# A kutatási adatok FAIR kezelésének alapjai és az adatkezelési terv

Holl András
MTA Könyvtár és Információs Központ

***ELKH Cloud / HRDA, 2021 május 20.***

***Az előadás bemutatja a FAIR kutatási adatkezelés legfontosabb lépéseit annak érdekében, hogy európai és hazai pályázatok számára megfelelő adatkezelési tervet (Data Management Plan) lehessen készíteni.***

***Részletes program:***
***10:00 – 10:45***
- ***Megnyitó***
- ***Bevezetés a FAIR adatkezelésbe***

***10:45 – 11:00 Kávészünet***
***11:00 – 12:00***
- ***Adatkezelési tervek***
- ***Adatkezelési tervek készítése ERC, Horizont Europe és NKFIH ("OTKA") pályázatokhoz***
- ***Kérdések***

# FAIR

| | |
|---|---|
| Data should be **Findable** | F1. (meta)data are assigned a globally unique and persistent identifier (DOI)<br>F2. data are described with rich metadata<br>F3. metadata clearly and explicitly include the identifier of the data it describes<br>F4. (meta)data are registered or indexed in a searchable resource |
| Data should be **Accessible** | A1. (meta)data are retrievable by their identifier using a standardized communications protocol<br>A1.1 the protocol is open, free, and universally implementable<br>A1.2 the protocol allows for an authentication and authorization procedure, where necessary<br>A2. metadata are accessible, even when the data are no longer available |
| Data should be **Interoperable** | I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.<br>I2. (meta)data use vocabularies that follow FAIR principles<br>I3. (meta)data include qualified references to other (meta)data |
| Data should be **Reusable** | R1. meta(data) are richly described with a plurality of accurate and relevant attributes<br>R1.1. (meta)data are released with a clear and accessible data usage license<br>R1.2. (meta)data are associated with detailed provenance<br>R1.3. (meta)data meet domain-relevant community standards |

# Hogyan kezeljük megfelelően az adatainkat?

## ~FAIR adatkezelés [nyílt tudományban] kezdőknek

**Kutatási adatkezelést minden, adatokkal foglalkozó kutató végez. A kérdés, hogy jól csinálja-e?**

PUBLICATIONS AND DATA

https://www.fosteropenscience.eu/content/cartoonpublication-and-data
Auke Herrema – Het Bouwteam

**...minden bizonnyal megfelel a szakma bevett szokásainak. Amivel a modern adatkezelés többet igényel: a rendelkezésre álló adatok és dokumentációjuk alapján független kutatóknak is reprodukálniuk kell tudniuk az eredményeket (ez korábban nem mindig valósult meg, az adatkezelés lépései, és maguk az adatok kevésbé voltak elérhetőek és dokumentáltak). Továbbá a dokumentáció olyan fokát kell elérni, hogy az adatok később, más kutatásokban potenciálisan használhatóak legyenek.**

**A digitális korban az adatok tárolása és megosztása új lehetőségeket biztosít.**

# 12 tanács kutatási adatkezeléshez

## 1.) Át kell gondolni, milyen adatok keletkeznek a projekt során. A H2020, ERC és OTKA adatkezelési terv minták jelzik, milyen kérdésekkel kell foglalkozni a tervezés során

- Szerzők (ORCID), projekt, támogatók, intézmények
- Az adatgyűjtés, megfigyelés célja
- Adatgyűjtés dokumentálása, felhasznált műszerek
- Alkalmazott adatfeldolgozás, szoftverek
- Szoftverek az adatok kezeléséhez, megjelenítéséhez
- Adatok mennyisége, ideiglenes tárolása
- Az adatokra alapozott v. a projektet ismertető publikációk
- Stb.

