III.07 DOI: 10.31915/NWS.2023.9



# Az RO-Crate alapú kutatási objektum csomagolás keretrendszere az ELKH ARP platformban

# The framework of research object packaging based on RO-Crate in the ELKH ARP platform

Tóth Zoltán
*SZTAKI – DSD*
toth.zoltan@sztaki.hu

**Absztrakt**

A FAIR irányelveknek való megfelelés több kutatási területen megkerülhetetlen tényezővé kezd válni, és ezzel a magyar kutatók is egyre gyakrabban szembesülnek publikációs tevékenységük kapcsán. Ezek az irányelvek alapvetően azt célozzák meg, hogy a kutatásokat alátámasztó adatok megtalálhatóak és feldolgozhatóak lehessenek számítástechnikai eszközökkel, akár emberi beavatkozás nélkül is. Előremutató lépés a kutatási adatok hagyományos adatrepozitóriumban történő tárolása és metaadatolása, ez azonban nem feltétlenül elegendő az irányelvek követéséhez, ugyanis az ott elvárt metaadatok jellemzően a repozitóriumba feltöltött adatcsomag egységekre vonatkoznak, a finomabb, akár fájl-szintű értelmezésnek nincsen ezekben a rendszerekben szabványosított, bevett módja. A FAIR Digitális Objektumok megoldást jelenthetnek erre a problémára. Ennek egyik lehetséges megvalósítása az RO-Crate kutatási objektum csomagolás, melyet az ELKH ARP (ELKH Adat Repozitórium Platform) projekt bevezet az Adat Repozitórium Platformba. Ismertetésre kerül az RO-Crate formátum, valamint az, hogy ezzel a kutatók mi módon találkoznak majd a repozitóriumban végzett tevékenységeik során.
**Kulcsszavak:** FAIR irányelvek, adatrepozitórium, FAIR Digitális Objektum, RO-Crate, metaadat

**Abstract**

Compliance with FAIR principles is becoming an unavoidable factor in multiple research areas, that Hungarian researchers also face in their publishing activities. These principles primarily aim to ensure that the data supporting research can be located and processed using computational tools, even without human intervention. Storing and providing metadata for research data in a traditional data repository is a progressive step. However, this alone may not be sufficient to adhere to the guidelines, as the expected metadata in these systems typically pertain to the units of data packages uploaded to the repository, and there is no standardized, established way to interpret finer details, such as file-level information. FAIR Digital Objects can provide a solution to this problem. One possible implementation is the research object packaging based on RO-Crate, which the ELKH ARP (ELKH Data Repository Platform) project introduces into the Data Repository Platform. The RO-Crate format will be presented, as well as how researchers will encounter it in their activities within the repository.
**Keywords:** FAIR principles, data repository, FAIR Digital Object, RO-Crate, metadata

## Bevezetés

A FAIR irányelvek [1], elsősorban az OpenScience kapcsán, egyre inkább előtérbe kerülnek. A betűszó a Findable, Accessible, Interoperable és Reusable angol szavakból áll

össze, és kicsit egyszerűsítve azt írják le, hogy a kutatási adatoknak ahhoz, hogy a kutatás tisztasága és ellenőrizhetősége biztosítva legyen, megtalálhatónak, egyszerű eszközökkel hozzáférhetőnek és feldolgozhatónak kell lenniük, valamint a kutatási eredményekhez vezető út ellenőrizhető módon megismételhető kell legyen. Belátható, hogy az OpenScience kezdeményezés örömmel tűzte zászlajára ezeket az elveket, hiszen ezek leegyszerűsítik a már publikálásra került adatok feldolgozását minden hozzáértő személy számára (még akkor is, ha a FAIR irányelvek követése nem garantálja a kutatási adatok közkinccsé tételét [3]). A FAIR irányelvek kapcsán egy fontos kitételt meg kell még említeni: az irányelveknek való megfelelést nem csak emberi feldolgozás esetén kell teljesíteni, hanem automatikus eszközök számára is biztosítani kell a hozzáférhetőséget és esetenként az automatikus feldolgozhatóságot is. Ez utóbbi kitétel megvalósulása viszont olyan szemantikus címkézését feltételezi a kutatási adatoknak, ami a jelenleg használt adatrepozitóriumokban nem, vagy csak nagyon sok kompromisszum árán valósulhat meg. Az ELKH ARP [13] repozitóriumában bevezettük az RO-Crate formátum olyan támogatását, amivel az adatrepozitóriumok ezen hiányossága áthidalható.

