# Embed Data Sharing Platforms into the Academic Evaluation System

**Main body:**

Recently, the NIH has issued a draft policy for data management and sharing aiming to strengthen public access to research data and open science goals (1). Similar efforts have been made by the European Commission through, among others, providing funding for various projects which aim to develop sustainable infrastructures for data sharing. One of the most recent examples was launching international flagship collaborations between the EU and Canada in 2017. Under this banner, projects such as euCanSHare and CINECA aim to facilitate data storage, interoperability and sharing, and have received funding for four-years to develop platforms for sharing data from disease and population cohorts across the EU and Canada.

Despite the massive efforts to develop these infrastructures for data sharing by the ongoing and completed initiatives, platform developers can be confronted with significant challenges in data sourcing due to reluctance of cohorts to broadly share their data. We stress that the current reward and crediting mechanisms embedded in academia are intensifying the challenges regarding the lack of incentives for data sharing which have been repeatedly voiced by researchers and policy makers (2). Building large-scale cohorts is labor-intensive, requiring entire teams of physicians, data curators, data managers and informaticians to assemble datasets over many months or years. These labors involved in data collection and curation have at times been described as “invisible”, as they can remain unrecognized in the academic reward system (3).

In response, the need for developing adequate approaches for crediting data sharing in the academic rewarding system has been put forward. Arguably, the traditional rewarding mechanisms including the granting of co-authorship in publications resulting from downstream analysis of data may not seem fit for purpose (4). Notably, systematically crediting all data generators has resulted in papers with hundreds of authors, contributing to the so-called “hyper-authorship” phenomenon and accelerating authorship inflation (5). This evolution has raised concerns over research integrity, such as the capacity of researchers to contradict prior conclusions of the data generators, how disputes over use of methodology should be resolved and the dilution of accountability (6). Furthermore, the undue influence of hyper-authorship on popular metrics of scientific productivity is a matter of concern (7).

Therefore, we suggest adopting an alternative approach, which leverages data-level metrics (hereafter “DLMs”) to capture and make data sharing efforts visible (see Fig. 1). The recording of these metrics, such as the number of downloads and metadata views can be integrated into emerging data sharing platforms and eventually be integrated into academic evaluations. DLMs can be seen as complementary to recent proposals for the specification of authorship roles on publications, such as the introducing the Data Authorship designation (6).

However, simply collecting DLMs through data sharing platforms is insufficient as it does not embed the platform within the broader academic ecosystem. The platform should systematically collect and transfer DLMs to digital spaces where they become visible and can be extracted by academic institution and funding organizations that are in charge of various types of academic evaluations. Without fulfilling these conditions, novel metrics will simply remain isolated in separate silos and not be put into practice. Thus, we make three recommendations on how connections between the platform, funders and academic institutions can be established to facilitate their use.

First, ORCID profiles should display metrics related to datasets researchers have contributed to, so that these can be used in evaluating academic performance. Researchers' identities, their academic work and metrics of research productivity should all be linked. Thus, datasets should be associated with a team of researchers or clinicians involved in data generation, curation or other pre-analytical roles within the data sharing platform. Scientific teams often collect cohort data over many years and the composition of the team might change over time. Therefore, the contributor roles of persons attached to datasets need to be dynamic. If data is re-used, this should contribute to dataset metrics.

Second, infrastructures that support Open Access/Open data such as OpenAIRE should receive metrics from data sharing platforms, and visualize DLMs for datasets over time. Notably, this option would fit well into the OpenAIRE Funder Dashboard that is designed to allow research funders and policy makers to monitor all their funded research outcomes. As such, this would provide funders with the possibility to see whether datasets have been uploaded, and to observe indicators of the scientific productivity of all datasets derived from their funding. This would address the problem that many funders with Open Data policies do not actively follow-up on sharing, primarily due to a lack of monitoring tools (8). Furthermore, it would also make the enforcement of mandates for sharing easier. Researchers can then be certain that sharing does not disadvantage them, as recording such metrics could increase their chances to acquire further funding (e.g. for expanding datasets with novel types of data or maintaining data curation services) (9).

Third, all collected data underlying DLMs should be made available for scientific research, so that they can be assessed, evaluated and refined (10). This is in line with the Open Science Policy Platform recommendation that: *"The data,* metadata and methods *that are relevant to research evaluation, including (...) citations, downloads and other potential indicators of academic re-use, should be publicly available for independent scrutiny and analysis by researchers, institutions, funders and other stakeholders"* (11). Without such availability, their credibility for use in research evaluation could be undermined and questioned (12). Thus, DLMs and information inherent to datasets such as cohort size, types of available data, phenotype richness and study type (e.g. disease cohort – population cohort) should to be made accessible. One way to realize this would be to pass on these data to the DataCite/Crossref Data Event service. This service is already collecting and collating similar metrics for datasets deposited within centralized repositories.

Despite DLMs offering novel opportunities to incentivize data sharing, they have their limitations as well. For instance, the use of data metrics may raise concerns about the possibility of the gaming (i.e. manipulation) of metrics, leading to unintended consequences (12). In addition, several relevant technical and governance issues also need to be addressed: If data re-use takes place over several data sharing platforms or central repositories, should these DLMs then be aggregated? Is it possible and desirable to attribute less credit for partial re-uses of the dataset? Should sharing alone, without re-use be in some way rewarded? These questions need to be discussed in view of the anticipated, downstream uses of DLMs in research evaluation.