**2.) [FI] Át kell gondolni, hogyan kell leírni az adatokat ahhoz, hogy a saját kutatás során, illetve később külsősök a kutatás validálása céljából, vagy saját kutatásaikhoz fel tudják használni az adatokat**

**El kell dönteni az adatrögzítésnél használt, fájlformátumokat, adatrendszerezést (pl. mappák), fájlok elnevezési rendszerét.**

**Válasszunk metaadat sémát! Egy lehetőség: Dublin Core (vagy minimális lehetőség: DataCite).**

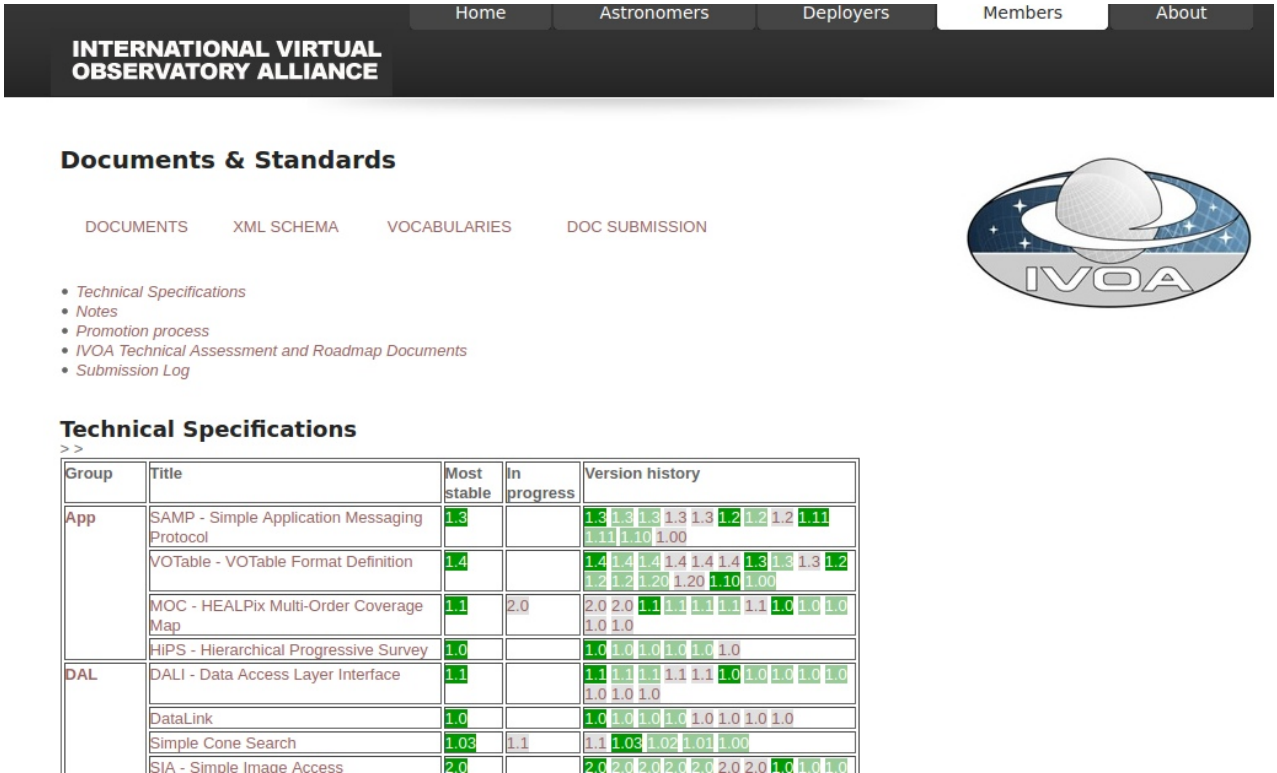
**Dublin Core leírás (OSZK):**
**https://mek.oszk.hu/html/irattar/dc.htm**

**DataCite metaadat séma:**
**https://schema.datacite.org/meta/kernel-4.3/**

**Az adatok szakmai/tartalmi leírásához a DC-nél bonyolultabb sémákra lehet szükség. Fontos szempont, hogy a metaadat-mezők kitöltésénél milyen értékeket engedünk meg, a mezőkbe beírt tartalmak definiáltak, szabványosak legyenek, megfeleljenek a szakma követelményeinek. Bár néhol szabadszavas mezőket szükséges használni, ajánlatos minél nagyobb mértékben jól definiált, publikus, szakmai szabványokon alapuló szótárakra, ontológiákra támaszkodni, és meg is adni, melyiket használtuk. (Ezeket nem csupán az adatcsomagot kísérő leíró adatokban kell használni, de sokszor magukban az adatállományokban is: pl. táblázatos adatoknál az oszlopok megnevezésében, és esetenként az egyes adatcellák lehetséges értékválasztásánál is.**