**FAIR digitális objektumok**

Az Európai Unió egy akciótervben 2018-ban bevezette a FAIR Digitális Objektumok fogalmát [2]. A FAIR Digitális Objektumok (FAIR Digital Object - FDO) olyan digitális objektumok, amik adott környezetben megvalósítják a FAIR irányelveket: „Az adatok, szoftver és más erőforrások reprezentációja.”... „Társítva vannak hozzá perzisztens azonosítók, metaadatok és kontextuális dokumentáció, ami lehetővé teszi a felderíthetőséget, idézést és újrahasznosítást.”

Ez a definíció, ismételten csak kicsit egyszerűsítve azt jelenti, hogy az FDO-nak az adott digitális környezetben meg kell tudnia mondania magáról mind emberek, mind pedig automatikus feldolgozó eszközök számára, hogy mi is valójában. Az újrahasznosítás és reprodukálhatóság kritériuma miatt olyan szintű metaadatolást kell tudni biztosítani, amivel formálisan meghatározhatóak a kutatási adatok feldolgozásához szükséges lépések, valamint megjelölhető mind a forrásadatok, mind pedig a feldolgozás eredménye is. Már ez a kritérium is olyan terhet ró az egyszerű generikus adatrepozitóriumokra, ami nehezen teljesíthető, ott ugyanis jellemzően a repozitálás és formális metaadatolás szintje az adatcsomag, amiben az egyes fájlokról nehéz megállapításokat tenni.

**RO-Crate**

Az RO-Crate (Research Object Crate) [4] csomagolástechnika az FDO egy lehetséges megvalósítása, amit az ELKH ARP projekt során kiválasztottunk a FAIR irányelvek támogatására. Kifejlesztése során a Kutatási Objektumok (Research Object - RO [11]) alapelveket ötvözték a DataCrate [5] csomagolással.
Az RO alapelvek előtérbe helyezik az azonosíthatóságot, az aggregációt és az annotációt:

- Azonosíthatóság
  Minden egyes objektumnak valamilyen egyedi azonosítója van.
- Aggregáció
  A kutatások eredménye nem csak maga a publikáció, hanem hozzá kell érteni a kutatás teljes folyamatában mindent, a forrásadatoktól a köztes lépéseken át a végső konklúzióig.
- Annotáció
  Mindenről, ami része a Kutatási Objektumnak, szemantikusan meg kell tudni mondani, hogy mi. Ez az eredeti Kutatási Objektum koncepció esetén Schema.org szemantikus annotáció társítását jelentette az egyes objektumokhoz.

A DataCrate egy csomagolástechnológia, ami az adatfájlok egységbe foglalását, tömörített tárolását (Bagit technológia [6], illetve JSON-LD [7] formátumú metaadatleírás), és szintén Schema.org annotációját jelentette.

