



Adatszörnyek megszelídítése

GYORSTALPALÓ A KUTATÁSI ADATOK
KEZELÉSÉRŐL BÖLCSÉSZEKNEK

Scheibner Tamás

tudományos főmunkatárs,
HUN-REN BTK TTI

ARP nagykövet

Kutatási adat

- táblázatok, grafikonok, térképek
- rajzok, fotók
- audio- és videó felvételek, azok szöveges leiratai
- jegyzetek, vázlatok (pl. fogalomtár)
- taxonómiák: olyan hierarchikus osztályozási rendszer, amely a fogalmakat vagy entitásokat közös jellemzők vagy attribútumok alapján csoportokba vagy kategóriákba rendezi
- bibliográfiák
- kérdőívek, interjú vezérfonalak, kódkönyvek, válaszokból létrehozott adatfájlok (pl. oral history)
- szövegtörzsek
- módszerek és munkafolyamatok leírásai

Mi az adat?

- ontológiák: a fogalmak, entitások és tulajdonságaik közötti kapcsolatok formális leírása (gépi olvashatóság)
- kutatási folyamat során megszerzett és/vagy előállított digitális – vagy nem digitális - tárgyak
- modellek, algoritmusok, kódok, szkriptek, kutatásra kifejlesztett szoftverek

Metaadat

adatok az adatról

Pl. cím, készítő, dátum, méret, fájl típus, eredettörténet, hozzáférés és felhasználás feltételei

Metaadat sémák

metaadat elemek meghatározott készlete (elemkészlet) és az ahhoz kapcsolódó szabályok

Dublin Core

1.	Cím	az adatfájl/adatcsomagnak adott név
2.	Létrehozó	az adatfájl/adatcsomag szellemi tartalmának alkotója vagy alkotói
3.	Tárgy- és kulcsszavak, jelzetek	az adatfájl/adatcsomag tartalmát leíró tárgyszavak, kifejezések
4.	Leírás	az adatfájl/adatcsomag tartalmának ismertetése, összegző leírása
5.	Kiadó	az adatfájl/adatcsomag nyilvánosság felé való közvetítéséért felelős entitás
6.	Közreműködő	az adatfájl/adatcsomag szellemi tartalmának létrehozásában részt vevő, vagy annak valamely részét létrehozó személy(ek) vagy intézmény(ek)
7.	Dátum	az adatfájlhoz/adatcsomaghoz kapcsolódó esemény(ek) időpontja
8.	Típus	az adatfájl/adatcsomag típusa, műfaja, általános sajátosságai
9.	Formátum	az adatfájl/adatcsomag fizikai vagy digitális megjelenési formája, kiterjesztése
10.	Azonosító	az adatfájllra/adatcsomagra utaló egyértelmű hivatkozás, az adatfájlt/adatcsomagot egyértelműen azonosító kód
11.	Forrás	hivatkozás arra az eredeti információforrásra, amelyből a jelen adatfájl/adatcsomag származik
12.	Nyelv	az adatfájl/adatcsomag tartalmának nyelve
13.	Kapcsolat	hivatkozás az adatfájllal/adatcsomaggal kapcsolatban lévő más információforrásra (pl. publikációra)
14.	Térbeli és időbeli lefedettség	az adatfájl/adatcsomag tartalmának térbeli vagy időbeli vonatkozásai, kiterjedése
15.	Jogok	információk az adatfájl/adatcsomag felhasználásával kapcsolatos jogokról

Miért fontos az adatkezelés?

A helyes adatkezelési gyakorlat előnyei:

- ✓ lehetővé teszi a metaadatok révén mások számára a kutatási projekt felfedezését,
- ✓ megőrzi a projekt egységét és érthetővé teszi a kontextust,
- ✓ javítja a kutatómunka hatékonyságát (pl. átgondolt munkamenet),
- ✓ elkerülhetővé teszi az adatvesztést.

A kutatás életciklusának *minden szakaszában* megfelelő adatmenedzsmentre van szükség!

