Rewarding Research Data Management

Joachim Schöpfel, Otmane Azeroual

Rewarding Research Data Management

Joachim Schöpfel*
GERiiCO Laboratory, University of Lille
Villeneuve-d'Ascq, France
joachim.schopfel@univ-lille.fr

Otmane Azeroual
German Center for Higher Education Research and Science Studies (DZHW)
Berlin, Germany
azeroual@dzhw.eu

ABSTRACT

In the context of open science, good research data management (RDM), including data sharing and data reuse, has become a major goal of research policy. However, studies and monitors reveal that open science practices are not yet widely mainstream. Rewards and incentives have been suggested as a solution, to facilitate and accelerate the development of open and transparent RDM. Based on relevant literature, our paper provides a critical analysis of three main issues: what should be rewarded and incentivized, who should be rewarded, and what kind of rewards and incentives should be used? Concluding the analysis, we ask if it is really necessary and appropriate to consider RDM as an individual (behavioral) issue, as the main challenges are elsewhere, not personal, but technological, institutional and financial.

CCS CONCEPTS

• **Information systems → Data management systems;** ; • **General and reference → General literature**.

KEYWORDS

research data management, open science, incentives, rewards

ACM Reference Format:
Joachim Schöpfel and Otmane Azeroual. 2021. Rewarding Research Data Management. In *Proceedings of Sci-K 2021: 1st International Workshop on Scientific Knowledge Representation, Discovery, and Assessment, Apr 19 – 23, 2021, Ljubljana, Slovenia.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/1122445.1122456

1 THE CHALLENGE

Open science has become the new paradigm for research. Research is not an end in itself. Researchers should aim to make their research results available to others. Traditionally, research results have been disseminated in writing, as journal articles, in books, proceedings, reports… Collected and produced research data were usually not published. Thus, data could not be checked by experts. The open science paradigm requires that research data should not only be archived but also made freely available to third parties – "as open as possible, as closed as necessary".

The long-term archiving and the sharing of research data open up further perspectives for researchers to obtain the knowledge they need for their own research. If a researcher publishes the entire database accumulated during the research process, other researchers can build on it, pursue other questions and draw further conclusions without necessarily having to collect data themselves. Last but not least, the interpretation approach of the research results can be examined and a qualitative assessment can be given.

In the context of open science, research institutions, funding agencies, science policy committees and publishers started to formulate "data policies" as guidelines for research data management (RDM). The guidelines contain statements on the secure storage and sustainable archiving of the data, on the publication and reuse, and on the persistent identification and citability of the data. The general aim is the transparency of research, especially traceability and verifiability. More and more funders and publishers require scientists to share their data, like the EU programs Horizon 2020 and Horizon Europe which expect funded projects to create a data management plan and call for the sharing of the resulting research data [8].

However, studies and monitors reveal that open science practices are not yet widely mainstream. In particular regarding RDM, researchers have not yet universally embraced open science; they partly worry over "being scooped if they share their data too early", and they are concerned that they "simply will not be rewarded for engaging in Open Science practices in the processes that matter the most - funding evaluation, promotion and tenure reviews - to the same degree that they are rewarded for engaging in conventional closed practices" [12]. This concern is not new. Back in 2007, the Organisation for Economic Co-operation and Development (OECD) alerted that insufficient incentives for researchers may lessen their efforts on data-related activities, i.e., RDM, and suggested the development of new reward structures and the adaptation of existing ones, including recognition of RDM activities in tenure and promotion review, as a way to address this problem [17].

A couple of years later, the San Francisco Declaration on Research Assessment (DORA) claimed that the value and impact of research data should be included in research assessment; "best efforts should be made to integrate the recognition and rewards for researchers engaging in Open Science into existing and future funding mechanisms", at national, regional and institutional level [9]. The DORA proposal was indorsed by a report addressed to the European Commission, on the evaluation of research careers: "In order to increase the practice of Open Science, it is critical that researchers, who are the key agents of change towards (open science), are encouraged and incentivized" [16]. This position, i.e., the

*Corresponding author

Sci-K 2021, Apr 19 – 23, 2021, Ljubljana, Slovenia

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

acknowledgment that open data objectives depend on researchers' participation, has been transposed into national science policy, like The Netherlands whose National Plan Open Science outlines as one key objective the adaptation of evaluation and award systems to bring them into line with the objectives of open science (reward systems), recognizing that the present incentive system is inadequate for supporting this dependency [25].