A RO-Crate tehát az RO és DataCrate egyfajta evolúciója. Ez az eredeti RO elvekhez képest praktikus egyszerűsítéseket tartalmaz (elegendő csak a kutatás célirányos leírásához szükséges fájlokat/objektumokat megfelelően annotálni), valamint nem ragaszkodik kifejezetten a Schema.org annotációkhoz, bármilyen publikus séma szerinti annotációt megenged a szemantikus leírásokhoz. Technikailag egy tömörített hierarchikus fájl-halmazt jelent, melyet egy JSON-LD leírás lát el akár fájl szintű metaadatokkal. Koncepcionálisan a következő elemekből áll:

- Adat entitások
  - Könyvtárszerkezet (kezdve egy lokális root elemmel);
  - Fájlok ebben a könyvtárszerkezetben;
  - Távoli URI-kkal beazonosítható objektumok.
- Kontextuális entitások
  - Olyan entitások, amik a digitális világon kívül is léteznek (pl. emberek, helyek);
  - Elsősorban metaadat formájában létező leírások (pl. geokoordináták).
- JSON-LD leírás
  - Összekapcsolja az adat és kontextuális entitásokat valamilyen publikus séma szerint tipizálva azokat.

Az RO-Crate felépítése a következő [8]:

```
<RO-Crate gyökér könyvtár>/
|  ro-crate-metadata.json    # RO-Crate Metadata Fájl – kötelező elem
|  ro-crate-preview.html     # RO-Crate Website honlap – javasolt elem
|  ro-crate-preview_files/   # Javasolt elem(ek)
|   | [other RO-Crate honlap fájlok]
|  [fájlok és könyvtárak]  # 0 vagy több
```

Tipikus felhasználási esetben adott kutatás adatai egy hierarchikus fájlrendszerbe kerülnek, ehhez kapcsolódnak a teljes adathalmazt leíró metaadatok (RO-Crate gyökér szintű metaadatleírás), illetve az egyes könyvtárakat, fájlokat leíró metaadatok, valamint olyan erőforrás-leírások, amik URI-kon keresztül beazonosíthatóak és nem kerülnek közvetlenül bele a kutatási adatok közé.

**RO-Crate JSON-LD leírás**

A kutatási adatok valamilyen hierarchiába rendezése/rendeződése gondos adatmenedzsment, tervezés útján magától is kialakulhat, és nem különösebben különbözik attól, ahogy egy generikus adatrepozitóriumba történő feltöltés során az adatok formáját és hierarchiáját elképzelhetjük. Az annotáció formája viszont már jelentős hozzáadott értéket képvisel ehhez az alapkoncepcióhoz képest. Ez az annotáció az RO-Crate gyökér könyvtárában található ro-crate-metadata.json fájlban valósul meg. Ez egy egyszerű flat JSON-LD leírás (azaz az adat és kontextuális entitások vektorszerűen vannak felsorolva benne szerializálva, egyedi azonosítókkal ellátva, és a hierarchikus struktúrát ezeknek az objektumoknak a leírása, és egymásra történő hivatkozása adja). Ettől az egyszerűsített leírástól az RO-Crate technológia fejlesztői a konkrét implementációk megkönnyítését remélik. A fejlesztők továbbá azt a megkötést is alkalmazták, hogy bár bármilyen ontológia szerint annotálhatóak a leírásban található entitások, de az annotáció URI-ja mellett azok címkéjének a megadása is kötelező. Ez azt eredményezi, hogy külön navigáció nélkül is értelmezhető az RO-Crate-ek tartalma emberi olvasók számára is.

