Adatrepozitóriumok általában: alapelvek, példák, jó gyakorlatok

Micsik András

SZTAKI DSD

Miről lesz szó?

- FAIR alapelvek: a nyílt, reprodukálható tudomány érdekében
- Metaadatok szerkezete, fajtái, fontossága
- Perzisztens azonosítók
- Repozitóriumok példákon keresztül

FAIR alapelvek

- Findable - meg tudjuk keresni
- Accessible - le tudjuk tölteni
- Interoperable - meg tudjuk érteni
- Reusable - fel tudjuk használni
- Az alapelvek részletes leírása:
  - https://www.go-fair.org/fair-principles/
- Első közlés 2016-ban:
  - 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

Megtalálható (findable)

- F1. A (meta)adatoknak egyedi, perzisztens azonosítójuk van
  - Pl. DOI, ORCID, URN, stb.
- F2. Az adatok leírása metaadatokkal bőséges (lásd R1 később)
  - Tartalmazza az automatikusan rögzített, valamint a kontextust leíró adatokat is
- F3. A metaadatok egyértelműen tartalmazzák a leírt adat azonosítóját
  - Vagyis az adat és a róla szóló metaadat legyen összerendelve
- F4. A metaadatok kereshetők valamilyen szolgáltatásban
  - Pl. a repozitórium saját keresője is megfelel erre a célra

Hozzáférhető (accessible)

- A1. A (meta)adatok az azonosító alapján letölthetők szabványos és elterjedt módon
  - Tehát nincs szükség különleges szoftverre vagy előkészületekre
  - A1.1 A letöltés módja és kódja nyíltan hozzáférhető, ingyenes és megvalósítható
  - A1.2 A letöltés/hozzáférés során alkalmazható azonosítás és jogosultság ellenőrzés
- A2. A metaadatok elérhetők maradnak akkor is, ha az adatok már nem nyilvánosak
  - A metaadatok akkor is hasznosak, ha az adatokhoz már nem férünk hozzá

Együttműködő (interoperable)

- I1. A (meta)adatok leírására formális, közzétett és széleskörűen alkalmazható tudásreprezentációs formát használunk
  - Példák: JSON-LD, XML séma, RDF
- I2. A (meta)adatok leírására használt szókészletek, leíró sémák maguk is eleget tesznek a FAIR ajánlásoknak
  - Vagyis van perzisztens azonosítója és könnyen megtalálható, letölthető
- I3. A (meta)adatok minősítetten hivatkoznak más (meta)adatokra
  - Minél pontosabban meg kell határozni az adatok kapcsolatát, és a hivatkozott adatot perzisztens azonosítóval megadni

Újrafelhasználható (reusable)

- R1. Az adatok leírása tartalmazzon minél több pontos és releváns leírót
  - A cél az, hogy a felhasználó (ember vagy gép) minél inkább el tudja dönteni, hogy hasznos-e számára az adat.
- R1.1. Egyértelmű és hozzáférhető felhasználási licenc megadása
  - Példák: Creative Commons, Licenses for Research Data
- R1.2. Az adatok eredete és keletkezése jól dokumentált (provenance)
  - Az adatok története, a felhasznált máshonnan származó adatkészletek, az elvégzett módosítások, korrekciók, stb.
- R1.3. Az adatok leírása megfelel a szakmai közösség szokásainak, elvárásainak

A nyílt adatok 5 csillagos osztályozása

Példa egy táblázat megosztása esetén

★	Weben elérhető, nyílt licensszel	PDF
★★	Strukturált, gépileg feldolgozható formában van az adat	Excel
★★★	Nyílt formátumban van az adat	CSV
★★★★	URI-val hivatkozunk a dolgokra	JSON-LD
★★★★★	Az adatokat összekapcsoljuk más adatokkal	RDF vagy JSON-LD

A Linked Open Data alapelvei FAIR adatok esetén is érvényesek!

Nyílt (meta)adat alapelvek

1. Amit lehet URI-val azonosítsunk és hivatkozzunk
   - Pl. ”Cím” helyett *http://purl.org/dc/terms/title*
   - URI = URL vagy URN
2. Az URI legyen feloldható, legyen mögötte tartalom
3. Az URI adjon információt ember és gép által feldolgozható módon
   - Pl. születési hely megadása esetén: https://www.w3.org/ns/person#placeOfBirth
     - Szöveges leírás embereknek
     - RDF séma gépi feldolgozásra
     - Innen továbbnavigáláshoz linkek: Person, Place

Eredmény: az adatok automatikusan is értelmezhetők (nagyjából ☺)

FAIR adatok a CONCORDA-ban

- A CONCORDA a FAIR elvek alapján készült
- Vezeti a kezet a metaadatok kitöltésénél
- Témakör szerinti metaadatleírók
- Választás listából (pl. licenc)
- Néha nem elég részletes a leírás lehetősége (pl. eredet, keletkezés)

Adatleíró sémák

- Más néven: metaadat-sémák
- Főbb dokumentálási feladatok
  - Kereshetőség: ki-mit-mikor-hogyan
  - Csoportosítás: témakör, kulcsszavak, fájltípus, stb.
  - Frissesség: dátumok a dokumentum történetében, frissítési információk
  - Újrafelhasználás: jogok, licenc, felhasználási feltételek
  - Hosszútávú megőrzés (preservation metadata)
  - Történet: verziók, források, felhasználások (provenance)