# Példa a csillagászat területéről

# – International Virtual Observatory Alliance



Home | Astronomers | Deployers | Members | About

INTERNATIONAL VIRTUAL OBSERVATORY ALLIANCE

## Documents & Standards

DOCUMENTS XML SCHEMA VOCABULARIES DOC SUBMISSION

- *Technical Specifications*
- *Notes*
- *Promotion process*
- *IVOA Technical Assessment and Roadmap Documents*
- *Submission Log*

## Technical Specifications

> >

| Group | Title | Most stable | In progress | Version history |
|---|---|---|---|---|
| App | SAMP - Simple Application Messaging Protocol | 1.3 | | 1.3 1.3 1.3 1.3 1.3 1.2 1.2 1.2 1.11 1.11 1.10 1.00 |
| | VOTable - VOTable Format Definition | 1.4 | | 1.4 1.4 1.4 1.4 1.4 1.4 1.3 1.3 1.3 1.2 1.2 1.2 1.20 1.20 1.10 1.00 |
| | MOC - HEALPix Multi-Order Coverage Map | 1.1 | 2.0 | 2.0 2.0 1.1 1.1 1.1 1.1 1.1 1.0 1.0 1.0 1.0 1.0 |
| | HiPS - Hierarchical Progressive Survey | 1.0 | | 1.0 1.0 1.0 1.0 1.0 1.0 |
| DAL | DALI - Data Access Layer Interface | 1.1 | | 1.1 1.1 1.1 1.1 1.1 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 |
| | DataLink | 1.0 | | 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 |
| | Simple Cone Search | 1.03 | 1.1 | 1.1 1.03 1.02 1.01 1.00 |
| | SIA - Simple Image Access | 2.0 | | 2.0 2.0 2.0 2.0 2.0 2.0 2.0 1.0 1.0 1.0 |

