Datu zientifikoen argitalpen eraginkorra FAIR printzipioen bidez

2021-ko Urtarrila

1 Datu zientifikoen eskuragarritasuna kolokan

Zientziaren garapen egokia experimentuen erreprodukzioan datza: alegia, edozeinek beste ikerlariek argitaratutako experimentuak erreproduzitzeko gaitasuna edukitzea. Experimentua erreproduzitzeko pausoz-pauso berregin behar da, jatorrizko experimentua baieztatu edo baliogabetzeko. Beraz, experimentu zientifiko bat ezin bada erreproduzitu, ez dauka balio zientifikorik.

Gaur egun komunitate zientifikoko ikerlariek garatzem dituzten experimentu gehienek datuen prozesamendua dute abiapuntu eta helmuga. Horregatik, datu zientifikoak modu egokian eskuragarri egotea berebiziko garrantzia du, prozesu zientifikoak aurrera egin dezan. Tamalez, gaur egun datu zientifikoen bererabilgarritasuna kolokan dago [1], eta beste faktore batzuekin batera, eskuragarritasun egokiaren faltak "erreprodukzio krisia" sortu du [2]: experimentu gehienak ezin dira erreproduzitu, komunitate zientifikoak ontzat emandakoak barne.

Egoera honen aurrean, ikerlari talde batek datu zientifikoak modu bererabilgarrian argitaratzeko printzipioak definitu zituen. FAIR printzipioak deitu zituzten[1]: "**F**indable", "**A**ccesible", "**I**nteroperable", "**R**eusable". FAIR printzipioak lehenengo aldiz *Scientific Data* aldizkari zientifiko garrantzitsuan argitaratu ziren [3].

2 FAIR printzipioak

FAIR printzipioen laburpen ofiziala 1 irudian ikus daiteke, hurrengo azpi-ataletan printzipio bakoitza azaltzen delarik.

Kontuan hartzekoa da FAIR printzipioek ez dutela estandar bat osatzen. Alegia, printzipio **orokorrak** dira, ezin dira bete modu binarioan: ez gara inoiz helduko FAIR printzipioak osoki betetzera, beti egongo da aukera datuak modu FAIR-"ago" batean argitaratzeko. FAIR printzipioen xedea ikerlarientzako datuen argitalpenerako gida izatea da, ikerlarien erakundeek dituzten baliabideen arabera FAIR maila ezberdinak lortuz argitalpen prozesuaren bidez.

[1] `https://www.go-fair.org/fair-principles/`.

Argitu beharreko beste ideia garrantzitsua FAIR printzipioek teknologiarekiko duten erlazioari dagokio. FAIR printzipioak hamarkadetan zehar baliagarriak izateko diseinatuak izan direnez, FAIR printzipioek ez dituzte teknologia zehatzak gomendatzen printzipioak gauzatzeko, nahitaez teknologiak aldakorrak baitira (Hurrengo ataletan teknologia zehatzak aipatzen direnean, implementazio adibide gixa aipatzen dira, ez FAIR printzipioen parte gixa). Hala ere, badago teknologia aldetik FAIR printzipioetan irizpide garrantzitsua: datuen erabilgarritasuna makinentzako erraztea dute xede (eta beraz, ondorio bezala, gizakiontzat ere). Ondorioz, FAIR printzipoek datuen kontsumitzaile nagusitzat makinak dauzkate.

Azkenik, FAIR printzipioak datu eta metadatuei dagozkiela aipatu behar da. Metadatuak ("Datuei buruzko datuak") datuen ezaugarri orokorrak adierazten dituzte, adibidez sortze-data, egilea, datuen gaiak, eta beste hainbeste. Beraz FAIR printzipioetan "(meta)datu" hitza erabiltzen da gehienetan, "datu" eta "metadatu" hitzak banaturik soilik beharrezkoa denean aipatzen direlarik.

