Anàlisi de la localització de les dades de recerca publicades pels investigadors de la UOC

Location analysis of the research data published by UOC researchers

Maia Francisco Borruel
Universitat Oberta de Catalunya (UOC) - Grup Operatiu de Ciència Oberta (GOCO)

Data de publicació: 2024-03-08
http://hdl.handle.net/10609/149964

Resum

Aquest estudi s'ha dut a terme per avaluar l'emmagatzematge i la publicació dels conjunts de dades dels investigadors *UOC* en repositoris de dades. El propòsit principal de l'estudi és investigar quins repositoris estan utilitzant els investigadors de la *UOC* per arxivar les seves dades i determinar en quina mesura aquests repositoris permeten el compliment de "la normativa".
Els resultats ens permeten saber-ne més sobre les característiques dels repositoris de dades, el coneixement del personal investigador sobre: la *Política institucional de coneixement obert* de la *UOC*, l'impacte dels conjunts de dades i les pràctiques relacionades amb les dades FAIR.

Els informes amb els resultats i les conclusions d'aquest estudi es fan a partir del punt de vista basat en el comportament dels investigadors en relació amb la *Política institucional de coneixement obert,* les dades *FAIR* i els repositoris de confiança (ex. amb certificat *CoreTrustSeal).*

Paraules clau: repositori, dataset, dades FAIR, CoreTrustSeal

Abstract

This study has been carried out to evaluate the storage and publication of *UOC* researchers' datasets in data repositories. The main purpose of the study is to investigate which repositories are being used by *UOC* researchers to archive their data and to determine to what extent these repositories allow compliance with "the regulations".
The results allow us to know more about the characteristics of the data repositories, the knowledge of the research staff about: the *UOC* *Institutional Open Knowledge Policy*, the impact of the datasets and the practices related to the *FAIR* data.

The reports with the results and conclusions of this study are made from the point of view based on the behavior of the researchers in relation to the *Institutional Open Knowledge Policy*, *FAIR* data and trusted repositories (e.g. with *CoreTrustSeal* certificate).

Keywords: repository, dataset, FAIR data, CoreTrustSeal

1. Introducció

Regles, protocols i directrius de gestió de dades

Diverses universitats catalanes, espanyoles i en l'àmbit europeu ja disposen de *Polítiques de dades*[1] [2]. Actualment, no existeix cap marc per a una política de gestió de dades de recerca a la *UOC*. No s'han elaborat regulacions, protocols ni polítiques sobre la gestió de dades. Per tant, no es pot tenir en compte cap normativa de gestió de dades que reguli la gestió, emmagatzematge i subministrament de dades de recerca.
La situació ideal seria que cada *Estudi* pogués treballar en una política pròpia per a la *UOC*, per tal de determinar a través d'un protocol de dades la manera de concretar la normativa dins dels departaments i dels centres de recerca.

No obstant això, la *UOC* sí que disposa d'una *Política institucional de coneixement obert* per a les responsabilitats del personal docent i investigador respecte a les dades de recerca.

- **Resum de responsabilitats del personal docent i investigador respecte a les dades de recerca** (*Política institucional de coneixement obert*):

[1] Polítiques de dades de diverses universitats
[2] https://cora.csuc.cat/ciencia-oberta/ciencia-oberta-a-les-universitats-i-centres-de-recerca-de-catalunya/

- “6. Publicar en obert les dades de recerca derivades de projectes de recerca”. (p.12)
- “7. Garantir que les dades obtingudes en l'activitat de recerca segueixin els principis *FAIR* per tal que siguin localitzables, accessibles, interoperables i reutilitzables”. (p.12)
- “9. Gestionar les dades d'acord amb les millors pràctiques, els codis ètics, la normativa i la legislació aplicables”. (p.12)
- “10. Conservar les dades i emmagatzemar-les de manera clara i precisa per permetre l'avaluació de resultats, la recuperació dels procediments i la reproducció de la recerca”. (p.12)
- “Sempre s'han d'assegurar els principis *FAIR* i s'ha d'actuar d'acord amb la legislació i el codi ètic vigent, i els requisits de les institucions finançadores”. (p.10)

La *Política institucional de coneixement obert* indica que sempre s'han d'assegurar els principis ***FAIR*** per tal que les dades siguin: localitzables, accessibles, interoperables i reutilitzables.