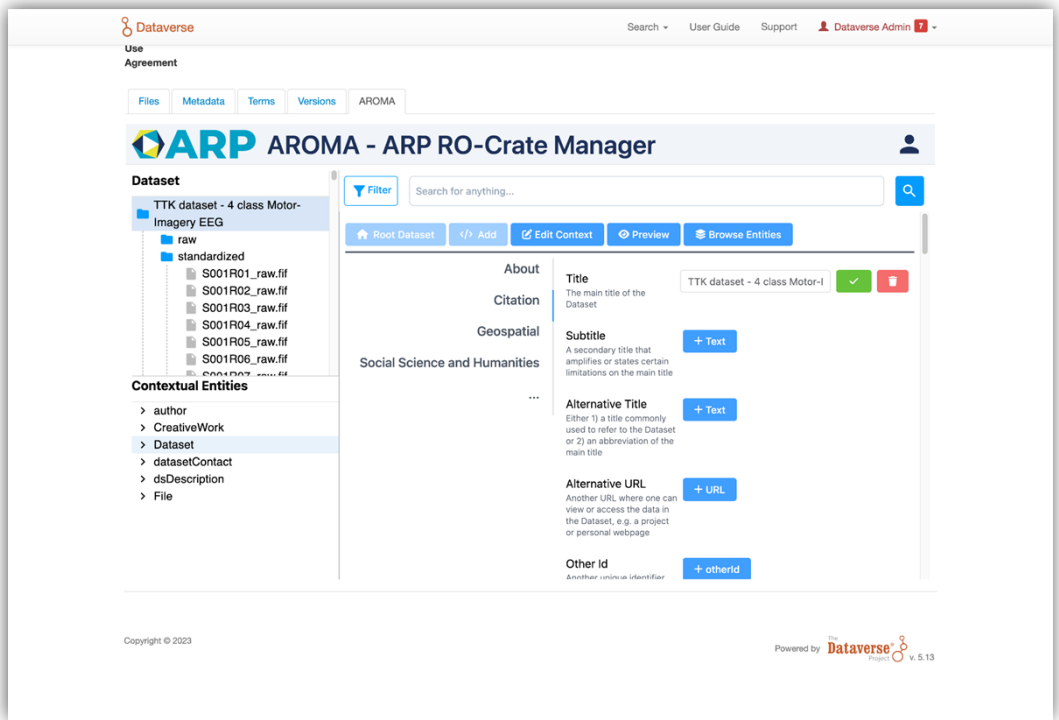
## RO-Crate csomagok támogatása az ELKH ARP repozitóriumában

A generikus repozitóriumok lehetőséget adnak, hogy abban a felhasználók adatcsomagokat hozzanak létre, az adatcsomagokban fájlokat helyezzenek el (akár valamilyen hierarchiába rendezve azokat), és az adatcsomag egészére metaadatleírást adjanak. Ebben az értelemben az RO-Crate adatcsomagok, mint fájlok is elhelyezhetőek bennük. Az RO-Crate adatcsomagok gyökér szintű metaadatleírása vonatkozik a teljes adatcsomagra, azaz ez pontosan megfeleltethető lehetne a repozitóriumokba kerülő adatcsomagok metaadatainak. A jelenleg használt adatrepozitóriumokba RO-Crate csomagokat feltöltve a fájl szintű metaadatok nem kerülnek értelmezésre (az RO-Crate adatcsomag egyetlen fájlnak fog látszani), az RO-Crate gyökér szintű metaadat leírása pedig teljesen különválik az adatcsomag metaadatleírásától. Az ELKH ARP repozitóriuma ezekre a hiányosságokra, valamint a fájl szintű metaadatok kereshetőségére ad megoldást.

A mi értelmezésünkben egy RO-Crate adatcsomag megfelel egy repozitóriumi adatcsomagnak, azaz egy-egy kutatás aggregált, publikálásra szánt eredményének. Az ELKH ARP repozitóriumának alapja egy Dataverse adatrepozitórium [9], és ebben a szokásos repozitóriumi funkciók mellett lehetőséget biztosítunk RO-Crate adatcsomagok importjára, exportjára, valamint helyben történő szerkesztésére is.

Import során létrejön egy olyan repozitóriumi adatcsomag, aminek az adatcsomag-szintű metaadatai kitöltésre kerülnek az RO-Crate gyökér szintű metaadataival. Ennek feltétele, hogy a metaadatolásra használt séma az általunk üzemeltetett, és egyébiránt a felhasználóink által is bővíthető sémaregiszterből (CEDAR [10]) származzon. Az importáláskor kibontásra kerül az RO-Crate, és létrejön a fájlok szokásos könyvtárhierarchiája a repozitóriumban. Ezzel együtt egy külön szerkeszthető objektumként megtekinthető az RO-Crate metaadatleírás, melyben elvégezhető a fájl-szintű metaadatolás.

