Rendező:

Hungarnet

Networkshop 2020 ONLINE

Házigazda:

PÉCSI TUDOMÁNYEGYETEM
UNIVERSITY OF PÉCS

„Esélyeink és kihívásaink a digitális transzformáció világában”

Országos Online Konferencia
2020. szeptember 2–4.

INNOVÁCIÓS ÉS TECHNOLÓGIAI MINISZTÉRIUM

Szerkesztette: Tick József, Kokas Károly, Holl András

Tipográfia és tördelés: Vas Viktória

Networkshop
2020. szeptember 2-4. Pécsi Tudományegyetem, (On-line)
konferencia előadásainak közleményei

ISBN 978-615-01-0376-1
DOI: 10.31915/NWS.2020

Kiadja a HUNGARNET Egyesület
az MTA Könyvtár és Információs Központ közreműködésével
Budapest
2020

Borítókép: freepik.com

DOI: 10.31915/NWS.2020.1

A kutatási adatok dokumentálását elősegítő szoftverek

Holl András
MTA Könyvtár és Információs Központ
holl.andras@konyvtar.mta.hu
ORCID: 0000-0002-6873-3425

A nyílt tudomány (Open Science, OS) a kutatási folyamat minden összetevőjét hozzáférhetőbbé, láthatóbbá kívánja tenni. A kutatók vállára ezzel újabb terhek kerülhetnek – mind a költségek, mind a feladatok szempontjából. A nyílt tudomány gyakorlata csak akkor terjedhet el, ha ezeket a terheket csökkenteni tudjuk.

A nyílt hozzáférésű tudomány kívánalmainak egyike a megfelelő dokumentálás, ami jelentős részben metaadatok alkalmazását jelenti. Mind a metaadatok létrehozását, mind a kutatási objektumok egyedi azonosítókkal való ellátását megkönnyíti, ha a kutatás során ezeket a feladatokat támogató szoftvereket alkalmaznak.

Előadásunkban az ilyen, OS-támogató vagy OS-barát szoftverek kérdését járjuk körül, a kutatási adatoktól egészen a tudományos publikációkig.

Open Science – meta-data – software

Open Science aims to transform science, increasing accessibility and transparency. Unfortunately, it potentially increases the workload of researchers and the research support personnel, and increases certain cost elements (while potentially reducing others). Open Science could be successful only if the workload and cost issues are managed.

Creation and management of meta-data is an important issue in OS, and another crucial point is the use of permanent identifiers. Handling these issues could only be effective using software that supports such functions.

We deal with questions of such, OS-supporting or OS-friendly software, used from data creation to publication and archiving of research results.

Keywords: Open Science, FAIR research data, workflows, software

Bevezető

A kutatási adatok megfelelő kezeléséhez – amennyiben meg akarunk felelni a nyílt hozzáférésű tudomány elvárásainak – változtatnunk kell a jelenlegi tudományos gyakorlaton. Számos fontos területen kell beavatkoznunk, úgymint:

- az adatkezelési tervek megkövetelése;
- a kutatási adatok archiválásának és megosztásának figyelembe vétele az értékelésben;
- az adatkezelési költségek finanszírozása;
- adatgazdászok alkalmazása a kutatók támogatására;
- minősített adatrepozitóriumok létrehozása és működtetése;
- oktatás;
- a hozzáférhető adatok megkövetelése a közleményekben.

Mindez azonban nem elég, ha az adatkezelés szempontjai nem épülnek be a kutatási folyamatba, az adatok dokumentálását nem segítik a folyamatban használt (szoftver) eszközök.

Nyílt tudomány – Nyílt hozzáférésű kutatási adatok – FAIR kutatási adatok

Az előadás a kutatási adatok kezelésének jó gyakorlatával foglalkozik. Számos tudományban adat-alapú a kutatás, a kutatói képességek, felkészültség része az adatok megfelelő kezelése. De vajon a külvilágnak is megfelelő-e az az adatkezelés, amit a kutató maga megfelelőnek tart? Az elmúlt évtizedben egyre nyomasztóbban merült fel a kutatási eredmények reprodukálhatatlanságának jelensége – mindez arra utal, hogy a publikációkban közölt információk és az adatok nem elégségesek a folyamat reprodukálásához. Még további dokumentációs igényeket vet fel az adatok esetleges újrahasznosításának megteremtése.

A fentiekből következik, hogy a kutatási adatoknak elérhetőeknek kell lennie – más kérdés, hogy mennyire szabadon elérhetőnek. Mi több, az adatok mellé dokumentáció kell. A legújabb kívánalomrendszert a FAIR betűszó: Findable (megtalálható), Accessible (hozzáférhető), Interoperable (szabványos) és Reusable (újrafelhasználható) írja le (Wilkinson et al., 2016).

FAIR – milyen leíró, dokumentáló adatokra van szükség?