Adattárolás

HERE	NEAR	FAR
saját számítógép	lokális adathordozó pl. külső merevlemez, intézeti szerver	távoli szerver pl. repozitórium

Az adattárolás módjának, eszközének kiválasztásakor figyelembe kell venni, hogy

- mennyi ideig szeretnénk vagy kell tárolni az adatokat,
- mekkora adatmennyiségről van szó,
- milyen körrel, hány emberrel, milyen eléréssel szeretnénk megosztani az adatokat,
- vannak-e szenzitív adatok az adatok között,
- mekkora erőforrás áll rendelkezésre,
- vannak-e intézményi vagy egyéb előírások az adatok tárolásával kapcsolatban.

Findable



MEGTALÁLHATÓ

Rendelkezik egyedi azonosítóval (pl. DOI)

Metaadatolt

Regisztráltak kereshető nyilvántartásban (pl. repozitórium)

Accessible



HOZZÁFÉRHETŐ

Elterjedt formátumú adatok

Szabványos kommunikációs protokoll teszi géppel olvashatóvá az adatokat (pl. HTTP)

Metaadatok adatvesztés esetén is elérhetőek

Interoperable

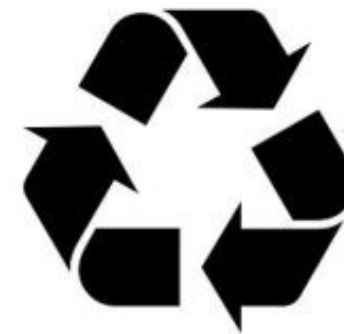


SZABVÁNYKÖVETŐ

Metaadatok közzététele elterjedt kódrendszerben (formális nyelven) történik (pl. RDF, OWL)

Más metaadatokra kereszthivatkozások

Reusable



ÚJRAFELHASZNÁLHATÓ

Egyértelmű felhasználási engedély

Adatok eredete és élettörténete jól dokumentált

Meta/adatok megfelelnek a tudományági szokásoknak

Miért jó adatrepozitóriumban tárolni az adatokat?

- ✓ Hosszú távú és biztonságos adatmegőrzés
- ✓ Biztonságos (nyílt vagy zárt) adatmegosztás lehetősége
- ✓ Adat újrafelhasználhatóságának biztosítása
- ✓ Kutatás láthatóságának növelése
- ✓ Pályázatoknak, intézményi előírásoknak, folyóiratok elvárásainak való megfelelés

Mit repositáljunk?

Olyan kutatási adatokat, amelyek önmagukban nem vagy nem megfelelően értelmezhetőek, de más kutatás számára hasznosak lehetnek **megfelelő kísézőmagyarázattal** ellátva.

A repositóriumba helyezendő **csomag** tartalma:

- kutatási adat,
- kutatási dokumentáció,
- readme fájl,
- metaadat,
- kutatási algoritmus, szoftver, modell (ha van).

Hogyan repositáljunk?

Olvashatóságot és újrafelhasználhatóságot segítő szempontok:

- megfelelő kutatási adatok és kutatási dokumentáció kiválasztása,
- megfelelő fájlformátum,
- megfelelő címkézés,
- megfelelő adat- és fájlstruktúra,
- verziókövetés jelölése,
- readme fájl,
- megfelelő metaadatolás,
- jogok és etikai szempontok (pl. GDPR).

Adatkezelési terv

Projekt és adatközlemény

Projekt profil

Projekt neve:

Projektmegvalósítás ideje:

Projekt finanszírozója:

Projekt rövid leírása:

Vezető kutató(k):

Kijelölt adatkezelő kutató(k):

Adatközlemény alapadatai

Cím:

Alcím:

Szerző(k) neve, email címe, ORCID azonosítója:

Kulcsszavak:

Adatközlemény mérete (várható):

Kapcsolódó publikációk (Chicago Style):

Kontextus és feldolgozás

Adatközlemény leírása

Mutassa be a létrehozni vagy gyűjteni tervezett adatokat, beleértve azok formátumát, szerkezetét, és hogy milyen célból jön létre az adategyüttes!

Adatok gyűjtése

Milyen módszerrel gyűjtik az adatokat (beleértve a meglévő adatok újrahazsnosítását)?

Milyen etikai és jogi megfontolások merülhetnek fel az adatgyűjtéssel kapcsolatban?

Adatok rendezése, előzetes feldolgozása

Hogyan rendezik az adatokat?

