Anàlisi de l'accessibilitat a les dades de recerca del Campus Nord UPC a partir de la informació recollida a articles de revista

Biblioteca Rector Gabriel Ferraté

Gener 2024

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Servei de Biblioteques, Publicacions i Arxius

BRGF

0. Introducció

L'objectiu del present treball és obtenir informació sobre l'accessibilitat de les dades de recerca generades o usades pel personal investigador del Campus Nord de la UPC a partir de la informació al respecte indicada als seus articles.

Concretament, l'anàlisi vol donar resposta a les següents qüestions:

- En quina mesura en els articles de revista s'informa de la disponibilitat de dades de recerca
- En quina mesura els investigadors estan publicant dades de recerca
- On es publiquen les dades de recerca i els codis font de programes informàtics
- Quins són els motius pels quals no es publiquen dades

1. Metodologia

1.1 Selecció dels articles:

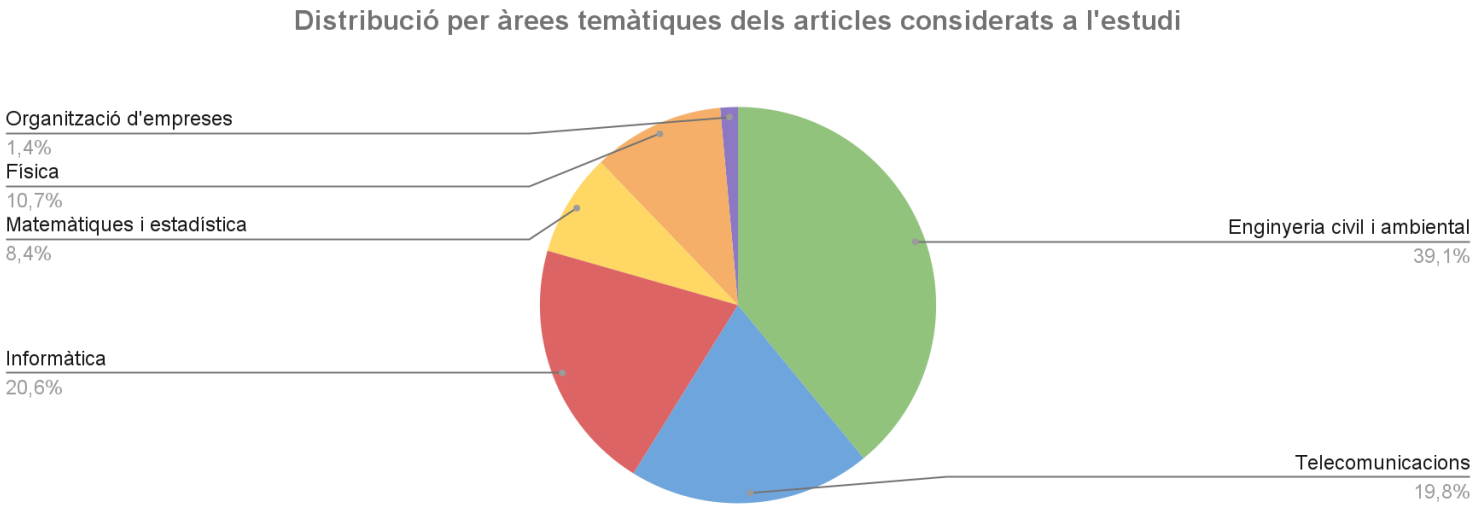
- S'han analitzat 705 articles de revista publicats el 2022, signats per, com a mínim, un autor vinculat a un centre docent del Campus Nord de la UPC[1].
- La font d'informació és FUTUR[2]. L'extracció de les dades es va realitzar el 4 d'octubre de 2023.
- Criteri d'exclusió: s'han exclòs de l'anàlisi els articles sense DOI.

[1] Escola Tècnica Superior d'Enginyeria de Camins, Canal i Ports de Barcelona (ETSECCPB); Escola Tècnica Superior d'Enginyeria de Telecomunicacions de Barcelona (ETSETB) i Facultat d'Informàtica de Barcelona (FIB)

[2] FUTUR. Portal de la Producció Científica de les Investigadores i Investigadors de la UPC.