- **Pràctiques relacionades amb les dades *FAIR***
  (Findable, **A**ccessible, **I**nteroperable, **R**eusable)
  (https://www.go-fair.org/fair-principles/)

  Cercables (Findable**)**
  - Identificador persistent (ex. DOI)
  - Metadades enriquides
  - Cercable i descobrible en línia

  Accessibles (Accessible**)**
  - Dipositat en un repositori de confiança
  - Les dades es poden restringir i seguir sent *FAIR* - “tan obertes com sigui possible, tan tancades com sigui necessari”

  Interoperables (Interoperable**)**
  - Formats de fitxer en obert i/o estandarditzats

  Reutilitzables (Reusable**)**
  - Ben documentat (ex. fitxers README), incloent-hi la procedència i les eines / instruments necessaris per reproduir els resultats
  - Llicència clara (ex. CC BY 4.0, CC0)

Per assegurar que les dades es poden reutilitzar de manera responsable i productiva, la millor opció és emmagatzemar aquestes dades en un repositori de dades de confiança preferiblement amb certificat ***CoreTrustSeal***. Aquesta certificació està relacionada amb la infraestructura organitzativa, la gestió d'objectes digitals, la tecnologia de la informació i la seguretat. Això fa que el repositori sigui fiable per arxivar els conjunts de dades.

- **Repositoris de confiança** (https://www.coretrustseal.org/) (https://amt.coretrustseal.org/certificates/)

    - Repositoris certificats (ex. *CoreTrustSeal*, nesto Seal DIN31644, ISO16363)
    - Repositoris disciplinaris i dominis comunament utilitzats i avalats per les comunitats de recerca internacionals
    - Repositoris generalistes (ex. Zenodo) o institucionals que presenten les característiques essencials dels dipòsits de confiança:
        - serveis, mecanismes i disposicions establertes per garantir l'exactitud, integritat, autenticitat i accés dels continguts.
        - ús de PID
        - metadades detallades, normalitzades i accionables per màquina (incloent-hi la procedència i la llicència)

Identificació de conjunts de dades

Per tal de reunir els conjunts de dades publicats per investigadors de la *UOC*, hem sol·licitat una llista dels investigadors principals (*IP*) per tal d'investigar per autors i veure un a un en quins repositoris dipositen. L'eina ideal per a dur a terme aquesta tasca hauria sigut *Data Monitor* d'Elsevier, però malauradament la *UOC* no hi té accés. Aquesta eina ens permetria fer un seguiment dels conjunts de dades de la nostra institució en diferents repositoris de dades.

2. Materials i mètodes

Mètodes per fer la recerca:

1. Entendre per "dataset" només conjunts de dades sense tenir en compte "figures" ni "col·leccions".
2. Només comptarem com a datasets dipositats els datasets que tinguin DOI.
3. Com a punt de partida disposem de la llista dels 50 investigadors principals (*IP*) dels grups de recerca *UOC* amb data d'abril de 2023 proporcionada per *ARI*.
4. Cerca dels datasets dels investigadors *UOC* en diferents repositoris.
    a. Detectar quins repositoris fan servir els investigadors *UOC*.
5. Tenim casos en què un dataset està associat a diversos investigadors *UOC*. En aquest cas, tindrem DOI de datasets duplicats però amb diferents "*UOC* researcher" associats. Cal detectar si els investigadors pertanyen a *Estudis* o centres diferents.
6. El repositori *Figshare* actualment funciona com a cercador i a la vegada com a repositori.
   Per detectar si el dataset està dipositat a *Figshare,* la URL del DOI assignat ha d'incloure el terme "figshare". Exemple:

https://doi.org/10.6084/m9.**figshare**.c.3846880.v1
També tenim casos en què el DOI inclou el terme “fpsyg” . Exemple:
https://doi.org/10.3389/**fpsyg**.2022.1040651.s003
A *Figshare*[3][4][5], els datasets associats a *PLOS ONE* tenen assignat un DOI que inclou *l’string* “journal.pone”. Exemple:
https://doi.org/10.1371/**journal.pone.**0277899.t001

7. Tenim casos en què el fitxer és part d’un conjunt de datasets associats al mateix article. Ho podem detectar a la informació del DOI.
   Exemple:
   https://doi.org/10.34810/data113/110 és part de
   https://doi.org/10.34810/**data113**.
   En aquest cas només tenim en compte el dataset amb el DOI arrel. (i.e,
   https://doi.org/10.34810/**data113**).
8. La URL del DOI dels datasets dipositats a CORA[6] inclou l’string ”data”. Exemple:
   https://doi.org/10.34810/**data186**
9. Tenim casos en què diversos datasets amb DOI assignat estan associats a un mateix article. A la informació del DOI veiem que el DOI arrel condueix a l’article i el DOI dels datasets associats afegeix un apèndix al DOI. Exemple: DOI arrel
   https://doi.org/10.3389/fpsyg.2022.**1040651** DOI dels datasets:
   https://doi.org/10.3389/fpsyg.2022.**1040651.s001** ,
   https://doi.org/10.3389/fpsyg.2022.**1040651.s002**
   En aquest cas tenim en compte tots els datasets.
10. La quantitat de visualitzacions i descàrregues dels datasets va augmentant amb el temps.
11. Apareixen casos en què l'autor no forma part de la llista dels 50 investigadors principals, però sí que ha dipositat datasets a CORA. En aquest cas es prioritza que l'investigador hagi dipositat a CORA, ja que considerem que aquesta informació és rellevant per a l'informe i, per tant, el dataset s'inclou en la llista.

[3] El 2013, Figshare va anunciar una associació amb PLOS per integrar l'allotjament, l'accés i la visualització de dades de Figshare amb els seus articles PLOS associats.
[4] A Figshare cal distingir entre els datasets dipositats al mateix repositori i els que apareixen a la cerca, però estan dipositats a altres repositoris.
[5] Glosari de Figshare: https://help.figshare.com/article/figtionary
[6] Crear i dipositar un dataset a CORA:
https://confluence.csuc.cat/display/RDM/Crear+i+dipositar+un+dataset

3. Anàlisi i resultats

Informes

Després de reunir tota la informació sobre els conjunts de dades publicats pels investigadors principals (*IP*), s'ha elaborat un informe per tal de facilitar una visió general específica de la disciplina. L'informe ens mostra tota la informació analitzada a escala universitària. L'informe permet activar filtres per visualitzar la informació de cada *Estudi* de la *UOC*, on els resultats poden ser més interessants per les seves particularitats i marc disciplinari.

L'informe es pot consultar a través d'aquest enllaç:

- https://app.powerbi.com/Redirect?action=OpenApp&appId=57a85203-b150-4864-a754-d9e20ae3c4d3&ctid=aec762e4-3d54-495e-a8fe-4287dce6fe69

La millor manera d'observar els resultats és interactuant amb l'informe PBI i les figures d'aquest informe. Algunes pàgines de l'informe PBI inclouen opcions de filtre per poder respondre necessitats concretes.