The French National Plan for Open Science requires in a general way that the assessment system for researchers and research institutions must be updated to reflect the principles and practices of open science, giving greater weight to quality rather than quantity; it also announces a "research data award" to showcase and reward teams that have performed outstanding work in this area [15]. The purpose is clearly stated: the development of open science and more specifically, as for the researchers, the development of good practice regarding RDM, through incentives and rewards. But what exactly should be rewarded? And who? And how should good RDM practice be rewarded, by means of which kind of incentives?

2 THE OBJECT OF REWARDING

Open data describes the concept of making data freely available to the public for use, dissemination and processing. The idea is that publication of the research data should be guaranteed promptly. Collected data and information should be made available just as quickly as they were collected and recorded. The faster research data is made available, the greater the benefit that can be derived from the publication. A constant update of the published research data should be ensured, as a culture of data sharing has become established in the social and economic sciences in recent years.

Sharing and exchanging data or knowledge is possible because multiple use does not destroy it. The systematic argument in favor of data sharing is that only the possibility of re-analyzes of published research results turns them into scientific findings. Science means that results are verifiable.

that arise in the context of public funding, i.e., funded by tax payers, should be made available for broad research and not be monopolized by a single researcher.

The verifiability of research results through re-analyzes is one of the formalized criteria of good scientific practice that were developed by the research community. In Germany, for instance, enabling reuse of research data by transferring it to suitable data repositories are part of the funding guidelines of the National Research Foundation (DFG) and the Federal Ministry of Education and Research (BMBF). The consistent implementation of this obligation depends on the scientific discipline. Data sharing enables scientifically valuable feedback processes, so that the data producers can increase the quality of their data and the effectiveness of their data collection and analysis. Moreover, the researchers' results become better known through external data evaluation and thus also their reputation.

Numerous starting points for increasing efficiency and effectiveness in connection with open data, data sharing and data reuse can be discovered in the literature. The implementation of the approach requires internal preparation. This usually includes the development of the existing data and the establishment of an internal data management system. This also creates transparency within the facility. The newly gained transparency makes it possible to determine where which data is available, who is responsible for it and whether the same data was collected simultaneously at different locations in the past. This information is very valuable, because in the future the data can be collected once and used several times thanks to the internal cross-departmental accessibility.

The definition of open standards (e.g., formats, interfaces, metadata) makes it easier to find research data and thus creates interoperability [2]. Interoperability, in turn, allows processes (especially IT processes) to be optimized and data to be integrated directly into applications or software, for example via an open interface for application programming (API). Services can thus be provided more cost-effectively.

Networking data creates a new quality of data. New evaluations can be carried out and a broader knowledge base is created. This allows a better assessment of the current situation and contributes to targeted action. In addition, the proactive provision of research data, in addition to internal quality control, enables external quality control. Errors, contradictions and information gaps are discovered more quickly and can be rectified more quickly.

At the same time, the institutions can benefit from the open data offerings of other institutions. The open research data offered can be easily accessed and used or integrated into your own offers. Open data also makes it easier to use benchmarking as an instrument and consequently to contribute to internal administrative controlling. All of these factors ensure that processes can be optimized through better cooperation and thus working time resources can be saved. The administration can thus increase its efficiency.

Coming back to the question of incentives, rewarding (as a kind of conditioning) needs a more or less precise definition of the target behavior. What should be rewarded, which behavior should be incentivized? The OECD report mentions, in a rather general way, "data-related activities" and "data management activities", including planning for and execution of the proper documentation and archiving, "essential organization" and curation of data sets [17]. The European Open Science Career Assessment Matrix OS-CAM [16] describes four criteria regarding open science activities in the field of RDM:

- Using the FAIR data principles;
- Adopting quality standards in open data management and open datasets;
- Making use of open data from other researchers; and
- Being aware of the ethical and legal issues relating to data sharing, confidentiality, attribution and environmental impact of open science activities (integrity).

Other papers describe the target behavior simply as data sharing [12], data reuse and data sharing [23], or data activities related to research projects [24].