1.2 Característiques de la mostra d'articles analitzats

- La distribució segons àrees temàtiques reflecteix el pes de les àrees d'especialització del Campus Nord:



1.3 Selecció i anàlisi de la informació

- Per a cada article s'ha extret la següent informació:
  - Existència o no d'informació específica sobre disponibilitat de dades de recerca o codi informàtic.
  - Existència o no de dades annexades a l'article, ja sigui com a annex, apèndix, com a material suplementari o incrustat.
  - Nom dels repositoris, servidors, portals, etc. que allotgen les dades creades o utilitzades.
  - Les URLs dels datasets o codis accessibles.

- La informació analitzada s'ha extret de les següents seccions dels articles:
  - Seccions específiques sobre disponibilitat de les dades (amb diferent nom en funció de l'editorial: *Data availability*; *Code availability*; *Data availability statement*; *Availability of data and materials*, etc.).
  - Apèndixs, annexos i seccions amb materials addicionals que poden incloure dades o bé enllaços a repositoris: (*Supplementary material*;

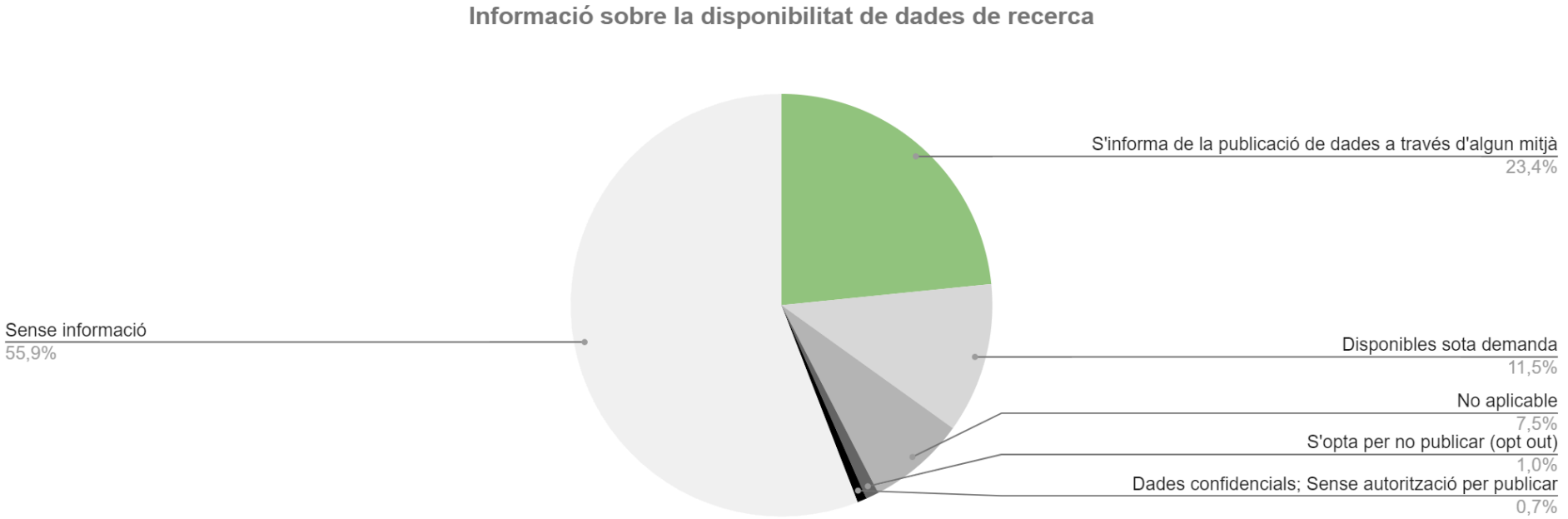
Supplementary files, *Supplementary data*, etc.).

- El cos de l'article: de vegades s'esmenta dins del text de datasets o codi en forma de citació o amb una nota al peu. Algunes editorials com PloS inclouen dades enllaçades, que apareixen incrustades a les gràfiques o taules del cos de l'article.
- Conclusions: sovint la disponibilitat de les dades s'esmenta entre les consideracions finals.
- Referències: En aquest cas només s'han tingut en compte les dades i els codis font usats o creats en la recerca descrita en l'article. No s'han considerat altres datasets o codis esmentats com a part dels antecedents (*background*), estat de la tècnica, o bé com a eines, programari o aplicacions desenvolupats per tercers que han estat emprats de forma instrumental.