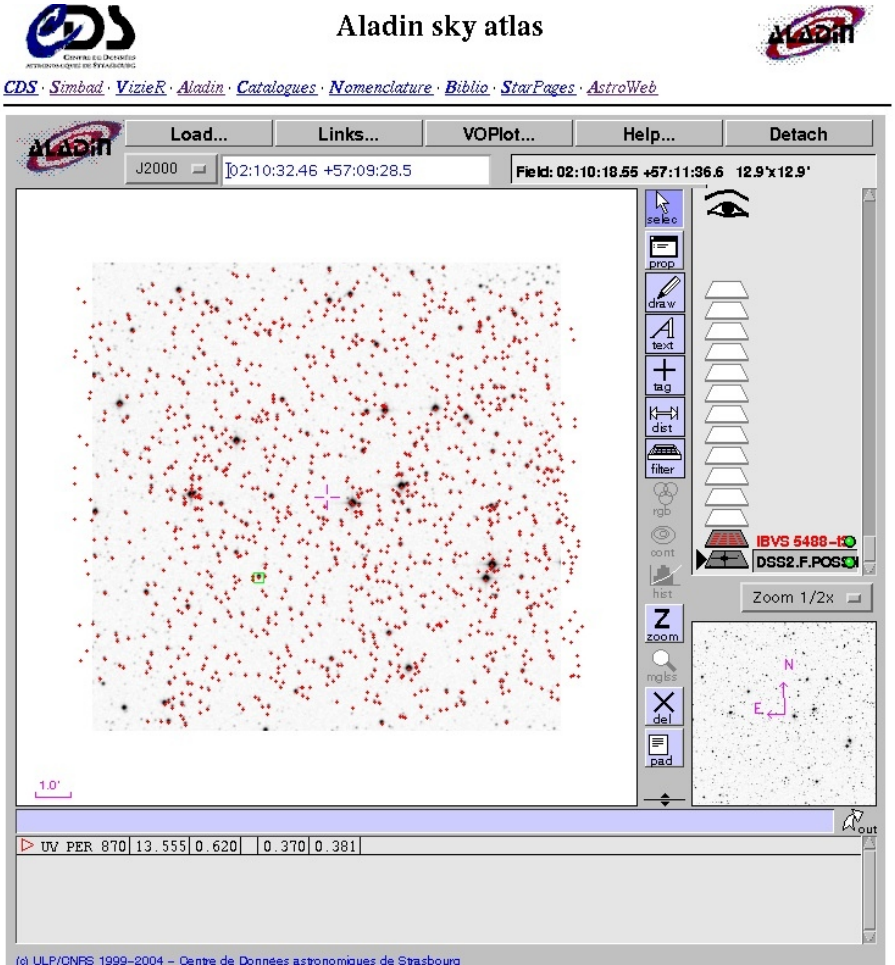
3

Table 2. Photometry of stars in the field of V418 Cassiopeiae

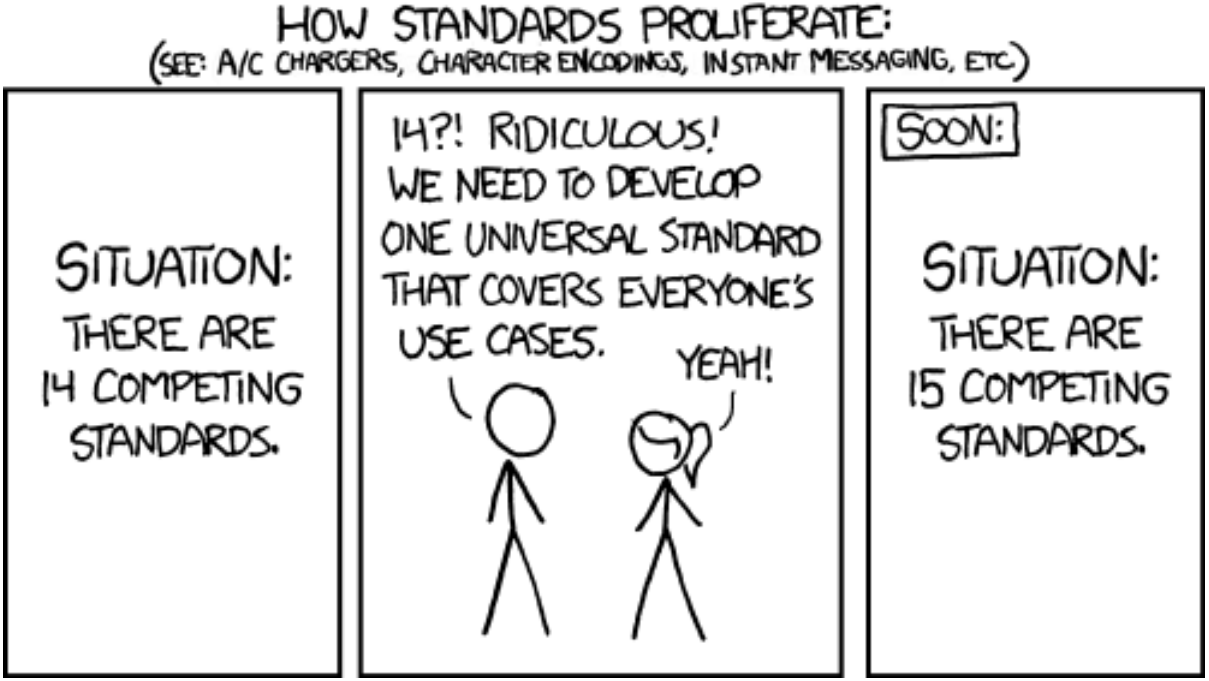
| Name | RA (2000) | Dec (2000) | V | $b-y$ | n |
|---|---|---|---|---|---|
| GSC 4034-0775 | $1^h12^m58^s.9$ | +62°06'56" | 10.91 | 0.44 | 1 |
| | | | .03 | .03 | |
| GSC 4034-0673 | 1 13 34.3 | +62 15 05 | 11.39 | 0.98 | 1 |
| | | | .02 | .02 | |
| GSC 4034-0841 | 1 13 31.1 | +62 09 56 | 11.63 | 0.71 | 2 |
| | | | .00 | .03 | |
| GSC 4034-0873 | 1 12 19.4 | +62 08 57 | 12.26 | 1.12 | 1 |
| | | | .03 | .04 | |
| GSC 4034-1602 | 1 13 33.5 | +62 05 56 | 12.53 | 0.78 | 1 |
| | | | .03 | .04 | |
| GSC 4034-1203 | 1 13 30.5 | +62 04 21 | 12.95 | 0.53 | 1 |
| | | | .04 | .04 | |
| GSC 4034-1542 | 1 13 12.5 | +62 08 57 | 13.27 | 0.68 | 1 |
| | | | .04 | .06 | |

```
#ID: IBVS 3967
# IBVS 3967-t2.txt
#Author: Skiff, B.
#IBVSdataKey: sequence
#Title: PHOTOMETRY OF STARS IN THE FIELD OF THE MIRA V418 CASSIOPEIAE
#UCD:ID_MAIN   POS_EQ_RA_MAIN POS_EQ_DEC_MAIN     PHOT_JHN_V
     PHOT_STR_B-Y
#unit:--- h:m:s     d:m:s     mag  mag
#name:Name     _RAJ2000  _DEJ2000  V         b-y
GSC_4034-0775  1:12:58.9 +62:06:56 10.91          0.44
GSC_4034-0673  1:13:34.3 +62:15:05 11.39          0.98