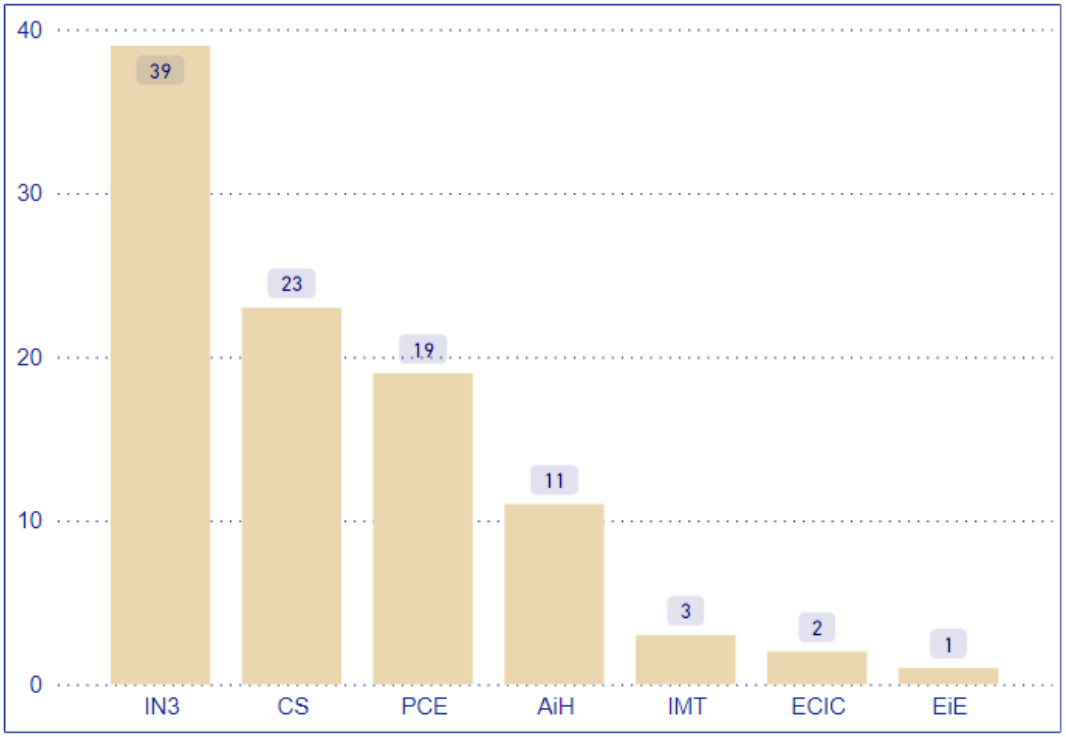
A l'informe PBI podem trobar la següent informació:

1. Repositoris utilitzats pels investigadors de la *UOC*.
    - Llista de repositoris utilitzats pels investigadors i conjunts de dades relacionats amb cada *Estudi*.
    - Certificació dels repositoris.
    - Compliment dels investigadors amb la *Política institucional de coneixement obert*: arxivament de dades en repositoris certificats.
    - Tipus de repositoris utilitzats pels investigadors.
2. Recompte de conjunts de dades dipositats a la *UOC* per any, per *Estudi* i per repositori.
3. Mètriques dels conjunts de dades.
4. Conjunts de dades relacionats amb articles.
5. Autoria dels conjunts de dades.

4.Resultats

Identifiquem que els investigadors de la *UOC* han publicat **90 conjunts de dades** (registrats fins al novembre de 2023). Al següent gràfic podem veure que la majoria dels conjunts de dades han estat publicats per **investigadors d'*IN3*** (39/90 datasets, 43,33% del

conjunt de dades). És un resultat que no ens sorprèn, ja que el centre de recerca *IN3* és l'amfitrió del nombre més gran de projectes de recerca intensius en dades.



Publicació de datasets per Estudi o Centre de recerca.

Al llarg d'aquest informe no es mencionen els *Estudis Dret i Ciències Polítiques (EDCP)* donat que no s'han detectat dipòsits de datasets per part d'aquests estudis.