- A efectes d'anàlisi, al present treball es consideren dades:
  - Les dades de recerca, segons la definició de CODATA[3]. S'inclouen les dades reutilitzades provinents de servidors i repositoris.
  - Els codis font de programes informàtics.

[3]Les dades de recerca són els registres numèrics, textuals, imatges o sons que s'utilitzen com a fonts primàries per a la investigació científica i que són comunament acceptats per validar les conclusions i els resultats d'una investigació per la comunitat científica. Les dades de recerca poden ser dades experimentals, dades d'observació, dades operatives, dades de tercers, dades del sector públic, dades de seguiment, dades processades o dades reutilitzades.

2. Resultats



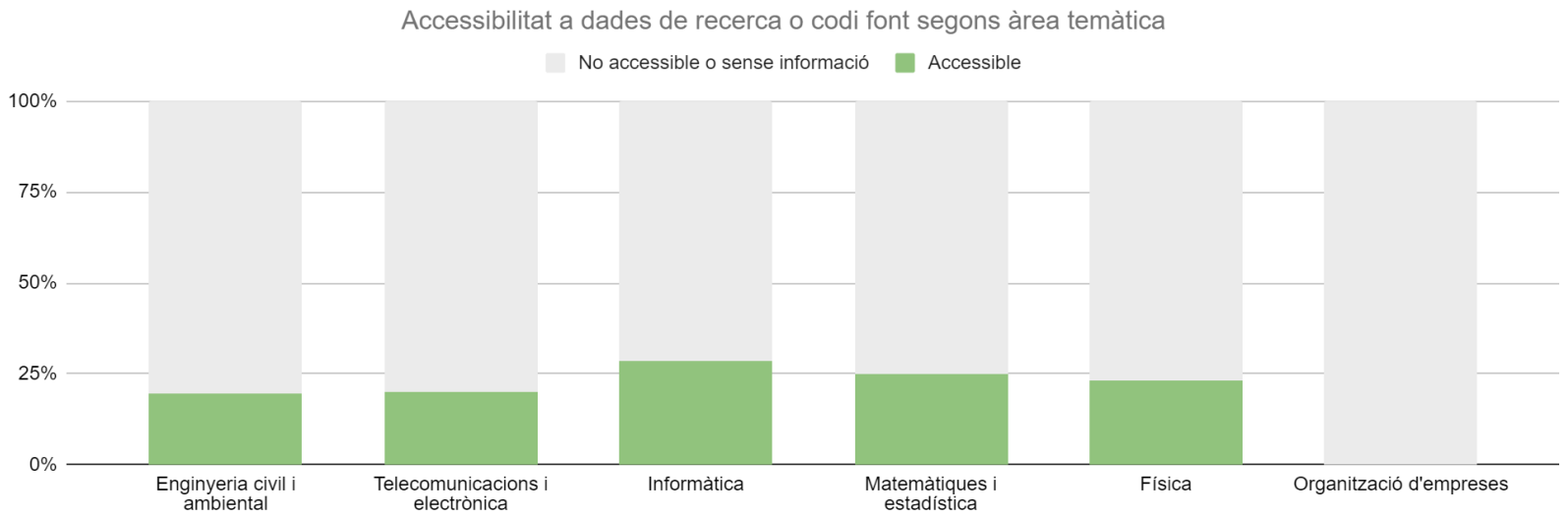
2.1 Informació sobre la disponibilitat de dades de recerca

- Un 44% dels articles estudiats inclou informació sobre la disponibilitat de les dades de recerca, ja sigui per informar de la seva ubicació i condicions d'accés, per indicar que les dades no són accessibles, o bé per indicar que no s'han generat datasets en el curs de la recerca.
- El 56% dels articles no fan esment a la disponibilitat de dades de recerca o codis font.

2.2 Publicació de dades de recerca o codi font

- Al 23% dels articles s'informa que les dades de recerca o els codis font relacionats són accessibles, i es proporcionen enllaços o indicacions per poder accedir-hi.
- Al 77% dels articles restants no s'indica cap informació sobre publicació de dades o codi font, o bé s'informa de la seva no accessibilitat.