GSC_4034-0841  1:13:31.1 +62:09:56 11.63          0.71
GSC_4034-0873  1:12:19.4 +62:08:57 12.26          1.12
GSC_4034-1602  1:13:33.5 +62:05:56 12.53          0.78
GSC_4034-1203  1:13:30.5 +62:04:21 12.95          0.53
GSC_4034-1542  1:13:12.5 +62:08:57 13.27          0.68
```

```
−<VOTABLE>
 −<DESCRIPTION>
    Author: Skiff, B. Title: PHOTOMETRY OF STARS IN THE FIELD OF THE MIRA V418 CASSIOPEIAE
    - a Perl script written by A. Sragli, Konkoly Obs.
  </DESCRIPTION>
 −<RESOURCE>
  −<TABLE>
     <FIELD unit="---" datatype="float" name="Name" ucd="ID_MAIN"/>
     <FIELD unit="h:m:s" datatype="float" name="_RAJ2000" ucd="POS_EQ_RA_MAIN"/>
     <FIELD unit="d:m:s" datatype="float" name="_DEJ2000" ucd="POS_EQ_DEC_MAIN"/>
     <FIELD unit="mag" datatype="float" name="V" ucd="PHOT_JHN_V"/>
     <FIELD unit="mag" datatype="float" name="b-y" ucd="PHOT_STR_B-Y"/>
    −<DATA>
     −<TABLEDATA>
      −<TR>
         <TD>GSC_4034-0775</TD>
         <TD>1:12:58.9</TD>
         <TD>+62:06:56</TD>
         <TD>10.91</TD>
         <TD>0.44</TD>
       </TR>
      −<TR>
         <TD>GSC_4034-0673</TD>
         <TD>1:13:34.3</TD>
         <TD>+62:15:05</TD>
         <TD>11.39</TD>
         <TD>0.98</TD>
       </TR>
```