2.1 Aurkigarri ("Findable")

(Meta)Datuek aurkituak izateko identifikatzaile globala behar dute, bakarra, eta ez dena gailenduko. Gaur egungo teknologiarekin, URI-ek betetzen dute funtzio hori (Uniform Resource Identifier[2]). Gainera, (meta)datuak argitaratzen dituen erakundeak bere gain hartu behar du sortutako identifikatzaileak mantentzeko erantzunkizuna (Ikus, adibidez, W3C erakundeak egindako promes publikoa[3]).

Bestalde, datuek metadatu aberatsak izan behar dituzte, kontsumitzaileek datuak aurkitzeko duten ahalmena handiagotzeko. Metadatuek datuen identifikatzailea eduki behar dute erreferentzia gixa, kontsumitzailea erraz joan dadin metadatuetatik datuetara, metadatuen bidez datuak aurkitu ondoren.

Azkenik, (meta)datuak indexatuak egon behar dute, bilatzaile orokorretan, edo zerbitzu espezializatuetan.

2.2 Eskuragarri ("Accesible")

(Meta)Datuak eskuragarriak egon behar dute, beraien identifikatzailea erabiliz, komunikazio protokolo estandar, libre eta unibertsal baten bidez. Gaur egun HTTPS da baldintza hori betetzen duen protokolo erabiliena[4].

Gainera, protokoloak autentifikazioa eta erabiltzaileen baimentzea bermatzeko mekanismoak behar ditu, argitaratuko diren datuek erabilpen mugak izan baititzakete (Adibidez datu klinikoak).

Azkenik, nahiz eta datuak desagertu, metadatuak eskuragarri egon behar dute, beharrezkoak ez diren bilaketak ekiditze aldera. Normalean metadatuak gordetzea datuak gordetzea baino merkeagoa eta errazagoa da, batez ere bolumen aldetik.

[2] `https://tools.ietf.org/html/rfc3986.`
[3] `https://www.w3.org/Consortium/Persistence.html.`
[4] `https://tools.ietf.org/html/rfc2818.`

To be Findable:

F1. (meta)data are assigned a globally unique and persistent identifier

F2. data are described with rich metadata (defined by R1 below)

F3. metadata clearly and explicitly include the identifier of the data it describes

F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:

A1. (meta)data are retrievable by their identifier using a standardized communications protocol

A1.1 the protocol is open, free, and universally implementable

A1.2 the protocol allows for an authentication and authorization procedure, where necessary

A2. metadata are accessible, even when the data are no longer available

To be Interoperable:

I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

I2. (meta)data use vocabularies that follow FAIR principles

I3. (meta)data include qualified references to other (meta)data

To be Reusable:

R1. meta(data) are richly described with a plurality of accurate and relevant attributes

R1.1. (meta)data are released with a clear and accessible data usage license

R1.2. (meta)data are associated with detailed provenance

R1.3. (meta)data meet domain-relevant community standards

Irudia 1: FAIR printzipioen adierazpen ofiziala [3].

2.3 Interoperable ("Interoperable")

(Meta)Datuek beste datuekin integratzerakoan lortzen dute baliorik handiena. Horregatik, (meta)datuak kodifikatzeko hizkuntza konputazional formala, eskuragarria, irekia, eta komunitatean erabilera zabala duena erabili behar da: adibidez, RDF (Resource Description Framework[5]) edo OWL (Web Ontology Language[6]).

Gainera, (meta)datuek FAIR printzipioak jarraitzen dituzten hiztegiekin anotatuak egon behar dute. Horrelako hiztegiak, edo ontologiak, Linked Open Vocabularies (LOV) zerbitzuan aurki daitezke[7].

Azkenik, (meta)datuek beste (meta)datuengana jotzen duten erreferentzia kualifikatuak izan behar dituzte, beste datuekin integrazioa errazteko. Alegia, erreferentzia bezala erabiliko den estekaren bi aldeen artean zer erlazio dagoen argi egon behar da (Adibidez, "parte-da", "katalizatzen-du", ...). Erlazioa makinentzako ulerkorra izateko aipatutako hiztegiak erabili behar dira.

