Reproductive Sciences (2021) 28:1023–1025
https://doi.org/10.1007/s43032-021-00484-8

REPRODUCTIVE BIOLOGY: PERSPECTIVE, OPINIONS AND COMMENTARIES

Is Sharing Datasets the Answer to the New Challenges of Reproductive Biology Research?

A. Sixto-Costoya[1,2] · R. Lucas-Domínguez[1,2,3] · R. Aleixandre-Benavent[2,4] · A. Vidal-Infer[1,2]

Received: 15 September 2020 / Accepted: 31 January 2021 / Published online: 16 February 2021
© Society for Reproductive Investigation 2021

Abstract

Data sharing increases the speed of research and saves time and resources while ensuring transparency and reproducibility. We have analyzed this behavior through the reproductive biology community. Our study revealed that Q1 (44%) and Q2 (36%) JCR reproductive biology journals are the most active journals in data sharing.

Keywords Data sharing · Datasets · Reproductive biology · Supplementary material · Raw data

Both scientific and biomedical advances and life style changes in the population (postponement in the age of the first child, different family models, importance of the professional career for women, etc.) make reproductive biology an area of special interest not only to the scientific community but also all the society [1]. There is an interest in reproductive biology as a constantly evolving field because of its relevance in a changing society, making effective but quick research necessary.

The sharing of datasets among researchers enhances this combination of speed and effectiveness. Among other possibilities, data sharing practices save time, money, and effort. Moreover, this practice allows researchers to validate studies' results, thanks to the observation of the raw data shared [2]. Apart from being a useful and beneficious practice, international statements declarations sustain and highlight that the research data, or raw data, is an integral part of scholarly knowledge [3]. Some projects in reproductive biology, like the Minerva Initiative, have promoted the sharing of research data about computational phenotyping [4]. The GO-FAANG consortium was also conformed to advance the annotation of assembled genomes of different organisms [5].

A very recent example has been the rapid release to the public of the genome sequence of the new coronavirus SARS-CoV-2 responsible for COVID-19 [6, 7]. Related to this last crisis, which is showing to have impact also in the reproductive biology field, there is already publications that considered data sharing as basic for public health action, including all the types of data related to health research from clinical trial to observational studies, operational research, genetic sequences, monitoring of disease control programs, survey results, etc. [8, 9].

Until now, other areas linked to reproductive biology, such as cell tissue engineering, are carrying out data sharing [10, 11], but it is also used in different disciplines, such as substance abuse research [12] that shows the spreading of this practice. It was found that the type of raw data that was shared varied by discipline. For example, in cell tissue engineering, it is especially common to share spreadsheets, but images and videos are also frequent. The way to share these data may change, but in the original articles, the possibilities to share data are usually related to the journal policies in which the papers will be published.

Despite the evidence of studies related to sharing comprehensible raw data and its advantages in several disciplines [13], no precise study about the journals in the specific area of reproductive biology was found. For this reason, we analyzed the presence of research data sharing in reproductive biology journals indexed in Journal Citation Reports (JCR) and PubMed Central repository.

✉ A. Sixto-Costoya
Antonio.Vidal-Infer@uv.es

1 Department of History of Science and Documentation, School of Medicine, University of Valencia, Avda. Blasco Ibañez 15, 46010 Valencia, Spain

2 UISYS, Joint Research Unit, CSIC-University of Valencia, Pza. Cisneros 4, 46003 Valencia, Spain

3 CIBERON, Valencia, Spain

4 Ingenio, CSIC-Politechnic University of Valencia, Ciudad Politécnica de la Innovación, Edif 8E 4º, Camino de Vera s/n, 46022 Valencia, Spain

Following the methods from previous studies [11, 12], we selected the 29 journals of the reproductive biology category of Web of Science (WoS), and they were organized in quartiles according to the Journal Citation Reports (JCR) ranking. After that, we executed a search equation in PubMed/Medline (PM) to retrieve all the articles published in the 29 journals from December 2018 to February 2019. Then a secondary search was carried out in the PubMed Central repository (PMC), based on the methodology of previous studies [11, 12, 14], to complete an analysis of the articles' supplementary material. The research strategy used in PMC was designed to retrieve only articles with supplementary material: "journal name"[Journal] (<supplementary-material> or <supplemental-information>). The number and types of files located on the articles with supplementary material were registered even if a single article included several different files. In cases where there was a compressed file, such as a .zip or .rar file, it was opened to check what types of files it contained.

The main results confirmed that 24 out of the 29 journals contained in the JCR Reproductive Biology category are indexed in PM and PMC. A total of 109,202 articles were found in PM and 10,928 in PMC repository (10.01%). From this 10.01%, 1841 documents contained supplementary material, which represents 16.85% of the articles in PMC, distributed only in quartile 1 (Q1) (44%), Q2 (36%), and Q3 (20%) journals.