Aladin sky atlas
CDS · Simbad · VizieR · Aladin · Catalogues · Nomenclature · Biblio · StarPages · AstroWeb
Load...
Links...
VOPlot...
Help...
Detach
J2000
02:10:32.46 +57:09:28.5
Field: 02:10:18.55 +57:11:36.6 12.9'x12.9'
select
prop
draw
text
tag
dist
filter
rgb
cont
hist
zoom
mglss
del
pad
IBVS 5488-1
DSS2.F.POSS
Zoom 1/2x
N
E
1.0'
out
UV PER 870 13.555 0.620 0.370 0.381
(c) ULP/CNRS 1999-2004 – Centre de Données astronomiques de Strasbourg

# 3.) [I] Szabványos, nyílt, elterjedt, szabad fájlformátumokat kell választani




https://xkcd.com/927/

**4.) [F] Állandó azonosítókról kell gondoskodni (pl. DOI)**

**A DOI azonosító biztosítása pénzbe kerül. Az MTA KIK a DataCite ügynökségen keresztül tud DOI-kat biztosítani adatoknak. Ehhez megállapodást kell kötni az igénylő intézménnyel. A DOI-val azonosított adatállományt vagy (adat)repozitóriumban kell elhelyezni, vagy biztonsági kópiát kell ott letenni.**

**https://openaccess.mtak.hu/doi/**

**DOI előnyök: i.) állandó elérés; ii.) DOI metaadat-tár; iii.) hivatkozások gyűjtése**

## 5.) [FA] Végleges elhelyezési helyet, adatrepozitóriumot kell találni

- Concorda (SZTAKI) : https://www.sztaki.hu/innovacio/projektek/concorda
- REAL (MTA KIK): http://real.mtak.hu

Ide az egyedi, publikációhoz kapcsolódó, egyszerű formátumú (pl. szöveges), kis mennyiségű adatot lehet elhelyezni.

Nemzetközi repozitóriumok kereshetőek a re3data segítségével: https://www.re3data.org/

Repozitórium választásnál a szakmai szemponton túl érdemes minősített repozitóriumot választani (MTMT, CoreTrustSeal).

**6.) [A] Meg kell választani a hozzáférhetőséget, embargót.**

**Hozzáférhetőség: zárt, korlátozott, nyílt**

**Zárt hozzáférhetőség FAIR alapon: pl. a szerzőktől kell hozzáférést kérni.**

**Az embargó időszak kitolhatja a hozzáférhetőség kezdetét.**

## 7.) [R] Licencet kell választani

Pl.: Creative Commons licencek
https://creativecommons.org/licenses/

Gnu GPL https://www.gnu.org/licenses/gpl-3.0.html

How to License Research Data

A. Ball

https://www.dcc.ac.uk/guidance/how-guides/license-research-data

**8.) Meg kell gondolni a lehetséges kockázatokat (etikai, GDPR, stb.), és ezek kezelését.**

**9.) Mindezt költségelni kell.**

**(2021: ELKH–HRDA támogatási lehetőség!)**

**10.) Meg kell gondolni, mire használhatóak az adatok másoknak?**

**11.) [F] A publikáció(k)ban hivatkozni kell az adatokra, és az adatok dokumentációjában a publikációkra! (Ha csak lehet, DOI–val!)**

# 12.) [F] A publikusan hozzáférhetővé vált, DOI-val rendelkező adatokat nyilvántartásba kell venni az MTMT-ben, és a rájuk érkezett hivatkozásokat is fel kell vinni!

| MTMT közlemény és idéző összefoglaló táblázat | | | | |
|---|---|---|---|---|
| Holl András adatai (2021.05.07) | | | | |
| **Közlemény típusok** | **Száma** | | **Hivatkozások[1]** | |
| Tudományos közlemények | Összes | Részletezve | Független | Összes |

| | | | | |
|---|---|---|---|---|
| **Közlemények összesen (I.-IV.)** | 61 | --- | 239 | 311 |
| **Absztrakt[3]** | 0 | --- | 0 | 0 |
| **Kutatási adat** | 2 | | 5 | 6 |
| **További tudományos művek[4]** | 10 | --- | 248 | 334 |
| **Összes tudományos közlemény** | 73 | --- | 492 | 651 |

# Adatkezelési terv
## (Data Management Plan, DMP)

**Általános tájékoztató előadás, 2020 HRDA meetup sorozat:**

**https://openaccess.mtak.hu/event/az-adatkezelesi-terv-data-management-plan/**

**Az adatkezelési terv élő dokumentum kell legyen, és az egyik módja az adatok dokumentálásának (az adatokkal együtt célszerű archiválni)!**

**Nyertes pályázatok adatkezelési tervei letölthetőek az EU Cordis rendszeréből.**

**H2020 és ERC korábbi ciklusokban pilot ill. opt-out**

# A.) H2020

**Az űrlap letölthető a Zenodo-ból. DOI: 10.5281/zenodo.2635768**

https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

## 1.) Az első blokkban általános információkat kell megadni a projektben keletkező kutatási adatokról.

- Meg kell adni az adatok célját. Itt célszerű azt leírni, amit egy cikkben a bevezetőben megfogalmaznánk: az adott kutatási cél eléréséhez miért volt szükség új adatok előállítására, begyűjtésére?
- A következő táblázat lehetséges adat-

típusokat sorol fel. A „form” az adatok jellege, a „format” a fájl formátum.

