - csak 64 bites bináris van
 - libc2.32-re fordítva
 - redis szerver is kell neki
- alfa verzió
- Nincs metaadat-feltöltés
 - és nem is terveznek

11

2023.
04. 13.

The screenshot displays the 'Demo Dataverse' interface. At the top, there's a navigation bar with 'Home' and 'Quit' links. Below it, a control bar includes a '<< Return' button, a 'Bulk select to:' dropdown, and buttons for 'Mirror', 'Copy', 'Clear selection', and 'Submit >>'. The main content area is titled 'Local filesystem' and shows a list of files. A dropdown menu is open over the 'Local filesystem' header, listing comparison status options: 'Comparison status', '(New files)', '(Changed files)', '(Unchanged files)', and '(Files only in RDR)', each with a checked checkbox. A 'Show all...' link is at the bottom of the menu. The file list includes: 'CODE_OF_CONDUCT.md', 'CONTRIBUTING.md', 'Dockerfile', 'LICENSE.md', 'README.md', 'Vagrantfile', 'checkstyle.xml', 'conf', 'docker-aio', and '0prep_deps.sh'. Each file row has a status icon (blue circle with 'i' or red circle with 'o'), a copy icon, and a corresponding file name. The 'README.md' row is highlighted in blue, 'Reports_PROD' in red, 'Vagrantfile' in white, 'bookmarks.html' in red, and 'checkstyle.xml' in green. The 'conf' and 'docker-aio' rows are collapsed with a downward arrow.

Köztes tároló az ELKH felhőben

- “Concorda Data Staging (SSH, btrfs)” image az ELKH felhőben (SZTAKI ág)
 - Helyi tároló helyett/mellett használható
 - Bárki indíthat ilyet, viszonylag nagy diszkkal (több száz TB)
 - Ide ideiglenesen gyűjthetőek adatok (6 hónap)
 - Futtatható a DataverseMassUploader vagy DVUploader
- Ki lehet próbálni!
 - <https://docs.google.com/document/d/1-zJdAHBUkq2Ob630ftIFRRoM6XNW-Tjz>
- Fejlesztési lehetőségek
 - Előre installált DataverseMassUploader és DVUploader, dependenciákkal

Adathozzáférés szabályozása a dataverse-ben

- Az adatcsomagok létrehozáskor DRAFT állapotba kerülnek
 - Az ilyen adatcsomagokat csak (befoglaló tárolóban) különleges hozzáféréssel rendelkezők láthatják
 - Egyedi hozzáférés adható ilyen adatcsomagokhoz a “Private URL” funkcióval (pl. cikk bírálók számára)
- Közzététel után a metaadatok hozzáférhetőek mindenki számára
 - Az adatok hozzáférhetősége akár fájlonként állítható
 - Korlátozott fájlok esetén beépített workflow használható hozzáférés engedélyezésére

Fájlok és adatcsomagok letöltése

- Webes felületen
 - Egyesével
 - ZIP formátumban
 - 100MB maximum (szerveroldalon konfigurálható, de pár GB fölött gond lehet vele)
 - metaadatokat nem tartalmaz
- RO-crate formátumban
 - vagy egy kiterjesztett ZIP vagy könyvtárstruktúra
 - könyvtárstruktúra letöltéséhez kliensoldali scriptelés kell
 - metaadatokat tartalmaz
 - fejlesztés alatt
- API-n keresztül
 - curl/wget
 - python/ruby/R/java könyvtárak
- DataverseMassUploader
 - `downloadDatasetFiles.py`



Köszönöm a figyelmet!

SZTAKI: www.sztaki.hu

DSD: dsd.sztaki.hu

Adatrepozitórium projekt: science-research-data.hu

CONCORDA: science-data.hu

ELKH CLOUD: science-cloud.hu