The specific analysis of supplementary material showed that the most frequent files were pdf (n=1573; 38.4%) and word processor (doc/docx) documents (n=1068; 26.1%). The materials contained in these files are mainly tables and figures, which completed the information of other tables and figures in the article. On the other hand, spreadsheets and raw data in general, associated with xls/xlsx or csv files, represented a 12.1% (n=496). Finally, the rest of the files analyzed were images (jpg/tif/png) (n=639; 15.6%), presentations (ppt/pptx) (n=119; 2.9%), and videos (mov/mp4/wmv/mpg/avi) (n=199; 4.9%) (Fig. 1).

The specific analysis of the raw data files showed that the spreadsheets were related mainly to the RNA and DNA sequences and the analysis of the expression of proteins with respect to the signaling pathways involved in embryology, the reproductive genetics, and gynecology. In addition, microarray analysis of significant genes involved in reproductive oncology through ovarian cancer lines was included. The raw data also included lists of participating patients' characteristics or, e.g., data on the embryonic development. Nevertheless, other raw data were more related to the results of statistical data analysis.

Several recent works have highlighted the growth of journals that allow the deposit of supplementary material, including datasets, due to the differential value that this possibility gives to the scientific community, related to a higher transparency and reproducibility, but it is not yet clear the motivation behind sharing data. Sometimes there is an obligation to deposit research data, both at editorial (Nature, PLoS) and official level (EU-H2020 guidelines and mandatory deposit in Zenodo). Thus, there is an increasing pressure in the open deposit of any type of file, which provokes also a certain lack of agreement in terms of the types of data available, as well as difficulties in the identification, format, and accessibility thereof, so that any quantitative analysis carried out in this regard must necessarily be accompanied by a manual review of deposited material [15–17].

One question arises about the acceptance of spreadsheets like raw data. There was in fact a limitation in this study, since the research team assumed that images, pdf, and word processor files are not raw data because they do not allow statistical analysis. However, an image can constitute a first glance into an issue; thus, it can be raw data as well. This limitation will be

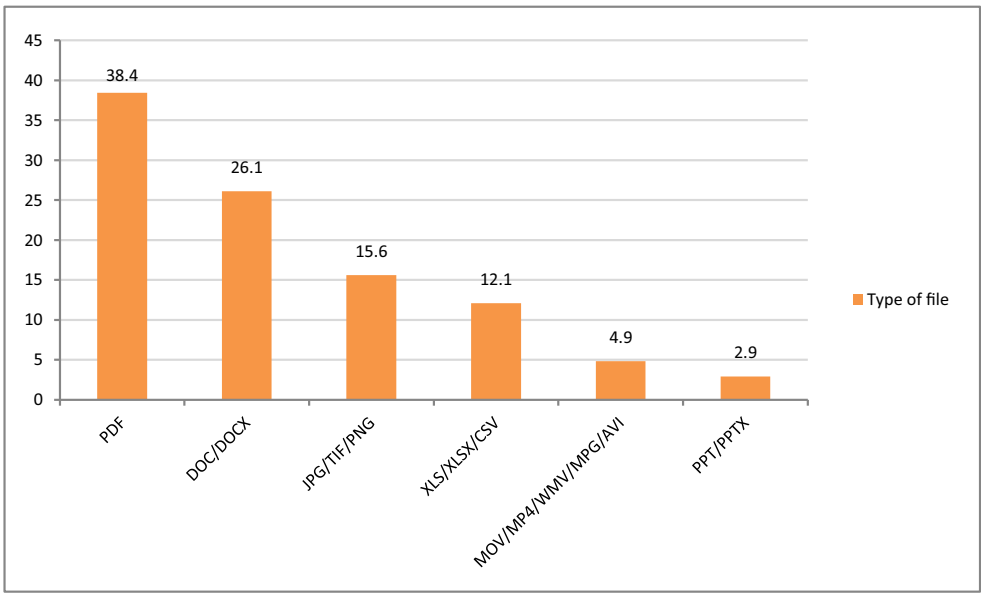

Fig. 1 Percentage of type of files found in supplementary material

corrected in further studies. There are some concerns that come up about the data sharing process being that nobody knows if the supplementary material is consulted or not, since the additional citations to the supplementary material are invisible. This concern could be solved by adding the supplementary material to the references section of the article.