- Az adatok eredetét kell megadni a következő táblázatban, úgymint: megfigyelési, kísérleti, szimuláció, származtatott/kompiláció, referencia/kanonikus. Az adatbázisból lekérdezéssel (pl. SQL) előállított adatok a legutóbbi kategóriába tartoznak.
- Az adatok életciklusáról kell információt adni: végleges, dinamikusan növekvő, frissíthető kategóriák szerint.
- Adatok mennyisége. A rubrikák fejléce és a megjegyzés rovat szerint itt a terjedelemre kíváncsiak, de javasolt közelítő „darabszámot” is megadni, amennyiben releváns. Pl. ~10000 kép, egyenként ~100 MB, összesen ~10 GB.

- **A következő kérdés az adattárolásra vonatkozik. Nem nyilvánvaló, hogy a tárolásnál a munkaterületre vagy a végleges, archiválási területre vonatkozik (bár ideális esetben a kettő megegyezhet). Javasoljuk mindkettőt megadni (pl.: munka: labor számítógép merevlemez; RAID lemeztömb. archiválás: adatrepozitórium; adatrepozitóriumi biztonsági politika (URL)).**
- **Az adatok értéke. Nem feltétlenül monetáris értékről van szó: azt kell megadni, kiknek és milyen újrafelhasználási célra lehet fontos az adat.**

**2.) A második blokk az adatok FAIR megfelelésáre vonatkozik, a négy alapkövetelmény szerint (Findable/megtalálható; Accessible/hozzáférhető; Interoperable/szabványos; Reusable/újrafelhasználható).**

**F**

- **Discoverability – itt azt kel megadni, hogy a leíró adatokban (metadata) milyen mezők szerepelnek, amelyekre kulcsszó vagy index-mező szerűen keresni lehet.**
- **Azonosíthatóság – állandó, egyedi azonosítók használata (pl. DOI).**
- **Az állományok, mappák, gyűjtemények elnevezésének alkalmazott szisztémája.**
- **Kulcsszavak használatának szisztémája.**
- **Verziókezelés.**
- **A metaadatok megadásánál használt szabványok**

(pl. szótárak, sémák, ontológiák).

**A**

- Kié az adat? Szabadon hozzáférhető –e?
- Ha nem nyilvános, kik férhetnek hozzá? Mi a korlátozás oka?
- Hogyan lesznek közzétéve az adatok?
- Milyen szoftverdokumentációra van szükség az adatok eléréséhez? (Bár célszerű az elérhetőséget standard, elterjedt, ingyenes, platformfüggetlen módon elérhetővé tenni.)
- Repozitórium.
- Alkalmazott hozzáférési korlátozások.
- Szabványossági szint.

**I**

- Standard szótárak vagy elterjedt ontológiák használata. (Ahhoz, hogy egy globális szolgáltatásban kereshetők legyenek az adatok, nem elegendő metaadatokat

biztosítani, de ezeknek a metaadatoknak valamilyen általánosan, szakmailag elfogadott rendszerbe (szótár, ontológia) illeszkedniük is kell. Pl. régészeti leletek koránál valamilyen nemzetközileg vagy legalábbis regionálisan elfogadott korszak-beosztást kell használni.)

- Az újrafelhasználást lehetővé tevő licenc (pl. Creative Commons, GNU GPL, stb.)

R

- Adatok nyilvánosságra hozatalának időzítése.
- Más kutatók/projektek számára való felhasználhatóság.
- Korlátozások.
- Minőségbiztosítási folyamat.
- Meddig lesznek az adatok újrafelhasználhatóak?