2.4 Bererabilgarri ("Reusable")

Bererabilgarriak izateko, (meta)datuak ahalik eta aberatsenak izan behar dira. Halaber, lizentzia argi eta formal bat izan behar dute (Adibidez Creative Commons[8]), eta datuen argitalpenerako aurretik existitzen diren komunitate zientifikoaren estandarrak jarraitu: adibidez, genomika funtzionaleko experimentuen datuak argitaratzeko, MINSEQE (Minimum Information about a high-throughput nucleotide SEQuencing Experiment) estandarra jarraitzea dago[9].

3 FAIR printzipoak eta zientziaren etorkizuna

Sortu zirenetik FAIR printzipoen erabilpena nabarmenki zabaldu egin da: ikerketarako dirulaguntzak jasotzeko, Europar Komisioak (EC) FAIR printzipiotan oinarritutako datu kudeaketarako planak eskatzen ditu. FAIR printzipioak sektore pribatuan ere agertzen dira, hain zuzen ere Bayer[10] edo Novartis[11] bezalako enpresa handiek datuen barne-kudeaketarako erabiltzen dituzte. Badaude ere FAIR printzipioak jarraituz datuak argitaratzen dituzten erakunde publikoak, UniProt datu-basea kasu [4].

Baina FAIR printzipioak betez datuak argitaratzea ez da erraza, azpiegitura tekniko konplexuak eraiki behar baitira [5], eta erakunde gehienek ez dituzte ez baliabide ez eta teknikari egokirik. Arazo honen aurrean badaude FAIR printzipioen arloan produktu eta zerbitzuak eskaintzen dituzten enpresak: The

[5] https://www.w3.org/TR/rdf11-concepts/.
[6] https://www.w3.org/TR/owl2-overview/.
[7] https://lov.linkeddata.es/dataset/lov/.
[8] https://creativecommons.org/.
[9] http://fged.org/projects/minseqe/.
[10] https://www.bayer.com/.
[11] https://www.novartis.com/.

Hyve[12], FAIR Data Systems[13], Eccenca GmbH[14], ...

FAIR printzipioak gero eta garrantzi handiagoa hartuko dute, eta zientzilariek beraien datuak kudeatzeko eta argitaratzeko ahalmen handiagoa beharko dute. Hala ere, ikerlarien ahalegina emankorra izango da, datu berriak kudeatzeko eta aurkitzeko ahalmen handiagoak ikerketa hobeagoak ekarriko baitituzte epe luzera.

Erreferentziak

[1] Dominique G. Roche, Loeske E. B. Kruuk, Robert Lanfear, and Sandra A. Binning. Public data archiving in ecology and evolution: How well are we doing? *PLOS Biology*, 13(11):1–12, 11 2015.

[2] Monya Baker. 1,500 scientists lift the lid on reproducibility. *Nature*, 533(7604):452—454, May 2016.

[3] Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3, 2016.

[4] Leyla García Castro, Jerven Bolleman, Sebastien Gehant, Nicole Redaschi, and María Martin. FAIR adoption, assessment and challenges at UniProt. *Scientific Data*, 6, 12 2019.

[5] Mark D. Wilkinson, Ruben Verborgh, Luiz Olavo Bonino da Silva Santos, Tim Clark, Morris A. Swertz, Fleur D.L. Kelpin, Alasdair J.G. Gray, Erik A. Schultes, Erik M. van Mulligen, Paolo Ciccarese, Arnold Kuzniar, Anand Gavai, Mark Thompson, Rajaram Kaliyaperumal, Jerven T. Bolleman, and Michel Dumontier. Interoperability and FAIRness through a novel combination of Web technologies. *PeerJ Computer Science*, 3:e110, April 2017.

[12] `https://thehyve.nl/.`
[13] `http://www.fairdata.systems/.`
[14] `https://eccenca.com/.`