Furthermore, the transition towards Open Science is described as a process of cultural change. This process involves the development of new policies, strategies and the evaluation of outputs and work against open criteria. To successfully realize these changes, an environment of trust, collaboration and commitment to a shared vision for the future is required (13). Notably, inertia against such cultural changes can be expected due to general conservatism in reward systems in academia, at times fueled by academics willing to preserve the system in which they have been successful (14). Therefore, community engagement with researchers, funders and institutions is necessary to raise support for the use of DLMs. All stakeholders involved should understand the uses, shortcomings and limitations of DLMs, and be committed to their development, application and fair use.

In recent years, the development of alternative credit systems in support of Open Science has been supported by many expert groups and organizations. The European Commission's Expert Group on Altmetrics encourages the development of new indicators to measure and support the development of Open Science (REC#2) (12). Additionally, they recommend that greater investment should be made into the field of 'meta-science' moving into Framework Programme 9. This includes research into the modeling of effects of these indicators, and into evaluation methods and practices (REC#3). Notably, REC#12 calls for *"next generation research data infrastructure[s], which can ensure greater efficiency and interoperability of data collection, and its intelligent and responsible use to inform research strategy, assessment, funding prioritisation and evaluation in support of open science".* In our view, data sharing platforms are examples of such next generation infrastructures and they could, in principle, be designed to advise research strategy and priorities. There are also indications that funders are open to other evaluation models for science. In the 2019 Scholarly Publishing and Academic Resources Coalition (SPARC) Report, approximately half of the funders have expressed support for or have signed the DORA Declaration, which calls for the abandonment of the Journal Impact Factor and to *"consider the* *value and impact of all research outputs (including datasets and software) in addition to research publications [for the purposes of research assessment]"* (8).

Finally, active collaboration and dialogue between researchers, metrics developers, (bio)informaticians and policy makers will be necessary to successfully tackle the incentive problems for Open Data. In addressing these issues, the onus should be on the manner in which these data sharing platforms can *inform* Open Data policies that will emerge in the coming years. By influencing and shaping policies at an earlier stage, it can be ensured that data are contributed to these infrastructures, while simultaneously providing scientists with proper credit. If DLMs can be designed with the support of the scientific community, and integrated into practice by policy makers, this would amount to the building of science policy around these platforms. Data sharing platforms are then rightfully recognized as indispensable components that can catalyze future data sharing and re-use in biomedical sciences.

**References**

1. NIH. Draft NIH Policy for Data Management and Sharing. 2019. Available from: https://osp.od.nih.gov/wp-content/uploads/Draft_NIH_Policy_Data_Management_and_Sharing.pdf
2. Ali-Khan SE, Harris LW, Gold ER. Motivating participation in open science by examining researcher incentives. *Elife*. 2017;6. Available from: https://doi.org/10.7554/eLife.29319.001
3. Ankeny RA, Leonelli S. Valuing Data in Postgenomic Biology: How Data Donation and Curation Practices Challenge the Scientific Publication System. In: Richardson SS, Stevens H. (eds.) *Postgenomics: Perspectives on Biology after the Genome*. Duke

University Press, USA: 2015. p. 126–149. Available from: https://doi.org/10.1215/9780822375449-007

4. Adams J, Pendlebury D, Potter R, Szomszor M. Global Research Report Multi-authorship and research analytics. 2019. Available from: https://clarivate.com/webofsciencegroup/campaigns/global-research-report-multi-authorship-and-research-analysis/

5. Cronin B. Hyperauthorship: A postmodern perversion or evidence of a structural shift in scholarly communication practices? *J Am Soc Inf Sci Technol*. 2001;52(7):558–69. Available from: https://doi.org/10.1002/asi.1097

6. Bierer BE, Crosas M, Pierce HH. Data Authorship as an Incentive to Data Sharing. *N Engl J Med*. 2017;376(17):1684–7. Available from: doi: 10.1056/NEJMsb1616595

7. Hu X, Rousseau R, Chen J. In those fields where multiple authorship is the rule, the h-index should be supplemented by role-based h-indices. *J Inf Sci*. 2010;36(1):73–85. Available from: https://doi.org/10.1177/0165551509348133

8. Fosci M, Richens E, Johnson R. Insights into European research funder Open policies and practices. 2019. Available from: https://zenodo.org/record/3401278#.Xn3-z4hKhPY

9. Sim I, Stebbins M, Bierer BE, Butte AJ, Drazen J, Dzau V, et al. Time for NIH to lead on data sharing. *Science*. 2020;367(6484):1308–9. Available from: DOI: 10.1126/science.aba4456

10. Hicks D, Wouters P, Waltman L, De Rijcke S, Rafols I. Bibliometrics: The Leiden Manifesto for research metrics. *Nature*. 2015;520(7548):429–31. Available from: doi:10.1038/520429a

11. Open Science Policy Platform. Open Science Policy Platform Recommendations. 2017;

Available from: https://ec.europa.eu/research/openscience/pdf/integrated_advice_opspp_recommendations.pdf

12. European Commission Expert Group on Altmetrics. Next-generation metrics: Responsible metrics and evaluation for open science. 2017. Available from: https://ec.europa.eu/research/openscience/pdf/report.pdf

13. Ayris P, López de San Román A, Maes K, Labastida I. Open Science and its role in universities : A roadmap for cultural change. *Leag Eur Res Univ*. 2018;24:13. Available from: https://www.leru.org/files/LERU-AP24-Open-Science-full-paper.pdf

14. Mutual Learning Excercise. Mutual Learning Exercise : Open Science – Altmetrics and Rewards Incentives and Rewards to engage in Open Science Activities. 2017;(3). Available from: https://rio.jrc.ec.europa.eu/sites/default/files/report/MLE%20OS_Final%20Report_0.pdf