Milyen adattranzformációkat vagy egyéb, a feldolgozást lehetővé tevő lépéseket végeznek az adatokon?

Adatformátumok

Szöveg: #pdf#

Kép: #jpg# vagy #png#

Audio: #mp3#

Video: #mp4#

Térkép: #shp/dbf/shx# vagy #json# (#geojson#)

Egyéb, mégpedig:

Adattárolás, adatmegosztás és adatbiztonság 1

Hol tárolják az adatokat a kutatás alatt?

Ki férhet hozzá az adatokhoz a kutatás alatt?

Hol tárolják (archiválják) majd az adatokat a kutatás lezárulta után?

A projekt lezárulta után mely adatok lesznek nyíltan hozzáférhetőek?

A projekt lezárulta után mely adatok nem lesznek nyíltan hozzáférhetőek?

Miért?

Léteznek-e olyan adatmegosztási megállapodások, licencek vagy szellemi tulajdonjogi korlátozások, melyek érintik az adatokat (pl. levéltári anyagokból létrehozott adategyűttesek esetében)? Ha igen, melyek ezek?

Adattárolás, adatmegosztás és adatbiztonság 2

Megfelelnek-e az adatok a vonatkozó adatvédelmi szabályoknak és törvényeknek vagy intézményi irányelveknek?

Milyen adathozzáférési korlátozásokat szeretne felállítani?

Szeretne-e, és ha igen, mely adatokra és milyen időtartamú embargót szeretne beállítani?

Ha az adatagyűttes érzékeny vagy személyes adatokat tartalmaz, milyen biztonsági és bizalmas kezelést biztosító intézkedéseket kíván életbe léptetni? Határozza meg az alkalmazandó anonimizálási vagy azonosítást megszüntető technikákat!

Adatminőség-biztosítás és workflow

Hogyan ellenőrzik az adatminőséget a projekt során? Hogyan történik az adatok validálása és javítása?

A kutatócsoportból ki dönti el, milyen adategyüttes érdemes a megőrzésre?

Ki metaadatolja a kutatási adatokat?

Milyen metaadatsablont használnak?

Ki szerzi be a fenti pontokhoz tartozó információkat?

Ki tölti fel, gondozza és frissíti az adatokat?

CONCORDA / ARP REPO

Alapok

- Regisztráció nélkül is látszanak a közzétett adatcsomagok (és metaadatok)
- Bárki regisztrálhat
- Regisztrálás után sincs sokkal több lehetőség automatikusan, de innentől lehet jogokat kérni és kapni
- Regisztrációval + tárolóhoz való hozzáféréssel lehet feltölteni, szerkeszteni, publikálni...

Felépítés:

tárolók (dataverse) > tárolók és/vagy adatcsomagok (dataset) > fájlok

- **Tárolók:**

- Alapvetően szervezeti egységek vagy projektek számára
- Hozzáférési/menedzselési jogosultságok kezelésére

- **Adatcsomagok:**

- Egy kutatáshoz vagy egy kutatás valamilyen egységéhez tartozó fájlok számára
- Részletes metaadatok

- **Fájlok:**

- Leírásuk lehet, de gazdag metaadatolásuk nincsen

A repozitálás lépései

0. Célok átgondolása, fájlok előkészítése

1. Regisztrálás
2. Hozzáférés kérése a saját intézményi tárolóhoz – Contact
3. Saját tároló vagy adatcsomag feltöltése – Add Data
4. Metaadatmezők kitöltése
5. Fájlok feltöltése (mappa beállítása)
6. Mentés
7. További beállítások – hozzáférések, személyek+szerepek hozzárendelése, privát URL, további metaadatok...
8. Publikálás

Lehet később módosítani a feltöltött anyagot? Lehet. Új verziószámmal.
Publikálás után lehet visszavonni, törölni? A metaadatok örökre megmaradnak.

<https://science-data.hu/>

Segítek!



Miben lesz más az ARP, mint a Concorda

Concorda -> ARP

- További metaadatsémák
- Fájlszintű metaadatolás és kereshetőség
- Közös Kereső – keresési lehetőség nem csak az ARP-ben, hanem más adatrepozitóriumok metaadatai között is
- RO-Crate import/export



**Köszönöm a
figyelmet!**

scheibner.tamas@abtk.hu