Akár RO-Crate-ként, akár egyszerű repozitóriumi műveletekkel lett létrehozva egy-egy adatcsomag, annak az RO-Crate jellegű exportja szintén megoldott a rendszerből. A leírás megengedő ebből a szempontból, azaz ha nem történt meg a fájl szintű metaadatolás, az egyébként kötelező RO-Crate gyökér-szintű metaadatolás akkor is létrejön az adatcsomag metaadataiból, és ez kerül bele az RO-Crate JSON-LD leírásába.




*1. ábra: A Dataverse installációba integrált ARP RO-Crate Manager*

A repozitórimban lehetőséget biztosítunk az RO-Crate metaadatok bármilyen szintű szerkesztésére az általunk fejlesztett AROMA (ARP RO-Crate Manager) komponens segítségével (1. ábra). Ez egy faszerű nézetet biztosít az adott adatcsomaghoz, melyben az egyes elemekhez megadhatóak a megfelelő metaadatok. Amellett, hogy így fájl szintű metaadatok megadása is lehetséges, az is megoldott, hogy a RO-Crate gyökér szintű metaadatainak módosítása közvetlenül módosítja a Dataverse adatcsomag szintű metaadatleírásokat is, illetve ez fordítva is megtörténik, azaz a Dataverse adatcsomagleírás változtatása közvetlenül módosítja a kapcsolódó RO-Crate metaadatokat.

A metaadatokban történő keresés a Dataverse lehetőségeinek felel meg a repozitórium felületén belül, azaz közvetlenül kereshetőek a teljes adatcsomagra vonatkozó metaadatok. Ezen felül viszont a platform részét képezi egy keresőfelület (közös kereső), ami tudásgráffá konvertálja az RO-Crate metaadatleírásokat, és amin keresztül kereshetővé válnak a fájl szintű metaadatok is.

**Felhasználói esetleírás**

Az ELKH ARP repozitóriumának előzménye a SZTAKI DSD (Számítástechnikai és Automatizálási Kutatóintézet - Elosztott Rendszerek Osztály) által üzemeltetett Dataverse alapú CONCORDA (Concentrated Cooperation on Research Data) adatrepozitórium [12]. Ebben felmerült az a felhasználói igény, hogy egy kutatás során létrejött nagy mennyiségű képi adatot kereshető formában el lehessen látni geokoordinátákkal. A geokoordináták társítására lehetőséget biztosít a Dataverse is, de csak adatcsomag szinten. Az alkalmazott “rossz gyakorlat” az volt, hogy felhasználóink egy fájlos adatcsomagokat hoztak létre, melyekhez adatcsomag szinten társították a kívánt adatokat. Ettől százas nagyságrendben jöttek létre olyan adatcsomagok, amik hivatkozása publikáció esetén meglehetősen nehézkes.

Ez a “rossz gyakorlat” teljes egészében kiváltható az új rendszerben az RO-Crate ábrázolással. Itt már lehetséges a publikálásnak megfelelő fájl-aggregáció, majd a fájlok egyenként történő metaadatolása, azaz a képállomány ellátása a megfelelő metaadatokkal. A kereshetőséget a geokoordinákra a közös kereső felülete biztosítja.

**Konklúzió**

A FAIR irányelvek követése a repozitóriumokkal szemben új elvárásokat támaszt. Az ezeknek való megfelelés új eszközkészletet igényel, és egy ilyen eszköz az FDO-k implementálása, az RO-Crate csomagolás bevezetése. Ezt, valamint a keresést segítő infrastruktúrát vezeti be az ELKH ARP repozitóriuma a kutatási adatok kezelésére Magyarországon, amivel világszinten is előremutató szolgáltatás jön létre. Az ELKH ARP platform jelenleg fejlesztési és tesztelési fázisban van, a fejlesztés várható befejezése 2023. december vége.