## 3.) Erőforrások

- **Költségbecslés a FAIR-megfelelőségre.**
- **Ki kezeli az adatokat? Ki felelős az adatkezelésért? (Itt projekten belüli felelőst várnak a megjegyzés szerint.)**

## Példa: FlowPhotoChem

## B.) ERC

**ERC tájékoztató:**
**https://erc.europa.eu/sites/default/files/document/file/ERC_info_document-Open_Research_Data_and_Data_Management_Plans.pdf**

**Az űrlap letölthető a https://erc.europa.eu/content/erc-data-management-plan-template oldalról.**
**Az űrlap hat nagy, szabad szövegdobozból áll, amelyek közül négy a FAIR alapelveknek felel meg.**

## 0.) Összegzés

**Itt kell megadni, szabad szöveges mezőben az adatok valamiféle megnevezését, eredetét,**

**várható méretét/mennyiségét, típusát és formátumát.**

**1.) Findable**

**Metaadatok, állandó és egyesi azonosítók (pl. DOI).**

**2.) Accessible**

**Szabadon hozzáférhető ill. zárt adatok, a hozzáférés korlátozásának oka, használt repozitórium, elérési mód.**

**3.) Interoperable**

**Standard (adat és) metaadat-szótárak, standard sémák és formátumok.**

**4.) Reusable**

**Milyen adatok és meddig maradnak újrafelhasználhatóak, lesz-e embargó, milyen licencek szerint lehet az adatokat újrafelhasználni, milyen minőségbiztosítási eljárásokat alkalmaznak?**

**5.) Erőforrások és adatbiztonság**

**Becsült adatkezelési költségek, potenciális újrafelhasználhatóság, biztonsági mentések, kényes adatok kezelése, hosszú távú adatkezelés.**

**Példa: NEUROFLIES**

# C.) OTKA

**Egyes kérdések megválaszolásához fűzünk megjegyzéseket.**

***Milyen típusú adatok keletkeznek a kutatás során?*** **Pl. JPG képek, spektrumok, TEI XML állományok, Excel táblázatok, stb.**

***Milyen eljárással végzik az adatgyűjtést?*** **Pl. megfigyelés, kísérlet, kompiláció, irodalmi adatgyűjtés, adatbázisból való lekérdezés, modellezés, stb.**

***Hol és hogyan tárolják az adatokat, dokumentációkat és gépi kódokat a kutatás lezárulta után?*** **Javasoljuk az**

**adatrepozitóriumban való tárolást. Gépi kód: szoftver, program, szkript.**

***Milyen hozzáférést biztosítanak ezekhez az adatokhoz?*** **Pl. nyílt hozzáférés, egy év embargó után.**

***Milyen metaadat szabványt használnak?*** **A DC általános leírásra szolgál. A DataCite DOI regisztrációhoz meg kell adni bizonyos adatokat, ez lehet egy minimum szabványos metaadat-készlet.**

***Hol és hogyan tárolják a metaadatokat a kutatás során?*** **A szerencsés megoldás, ha már a kutatás során adatrepozitóriumba kerülnek az adatok. Lehet lokális fájlokban, adatbázisban, excel**

**táblában tárolni az adatokat és metaadatokat.**

***Milyen hozzáférést biztosítanak a metaadatokhoz?*** **Javasoljuk a kérdést úgy értelmezni, hogy a hosszú távú tárolásra vonatkozik. Érdeklődni kell a használt adatrepozitórium kapcsolattartóitól. Lehetséges válasz pl.: OAI-PMH protokoll és HTTP protokoll segítségével nyilvánosan elérhetőek a metaadatok, stb.**

***A kutatási adatok kezelése megfelel-e a GDPR előírásainak?*** **A válasz természetesen „igen" kell legyen, de célszerű kibontani: pl. anonimizálás alkalmazásával, személyes adatok mellőzésével.**

**Az adatkezelési terv élő dokumentum, és az adatok dokumentálásának kulcsfontosságú része. Lehetőleg az adatokkal együtt kell archiválni.**

*

**Adatkezelési terv készítését segítő eszköz:**

**dmptool https://dmptool.org/**

**Száldobágyi Ádám diái, NKFIH – EKK HUNOR meetup: https://openscience.hu/wp-content/uploads/2021/01/Kutatasi_adatkezeles_202101_uj.pdf**

holl . andras @ konyvtar . mta . hu