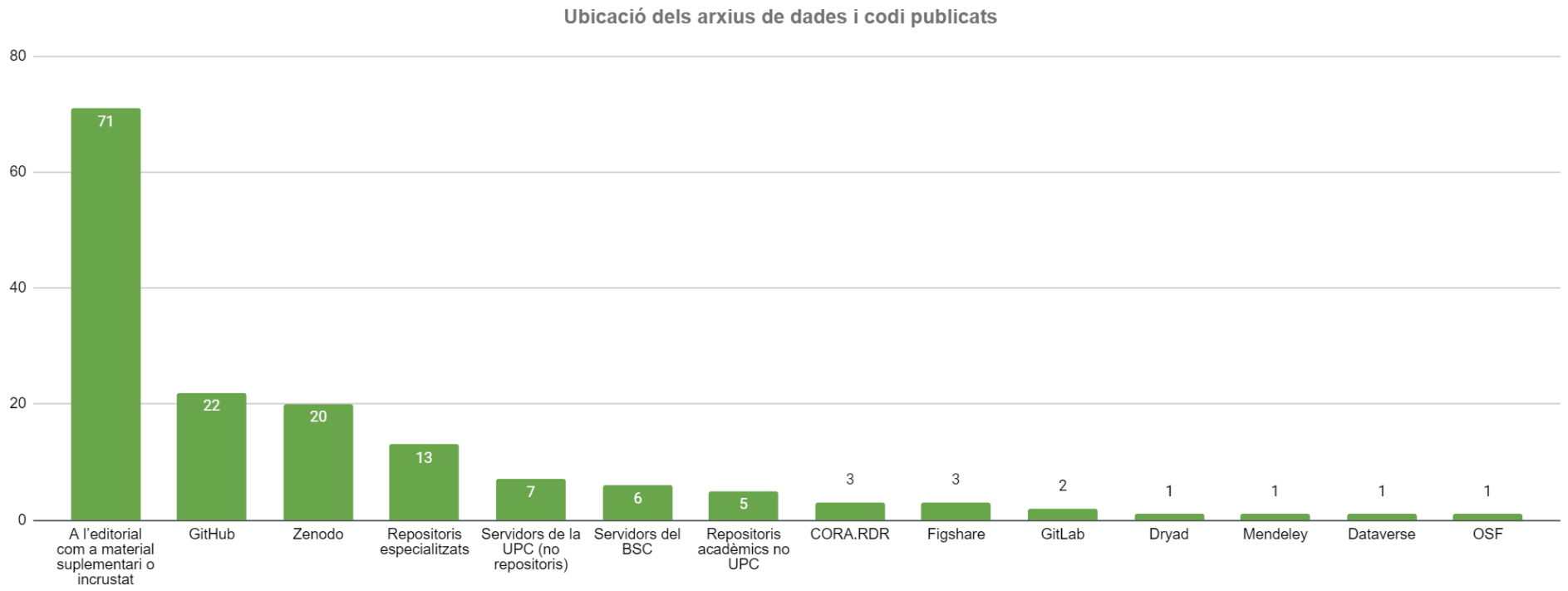
- Per àrees temàtiques el percentatge és lleugerament superior als articles de l'àrea d'informàtica.



2.3 Motius de la no publicació de dades o codi

- A l'11,5 % dels articles s'indica que els autors facilitaran les dades sota demanda (*on request; on reasonable request*).
- Al 7,5% dels articles s'afirma que la recerca exposada no ha generat o usat dades (*not applicable*).
- A l'1% dels articles els autors opten per no fer les dades públiques (de vegades s'utiltza l'expressió *opt out*). En alguns casos s'especifica que les dades s'estan utilitzant en un projecte en curs (*on-going project*).
- Al 0,7% dels articles s'informa de la confidencialitat de les dades, o bé de la impossibilitat de fer-les públiques per motius contractuals o de copyright.