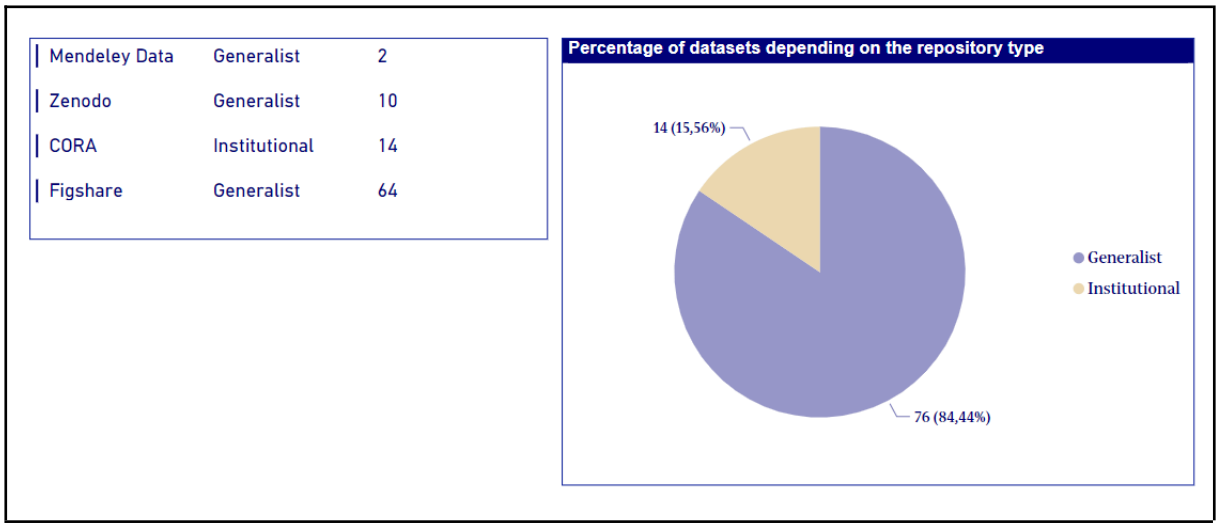
Per poder entendre els resultats de l'estudi és necessari conèixer una mica més els repositoris de dades. En el nostre informe distingim dos tipus de repositoris:

- **Institucional**: recull conjunts de dades d'autors d'una institució concreta. En el nostre cas, *CORA* serveix com a repositori institucional per als investigadors *UOC*. Com a institució, ja s'ha sol·licitat la certificació del repositori, donat que la preservació a llarg termini de les dades ja està garantida.

- **Generalista/Multidisciplinari**: quan els investigadors no tenen un repositori institucional o no troben un repositori disciplinari útil per a les seves dades, poden publicar el conjunt de dades en un repositori generalista, ja que accepta dades independentment del contingut, format, tipus de dades o enfocament disciplinari. Només els repositoris generalistes que donen servei a una comunitat

específica designada poden tenir una certificació (ex. Mendeley Data, Zenodo[7], Figshare).

El següent gràfic mostra el recompte de datasets dipositats per tipus de repositori.