**Bibliográfia**

[1] Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

[2] European Commission, Directorate-General for Research and Innovation, Turning FAIR into reality : final report and action plan from the European Commission expert group on FAIR data, Publications Office, 2018, https://data.europa.eu/doi/10.2777/1524

[3] Putu Hadi Purnama Jati, Yi Lin, Sara Nodehi, Dwy Bagus Cahyono, Mirjam van Reisen; FAIR Versus Open Data: A Comparison of Objectives and Principles. Data Intelligence 2022; 4 (4): 867–881. doi: https://doi.org/10.1162/dint_a_00176

[4] Stian Soiland-Reyes, Peter Sefton, Mercè Crosas, Leyla Jael Castro, Frederik Coppens, José M. Fernández, Daniel Garijo, Björn Grüning, Marco La Rosa, Simone Leo, Eoghan

Ó Carragáin, Marc Portier, Ana Trisovic, RO-Crate Community, Paul Groth, Carole Goble (2022): Packaging research artefacts with RO-Crate. Data Science 5(2) https://doi.org/10.3233/DS-210053
[5] The DataCrate specification for packaging research data, https://github.com/UTS-eResearch/datacrate (letöltve: 2023.06.18.)
[6] Kunze, J., Littman, J., Madden, E., Scancella, J., and C. Adams, The BagIt File Packaging Format (V1.0), RFC 8493, DOI 10.17487/RFC8493, October 2018
[7] JSON for Linking Data, https://json-ld.org/ (letöltve: 2023.06.18.)
[8] RO-Crate Structure, https://www.researchobject.org/ro-crate/1.1/structure.html (letöltve: 2023.06.18.)
[9] Harvard Dataverse Repository, https://dataverse.harvard.edu/ (letöltve: 2023.06.18.)
[10] CEDAR - Center for Expanded Data Annotation and Retrieval, https://more.metadatacenter.org/ (letöltve: 2023.06.18.)
[11] Bechhofer, S., De Roure, D., Gamble, M. et al. Research Objects: Towards Exchange and Reuse of Digital Knowledge. Nat Prec (2010). https://doi.org/10.1038/npre.2010.4626.1
[12] CONCORDA - Concentrated Cooperation on Research Data, https://concorda.hu/ (letöltve: 2023.06.18.)
[13] ELKH Adatrepozitórium Platform, https://science-research-data.hu/ (letöltve: 2023.06.18.)

# ÚJ TECHNOLÓGIÁKKAL, ÚJ TARTALMAKKAL A JÖVŐ DIGITÁLIS TRANSZFORMÁCIÓJA FELÉ

## 32. Networkshop: országos konferencia

2023. április 12–14.
Pannon Egyetem, Veszprém

# ÚJ TECHNOLÓGIÁKKAL, ÚJ TARTALMAKKAL A JÖVŐ DIGITÁLIS TRANSZFORMÁCIÓJA FELÉ

## 32. Networkshop: országos konferencia

2023. április 12–14.
Pannon Egyetem, Veszprém

Szerkesztette: Tick József, Kokas Károly, Holl András

HUNGARNET Egyesület
Budapest, 2023


Szerkesztette: Tick József, Kokas Károly, Holl András

Tipográfia és tördelés: Vas Viktória

Networkshop
2023. április 12–14. Pannon Egyetem, Veszprém konferencia előadásainak közleményei

ISBN 978-615-82243-1-4
DOI: 10.31915/NWS.2023

Kiadja a HUNGARNET Egyesület
az MTA Könyvtár és Információs Központ közreműködésével
Budapest
2023

Borítókép: freepik.com

## TARTALOMJEGYZÉK