Adatleíró sémák modellje

- Általános modell:
  - Elemkészlet: megfelel kb. az űrlapmezőknek (element set)
    - Egy elemnek lehetnek altípusai
      - pl. címnek: alcím vagy cím más nyelven
      - Közreműködő altípusai: operatőr, világosító, jelmeztervező, stb.
  - Szókészlet: mit írhatunk a mezőbe (vocabulary)
    - Pl. IMT (fájl típus), UDC (témakör, magyarul ETO)
  - Kódolás: hogyan írjuk a mezőbe (encoding scheme)
    - Pl. ISO 639-3 (nyelv), W3C-DTF (dátum)

Dublin Core 1.

- A legáltalánosabb leíróséma 1995-ből
  - Dublincore.org
- Érdemes áttekinteni az elemkészletet:
    - Title
    - Date
    - Type
    - Format
      - Extent, Medium
    - Language
    - Identifier
    - Creator
    - Contributor
    - Publisher
    - ... folyt. köv.

Dublin Core 2.

- Description
  - Abstract, tableOfContents
- Subject
- Source
- Rights
- Relation
  - Replaces, references, requires, conformsTo, hasPart, hasVersion
- Provenance
- Coverage
  - Spatial, temporal
- Audience

Adatleíró sémák formátuma

- Névtér fogalma (namespace, ns)
  - Rövidítési és csoportosítási módszer
    - Névtér prefix = Névtér URI
    - A csoportba tartozó elemek neve előtt URI helyett prefixet használunk,
    - pl. *http://purl.org/dc/terms/title* helyett *dc:title*
- Lehetséges formátumok
  - XML, JSON-LD, RDF, ...
- Datacite leíró séma (DOI igényléshez)
  - Dokumentáció
  - Példa használatra

Perzisztens azonosítók

- DOI – főleg dokumentumokra használt
  - Több DOI szolgáltató van, különböző használati feltételekkel
- Handle system
  - Általános mechanizmus, a DOI is ezt használja
- ARK: decentralizált megoldás
- Szerzőkre: ORCID
- Intézményekre: GRID
- Feloldás:
  - DOI: doi.org
  - Általános feloldó: https://n2t.net/
    - ARK, DOI, URN, Handle, PMID, PDB, Taxon, GRID, arxiv, ISSN, ...

Repozitórium példák

- Közleményrepozitórium: REAL
  - MTMT kapcsolat, Sword protocol
- Tematikus, vegyes repozitórium: KDK
- Kódrepozitórium: GitHub
- Figshare: all-in-one repository, DOI, 5GB limit
- Myexperiment: kísérletek, folyamatok publikálása
- OSF.io: teljes kutatási projekt-infrastruktúra
- EU által támogatott repozitóriumok
  - Zenodo: DOI, 50GB limit
  - B2SHARE: 20 GB limit

Repozitórium minősítés

- Minősítések, „pecsétek”
  - CoreTrustSeal.org – adatrepozitóriumokra részletes kritériumok
  - MTA KIK Repozitóriumminősítő Szakbizottság - közlemény repozitóriumokat minősít
- A minősítés célja
  - A repozitórium szervezeti, működési és műszaki hátterének átvilágítása
  - „Megbízhatóság garantálása”

Repozitórium választás

- Mi a cél?
  - EU projekt kéri (Data Management Plan)
  - Folyóirat, kiadó kéri
  - Közös munkához szükséges
  - Önkéntes „archiválás”
  - Stb.
- Választáshoz segítség:
  - A Nature kiadó adatrepozitórium ajánlasai: Scientific Data
  - Repozitórium listák, pl. Re3data - REgistry of REsearch data Repositories
  - FAIRsharing.org: repozitóriumok FAIR-sége és támogatottsága
  - Választási szempontok magyarul

Repozitórium választási javaslatok

- Lehetőségek
  - Előírt vagy ajánlott repozitórium használata
  - Tudományterületen szokásos adatrepozitórium használata
  - Intézményi vagy országos repozitórium használata
- Javaslatok
  - Az itthon keletkezett adatokat legalább egy hazai repozitóriumba töltsük fel
  - CONCORDA
    - ingyenes ELKH intézmények számára
    - DOI-t nem tud adni, de folyamatban van a PID megoldása
  - Zenodo
    - DOI automatikusan
    - 50 GB korlát

Adatok keresése

- Hogyan találhatjuk meg a minket érdeklő adatokat?
  - Nincs átfogó, teljeskörű megoldás
- Hazai kereső
  - oaikereso.sztaki.hu hazai repozitóriumokban közleményeket keres
- EU támogatott kereső szolgáltatások
  - https://explore.openaire.eu/search/find
  - http://b2find.eudat.eu/

Mit tudjon egy adatrepozitórium?

- Verziókezelés
- Nagy adattömeg feltöltése
- Metaadatolás (licenc megadása)
- Közzététel, visszavonás
- Hozzáférési jogok szabályozása
- Közös kereshetőség támogatása (pl. OAI alapon)
- Statisztika szolgáltatás

- Ezeket nézzük meg a továbbiakban a CONCORDA esetében...

Köszönöm a figyelmet

Kérdések?