2.4 Ubicació de dades de recerca i codis font



- A continuació es descriuen les diferents ubicacions on s'han publicat dades o codis relacionats amb els articles considerats en aquest estudi:
  - Al mateix article:
    - Les revistes de les editorials Elsevier, MDPI, Springer o PLoS publiquen sovint dades com a material suplementari o incrustat (*embedded*) de l'article.
    - És l'opció més freqüent. Un 45,5% dels articles analitzats fan públiques les seves dades a través d'aquesta opció.
    - En alguns casos, les dades incloses dins l'article es troben dipositades a repositoris vinculats a l'editorial, com és el cas de Code Ocean d'IEEE.
    - El grau d'accessibilitat en obert a les dades correspon al de l'article. En el cas de revistes híbrides les dades són accessibles en obert només si s'ha tramitat la corresponent APC[4]. En cas contrari únicament són accessibles per als lectors/institucions que disposen de subscripció.
    - Alguns articles considerats en aquest estudi s'han publicat a revistes de dades (*data journals*). En aquestes publicacions les dades es publiquen totalment o parcial dins de l'article.
  - Plataformes de desenvolupament de codi: GitHub; GitLab; PiPy
    - GitHub és la plataforma més utilitzada.
    - En alguns casos els codis s'allotgen a comptes genèrics corresponents a projectes o grups de recerca. Tot i això, en la majoria d'ocasions, els arxius es publiquen a comptes personals d'investigadors (de la UPC o d'altres institucions).
  - Repositoris de dades: Zenodo; Figshare; Mendeley-Data; Dryad; OSF; Dataverse
    - Zenodo és el repositori de dades més usat.
    - Tot i l'existència de comunitats que recullen les dades d'un grup de recerca, o les creades en el marc de projectes de recerca, és més freqüent que el dipòsit es faci aïlladament per algun dels autors dels datasets.
  - Repositoris especialitzats:
    - S'usen sobretot en articles d'àmbits com: meteorologia, teledetecció, enginyeria del terreny, oceanografia; i especialment en relació amb grans projectes de recerca d'abast global. GNSS Service de la NASA, és el més citat.
    - També es citen com a fonts de dades estadístiques diversos portals d'administracions i organismes internacionals.
  - Servidors o webs de la UPC (no repositoris):
    - Com ara, servidors de departaments, webs de grups de recerca, o portals que recullen resultats d'un projecte.
  - Servidors del Barcelona Supercomputer Center (BSC):
    - En el cas de projectes desenvolupats amb participació del BSC, o per investigadors UPC amb doble afiliació

[4] *Article processing charge*

    - Dipòsits acadèmics no UPC:
        - Habitualment quan un dels autors és membre de la institució que gestiona el dipòsit on s'han publicat les dades. És el cas, per exemple de Digital-CSIC.
    - Repositori UPC CORA.RDR:
        - 3 articles esmenten datasets publicats al repositori de dades de la UPC.
- Els articles que proporcionen informació sobre la disponibilitat de les dades, indiquen sovint més d'un repositori, plataforma o ubicació dels diferents datasets o arxius de codi creats o usats.

3. Conclusions i propostes

- A la majoria dels articles estudiats (55%) no es proporciona informació sobre la disponibilitat dels fitxers de dades de recerca o codi font generades en l'activitat de recerca. Seria convenient explorar més a fons les causes considerant diferents hipòtesis: diferències per disciplines, tractament i condicions de les editorials durant en procés de tramesa del manuscrit, hàbits dels investigadors, dinàmiques dels grups de recerca, etc.
- La important dispersió pel que fa al lloc de publicació de dades de recerca, no afavoreix la seva visibilitat, especialment en el cas de les dades publicades a les mateixes editorials com a materials suplementaris dels articles. És important tenir en compte que en molts casos aquests materials només són accessibles des de la versió HTML a la web de l'editorial, i només si l'article també és accessible en accés obert. Seria important recomanar als investigadors el dipòsit addicional d'aquest materials, d'acord amb els principis FAIR, a d'altres repositoris, preferentment a CORA.RDR.
- Igualment, caldria conèixer amb més detall si la signatura dels Copyright Transfer Agreements amb les editorials, implica també la cessió de drets les dades publicades com a material addicional o incrustat.
- Les dades o el codis allotjats a alguns repositoris com ara Zenodo i GitHub, sovint no mostren metadades correctes dels noms i les filiacions dels membres de la UPC, ni tampoc sempre queda clara la seva relació amb projectes de recerca competitius.
- Per evitar la dispersió i afavorir la visibilitat de les dades de recerca és convenient avançar en la introducció a DRAC de la informació relativa als datasets creats per membres de la UPC, d'acord amb els objectius del Full de Ruta 2025 de Ciència Oberta UPC[5].

[5] Universitat Politècnica de Catalunya. Acord CG/2023/07/08, de 5 de juliol de 2023, del Consell de Govern. per la qual s'aprova el Full de ruta 2025 de Ciència Oberta UPC. Disponible a: https://govern.upc.edu/ca/consell-de-govern/consell-de-govern/sessio-07-2023-del-consell-de-govern/comissio-de-recerca/aprovacio-del-full-de-ruta-2025-de-ciencia-oberta-upc