A FAIR szempontrendszer megvalósítása megkívánja, hogy a kutatási adatok megtalálhatósága érdekében egyedi, állandó azonosítókat (pl. DOI) alkalmazzanak. A DOI azonosítóhoz tartozó metaadatok fogódzót nyújtanak az adatok felfedezhetőségének növeléséhez, és megteremtik a lehetőségét, hogy azok akkor is elérhetőek maradjanak, ha az idők során más helyre kerülnek. Szükséges az, hogy az adatokat elterjedten használt, szabványos, ingyenesen megvalósítható módon lehessen elérni. Fontos a szabványos formátumok használata, a megfelelő dokumentáció, a jól értelmezhető felhasználási feltételek megléte (licenc).

Az adatok létrehozása és feldolgozása során lényegesen növekednek a dokumentációs feladatok, de a projekt kezdetén is alaposabban át kell gondolni a feladat megvalósítását. Ez utóbbi célt szolgálja az adatkezelési terv elkészítése. Nem csak az adatokat létrehozó és közlő kutatónak, de a projekt befejezése utáni archiválást végző repozitóriumra vagy adatbankra is feladatokat ró a FAIR követelmények megvalósítása.

Lehetőségek

A továbbiakban áttekintjük, megfelelő munkafolyamat szervezéssel és a célt segítő szoftverek alkalmazásával hogyan lehet a mérési jegyzőkönyv vezetését automatizálni. Először is vegyük sorra, milyen alapvető lehetőségek vannak a metaadatok automatikus generálására.

- Automatikus adatkitöltés

Talán a legtöbbet az automatikus metaadat-kitöltés használata járulhat hozzá a dokumentáláshoz. A legtöbb mérőrendszert látható vagy beágyazott számítógépek vezérlik manapság, a szimulációkkal létrehozott adatok eleve számítógépeken keletkeznek. Fontos a szoftvereket úgy megtervezni, hogy minél több beállítási vagy környezeti paramétert naplózzanak automatikusan a szabványos kimeneti adatstruktúrákba. Ma már a hétköznapi életben is elterjedtek az efféle eljárások: a mobil telefonok és digitális fényképezőgépek is rögzíteni tudják a fényképek készítésének idejét és helyét.

- Adat-öröklés

Amit a mérésvezérlő szoftver nem tud a mérőműszerből kiolvasni, azt a mérési kontextusból lehet megpróbálni kinyerni (pl. a mérést végző nevét a számítógépbe bejelentkezett felhasználó nevéből tippelheti meg a rendszer, az intézmény, labor nevét pedig eleve rögzíteni lehet). Ezek a funkciók is évtizedek óta jelen vannak a személyi számítógépeinken: a szövegszerkesztők a dokumentumok leíró adatai között elhelyezik a szerző nevét (amin a felhasználónévből vagy a licenc tulajdonos nevéből következtetnek ki).

- Adatbekérés

Amit a fenti eljárásokkal nem lehet kitölteni, azt az adatot a mérő-szoftver kérdezze meg a mérést végző kutatótól. Célszerű minél kevesebb rubrika kitöltését megkövetelni – a türelmetlen kutató különben ezeket egy-egy „X” beírásával fogja letudni.

Mindezeket az eljárásokat nem csupán a méréskor, de az adatfeldolgozáskor is alkalmazni kell. A feldolgozó szoftvereknek meg kell őrizniük a bemenő adatokkal érkező metaadatokat, sőt, gazdagítaniuk kell azokat az elvégzett adatfeldolgozási lépések dokumentálásával. Végül a kimeneti adatokat is szabványos módon kell elmenteni.

- Metaadat-ellenőrzés, FAIR-ség tesztelés

Végül szoftverek segítségével tesztelni lehet az elkészült adatállományt, mennyiben felel meg a FAIR alapelveknek. Az európai FAIRsFAIR projekt egyik eredménye a FAIR-Aware elnevezésű interaktív on-line értékelő eszköz[1], melynek segítségével a kutató ellenőrizheti, mennyire felelnek meg adatai az elvárásoknak. Az adatok tömeges minősítéséhez természetesen automatikus eszközök használatára is szükség lesz.

- Repozitóriumi szolgáltatások

Vannak olyan adatok, amelyek az adatok archiválását és hozzáférhetővé tételéről gondoskodó repozitóriumban jelennek meg, és az adatlekérdezéskor az eredmény ezeket a metaadatokat a repozitóriumtól örökli (lehet ilyen például a felhasználási licenc). A repozitóriumok szerepet játszanak az adatok és metaadatok ellenőrzésében, a metaadatok rendezésében, szükség esetén az adatállományok formátumának megváltoztatásában is. Ugyancsak jelentős könnyítés a kutatónak, ha az adatok repozitóriumi elhelyezése DOI azonosító generálásával jár.

Példa a csillagászat területéről

- FITS: egy rugalmas és innovatív adatformátum

A FITS egy negyven éves szabvány, ami kiállta az idők próbáját és ma is inspiratív (Wells et al., 1981). Számunkra annyi az érdekes, hogy a metaadatok és az adatok együtt kerülnek tárolásra. A metaadatok a bináris fájl elején egyszerű, szabványos rekordstruktúrában kerülnek elhelyezésre (kulcsszó „=” érték „/” megjegyzés), mennyiségük korlátlanul bővíthető.