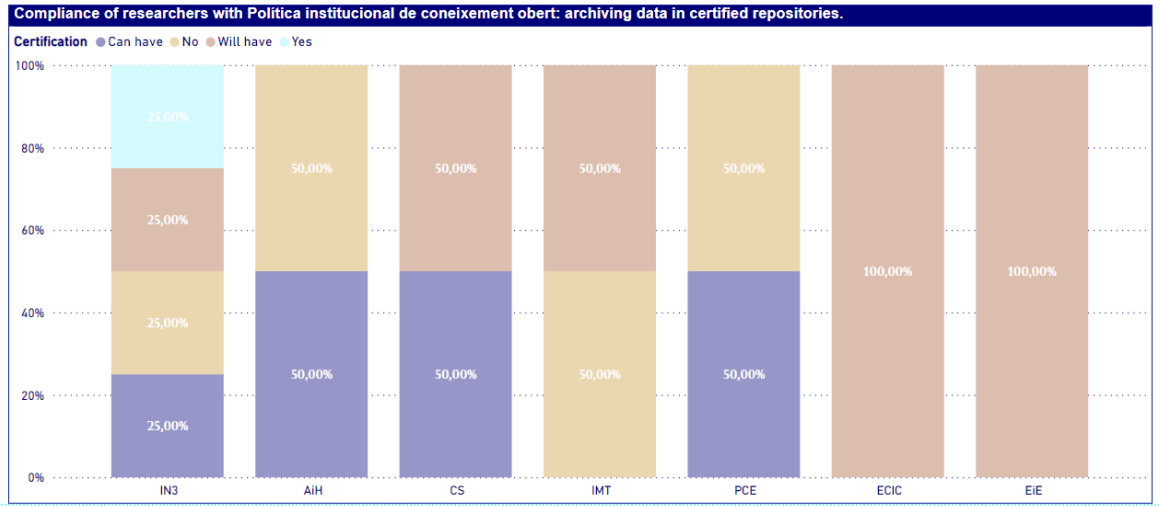


Publicació de datasets per repositori.

L'elecció de l'investigador a l'hora de cercar un repositori de dades pot afectar el compliment dels requisits de la universitat (i.e, *Política institucional de coneixement obert*), ja que el compliment està relacionat amb l'ús de repositoris de confiança (ex. amb certificat *CoreTrustSeal*).

El següent gràfic ens proporciona una idea del compliment esmentat mostrant l'estat de la certificació *CoreTrustSeal* dels repositoris utilitzats pels investigadors dels diferents *Estudis*.

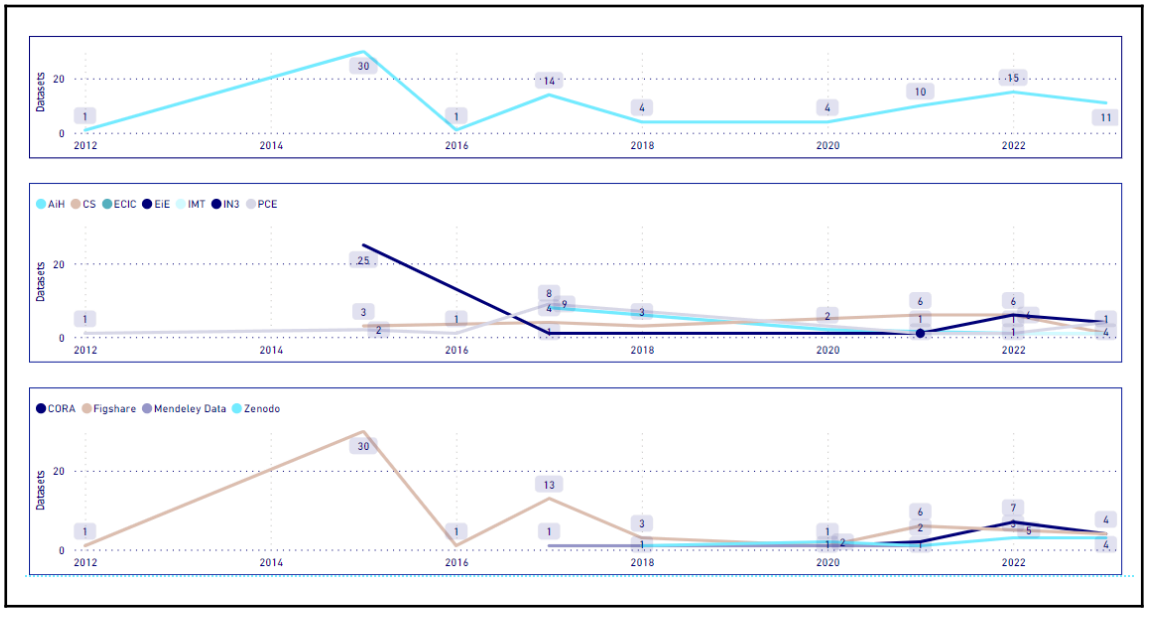
[7] Descripció de l'estat de la certificació del repositori Zenodo:
"Does *Zenodo* have the CoreTrustSeal or other certification?
Not yet since *CoreTrustSeal* at the moment only certifies repositories that serve a specifically designated community and therefore not generalist repositories like *Zenodo*. *CoreTrustSeal* is considering including generalist repositories in *CoreTrustSeal 2022*, at which point we would apply for certification. We have provided our input to *CoreTrustSeal*'s request for community feedback on this issue". (https://help.zenodo.org/faq/)