Trying to answer the question that entitles this commentary, we do not know whether promoting the sharing of research data is the response for the new decade challenges in reproductive biology; we can only ensure that the collective is stronger than the individual. This effort of sharing data must be accompanied of a pedagogical intervention as early as possible about the importance of sharing discoveries, as well as a standardization of rules and processes used by the scientific community to share data following the FAIR principles (Findable, Accessible, Interoperable, Reusable) [18]. In this way, we highlight the need for robust indicators that serve to adequately and reliably measure the practice of data sharing, including clear mechanisms for citing data and accounting for these citations as valuable material. We also highlight the need for effective incentives for researchers who share their data (in terms of prestige, positive scores in competitive projects, access to better funding, etc.) and intuitive infrastructures including data repositories and useful ways to share supplementary material. Thus, it is necessary to pay attention to the reproductive biology area, because the growth of this practice involves significant changes in the way of doing and communicating science.

Author Contribution AVI and RAB conceived the design of the study. ASC and RLD completed the data collection and the data analysis. AVI coordinated the results section. ASC, RLD, RAB, and AVI collaborated to write the paper.

Funding This work benefited from assistance by Spanish Ministry of Science and Innovation (PID2019-105708RB-C22, PID2019-108579RB-I00 and BES-2016-079394) and the CIBERONC (CB16/12/00350).

Data Availability The data generated and used during this research are openly available from Zenodo.org public repository at DOI: https://doi.org/10.5281/zenodo.4159392

Declarations

Conflict of Interest The authors declare no competing interests.

References

1. Aleixandre-Benavent R, Simon C, Fauser BCJM. Trends in clinical reproductive medicine research: 10 years of growth. Fertil Steril. 2015;104(1):131–137.e5.
2. Gøtzsche PC. Strengthening and opening up health research by sharing our raw data. Circ Cardiovasc Qual Outcomes. 2012;5(2):236–7.
3. Max-Planck-Gesellschaft. Berlin declaration on open access to knowledge in the sciences and humanities. In: Berlin Open Access Conference, Max Planck Society and Max Planck Institute for the History of Science. Berlin; 2003. p. 4.
4. Nellåker C, Alkuraya FS, Baynam G, Bernier R, Bernier FP, Boulanger V, et al. Enabling global clinical collaborations on identifiable patient data: the Minerva initiative. Front Genet. 2019;10:1–9.
5. Tuggle CK, Giuffra E, White SN, Clarke L, Zhou H, Ross PJ, et al. GO-FAANG meeting: a gathering on functional annotation of animal genomes. Anim Genet. 2016;47(5):528–33.
6. Smith E, Haustein S, Mongeon P, Shu F, Ridde V, Larivière V. Knowledge sharing in global health research—the impact, uptake and cost of open access to scholarly literature. Health Res Policy Syst. 2017;15:73.
7. Liu SL, Saif L. Emerging viruses without borders: the Wuhan coronavirus. Viruses. 2020;12:e130.
8. Moorthy V, Restrepo AMH, Preziosi MP, Swaminathan S. Data sharing for novel coronavirus (COVID-19). Bull World Health Organ. 2020;98:150. https://doi.org/10.2471/BLT.20.251561.
9. Dutta S, Sengupta P. SARS-CoV-2 and male infertility: possible multifaceted pathology. Reprod Sci. 2020:8–11.
10. Roberts L. A tussle over the rules for DNA data sharing. Science. 2002;298(5597):1312–3.
11. Aleixandre-Benavent R, Lucas-Domínguez R, Sixto-Costoya A, Vidal-Infer A. The sharing of research data in the cell & tissue engineering area: is it a common practice? Stem Cells Dev. 2018;27(11):717–22. https://doi.org/10.1089/scd.2018.0036.
12. Vidal-Infer A, Aleixandre-Benavent R, Lucas-Domínguez R, Sixto-Costoya A. The availability of raw data in substance abuse scientific journals. J Subst Use. 2019;24(1):36–40. https://doi.org/10.1080/14659891.2018.1489905.
13. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res. 2012;40(1):109–14.
14. Vidal-Infer A, Tarazona B, Alonso-Arroyo A, Aleixandre-Benavent R. Public availability of research data in dentistry journals indexed in Journal Citation Reports. Clin Oral Investig. 2018;22(1):275–80.
15. Price A, Schroter S, Clarke M, McAneney H. Role of supplementary material in biomedical journal articles: surveys of authors, reviewers and readers. BMJ Open. 2018;8(9):1–7.
16. Park H, Wolfram D. An examination of research data sharing and re-use: implications for data citation practice. Scientometrics. 2017;111(1):443–461.
17. Aleixandre-Benavent R, Vidal-Infer A, Alonso-Arroyo A, Peset F, Ferrer-Sapena A. Research data sharing in Spain: exploring determinants, practices, and perceptions. Data. 2020:1–14.
18. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR guiding principles for scientific data management and stewardship. Sci Data. 2016;3:1–9.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.