- CCD kamera szoftver

Az 1990-es évek elején gyors ütemben terjedő, CCD detektorokat alkalmazó csillagászati (távcsőre szerelhető) kamerák (a kamerát vezérlő szoftver) képesek voltak az adatokat FITS formátumban elmenteni, az adatállományba naplózva a fontos leíró információkat, mint a felvétel készítésének ideje, a megfigyelő, az expozíciós idő, az égi koordináták, az obszervatórium földrajzi koordinátái, stb.

- IRAF

Ezidőtájt szabadon felhasználható szoftverként rendelkezésre állt az IRAF (Image Reduction and Analysis Facility), amit az Egyesült Államokban a National Optical Astronomy Observatory készített és adott közre (Tody, 1986). A CCD képek feldolgozásánál az IRAF beolvasta a FITS állományt, és az abban rögzített leíró adatokat felhasználta a képfeldolgozás során. Végül a feldolgozási lépéseket a kimenő – ugyancsak FITS formátumú – állományba naplózta.

1 https://fairaware.dans.knaw.nl/

- SAADA

A Strasbourgi CDS-ben készítették a SAADA nevű „repozitórium-generáló" szoftvert, ami a Virtuális Obszervatórium kezdeményezés kívánalmainak megfelelően tudott csillagászati adatokat szolgáltatni (Nguyen et al., 2004). Képes volt FITS felvételeket az archívumba beépíteni, a bennük található metaadatok felhasználásával.

*

Mint a példák mutatják, okos programokkal (beleértve a mérőeszközöket vezérlő programokat) sokat lehet csökkenteni a kutatók dokumentációs terheit. A folyamatba integrált dokumentáció elve a számítógép-programozásban is régen megjelent, és onnan az adattudományba is átkerült (Knuth, 1984; Kery et al., 2018). A modern nyílt hozzáférésű tudomány gyakorlatában az interaktív laborjegyzőkönyvek (pl. Jupyter Notebook) is megoldást jelentenek a munkafolyamatba ágyazott dokumentálásra (Kluyver et al., 2016). Mindazonáltal bármilyen jó szoftvereket is használunk, lehetséges, hogy az adatkezelés támogatására külön személyzetet is alkalmazni kell: adatgazdászt (data steward) vagy adatkönyvtárost.

Hogy valósíthatjuk meg a kutatási adatok megfelelő kezelését?

A kutatási adatok és az adatokat tároló repozitóriumok minősítése *a posteriori* történik. Áttörést elérni meggyőződésünk szerint csak a teljes kutatási folyamat *a priori* átalakításával lehet. A nyílt tudomány és a FAIR követelményrendszerét figyelembe kell venni a laboratóriumok felszerelésénél, a kutatási projektek tervezésénél, a publikálásnál és a kutatásértékelésnél is. A kutatóknak mind inkább olyan integrált környezetben kell végezniük munkájukat, amely amellett, hogy segíti a megfelelő dokumentálást, a FAIR követelményeknek való megfelelést, terhet vesz le a kutató válláról. Már most is léteznek elemei az ilyen infrastruktúráknak.

Köszönetnyilvánítás: a dolgozat a Research Data Alliance támogatásával készült (RDA Europe 4.0 grant a HRDA számára).

Irodalomjegyzék

Nguyen, N.H., Michel, L., Motch, C., „SAADA: An Automatic Archival System for Astronomical Data", ADASS XIII., ASPC Conf. Ser. Vol. 314, 121. 2004. http://articles.adsabs.harvard.edu/pdf/2004ASPC..314..121N

Kery, M.B. et al., „The Story in the Notebook: Exploratory Data Science using a Literate Programming Tool", CHI 2018, 2018, doi: 10.1145/3173574.3173748

Kluyver, T. et al., „Jupyter Notebooks – a publishing format for reproducible computational workflows", In *Positioning and Power in Academic Publishing: Players, Agents and Agendas*. IOS Press. pp. 87-90. 2016. doi: 10.3233/978-1-61499-649-1-87

Knuth, D., „Literate Programming", Computer Journal, 27. 2. 97. 1984. https://doi.org/10.1093/comjnl/27.2.97

Tody, D., „The IRAF Data Reduction and Analysis System", in Proc. SPIE Instrumentation in Astronomy VI, ed. D.L. Crawford, 627, https://doi.org/10.1117/12.968154. https://iraf-community.github.io/doc/iraf.pdf

Wells, D. C., Greisen, E. W., and Harten, R. H., „FITS: A Flexible Image Transport System", *Astronomy & Astrophysics Supplement Series*, 44, 363-370, 1981. http://articles.adsabs.harvard.edu/pdf/1981A%26AS...44..363W

Wilkinson, Mark D. et al. "The FAIR Guiding Principles for scientific data management and stewardship", *Scientific Data* vol. 3, 160018. 15. Mar. 2016, doi: 10.1038/sdata.2016.18