Publicació de datasets per repositori segons l'estat de la certificació.

L'*IN3* és el centre de recerca on els investigadors compleixen més amb els requisits donat que el 25% dels datasets estan dipositats en repositoris amb certificació *CoreTrustSeal*.

El següent gràfic mostra el recompte de conjunts de dades dipositats per any, per departament i per repositori.



Publicació de datasets per any, per departament i per repositori.

IN3 és el centre de recerca que més ha dipositat i *figshare* és el repositori més utilitzat per a dipositar.

5. Conclusions

Un cop finalitzat aquest estudi, identifiquem algunes qüestions relacionades amb l'equitat de les dades de recerca (i.e, dades *FAIR*), el seu impacte i la importància dels protocols.

En general, hem vist que el compliment dels investigadors amb els "requisits universitaris" no és bo, tot i tenir en compte que el document de la *Política institucional de coneixement obert* no especifica els requisits que cal complir. Creiem que és realment important que els investigadors puguin tenir alguna mena de "protocols" com a guia en tot el procés de la seva investigació.

Veiem que hi ha algunes pràctiques que podrien millorar-se:

- Els investigadors emmagatzemen sovint les seves dades perquè "ho han de fer", però no les comparteixen o citen. En aquests casos, els conjunts de dades són difícils de trobar i també es crea una barrera a la reutilització.

- És important vincular els conjunts de dades a les publicacions de recerca perquè tots els interessats en la recerca puguin trobar fàcilment el conjunt de dades.

- Tots els articles haurien d'incloure una "Declaració de Disponibilitat de Dades"[8] (*Data Availability Statement*), fins i tot quan no hi hagi dades associades a l'article.

Mitjançant aquest estudi també en volíem saber més sobre l'impacte de la publicació de conjunts de dades. Hem vist que això no és tan fàcil de demostrar. Només alguns dels repositoris ens mostren algunes mètriques relacionades amb els conjunts de dades (generalment els repositoris generalistes). Podem veure quantes vistes o descàrregues tenen els conjunts de dades, però no és fàcil relacionar aquesta informació amb l'impacte de la recerca o dels investigadors. Per a poder saber més sobre l'impacte, hauríem de tenir informació sobre publicacions sense conjunts de dades publicades per tal de poder comparar les mètriques. Aquest aspecte es podria estudiar en futures recerques, ja que seria molt interessant disposar d'aquesta informació. Però, de moment, hi ha moltes coses que es poden millorar amb bones pràctiques per a fer que les dades de recerca siguin més accessibles, reutilitzables i augmentar la seva visibilitat.

Observem que el nombre de visites i de descàrregues dels datasets augmenten a mesura que passen els dies. Aquest augment progressiu de visites i descàrregues indica que els datasets són utilitzats per altres investigadors i que, per tant, podem confirmar que la funció de dipositar en repositoris és imprescindible per a contribuir a la *Ciència Oberta* com a recurs que beneficia a tota la comunitat científica.

[8] 00052_Infografia-consells-escriure-treballs-recerca.pdf Veure apartat núm. 9.

6. Dades de recerca relacionades

(*Data availability statement*)

The datasets generated and analysed during the current study are available on the CORA repository at https://doi.org/10.34810/data1162 .