All this remains rather general and is not operational for efficient rewarding. For the moment, we can't find a kind of classification of "data behaviors" that would be helpful for rewarding and incentives. We can't find, moreover, any differentiated approach for the rewarding of "closed" RDM. Should researchers sharing "easy data" like code or seismic data be more rewarded than those handling sensitive data with privacy issues, such as medical data? This, of course, makes no sense. What about volumes of data sets? Another

issue is related to the respect of FAIR principles defined as a target behavior. Initially, the FAIR principles have been developed not as behavioral criteria but for the assessment of systems and infrastructures [10]. In other terms, FAIR seems less a description of specific data behavior than of the availability and the use of appropriate tools.

Finally, research information management systems, at least so far, usually do not represent correctly the whole range of data-related activities. Snowball metrics, a project for standardized research assessment, simply do not include data assessment [7]. Regarding RDM, these systems usually assess "data performance" in terms of data management plans, assignment of DOI to datasets, indexing (rich and standardized metadata), deposit in a labelled data repository (DataSeal, FAIR…), and data sharing [20]. Some of these indicators can be useful for rewarding and incentivizing. Yet, these data-related criteria are often in charge of (or shared with) information professionals, e.g., academic librarians or data officers, rather than of scientists. Also, who should be rewarded?

3 WHO SHOULD BE REWARDED?

At first sight, the answer seems easy to the question who should be the beneficiary of RDM rewarding: researchers, of course. Researchers produce research data, they do RDM, they are in the heart of the research process and, above all, they have impact on the system.

However, researchers are not an amorphous community but fall into more or less clearly defined categories, such as first stage or early career researchers, recognized researchers, established researchers and leading or senior researchers [16]. The challenge is not the same; especially early career researchers may be disadvantaged by the adoption of open science practices, e.g., because of restrictions of flexibility and time cost, without appropriate incentive structures or reward systems [1]. Specific and priority rewards for young researchers can be seen as an investment in the future.

On the other hand, senior researchers are in a position to change the system, as their assessment is relevant for recruitment, career progression and, through peer review, for funding and publishing. So, should they take the lead, should they be rewarded first and foremost, for the adoption of open science practice and for incentivizing other (younger) researchers to practice open science [16]?

The background of the researcher should be taken into account. Researchers seeking a position in academia from industry probably have been less engaged in open science activities [16].

The background of the researcher should be taken into account. Researchers seeking a position in academia from industry probably have been less engaged in open science activities [16]. Beyond the category and the background, there is another issue. Who exactly are researchers? In a recent reference work on scientific evaluation, the term researcher is a denominator "for any faculty or staff member who could act as the principal investigator of a funding application and who spends >0% time on research", including not only researchers who engage in "traditional" laboratory work, but also clinicians who are doing even a small amount of research, and librarians and professional research staff [7]. Regarding RDM and data-related open science practice, other staff members must be added, like data officers, data engineers or data librarians. They contribute to RDM in universities, research laboratories and other structures; should they be excluded of new open science incentives and rewards, because RDM is their job? This leads to another, fundamental question. Is (all) RDM research work? Should data sharing, i.e., description, structuring and deposit of datasets, be considered as part of usual research activity? Is cataloging, indexing, formatting and shelving of scientific papers part of research work? Partly? All of it? Discussions with colleagues and staff reveal different and sometimes opposite viewpoints.

At least for two reasons, one should be careful about giving an answer. First, it often depends on the work environment who does the job; when specialized staff is missing (no data librarian, no data engineer…), RDM is on the researcher's agenda; but this does not mean that all this is genuine research work. Second, the relationship between researcher and data is conditioned by the scientific domain and equipment. Especially in social sciences and humanities, where researchers often produce "their own data" and where their personal knowledge about the context of data collection and production is required for interpretation reuse, data-related activities are intimate part of research [13], which is a quite different situation compared to other research disciplines, with other equipment, tools and infrastructures [5].

4 WHICH KIND OF INCENTIVES?

Years ago, the OECD alerted that attention should be paid to incentives and the development of professional expertise in all areas of RDM [17]. More recently, an empirical study suggested that journals should provide incentives for following open science practices, especially for sharing research data [21]. But which kind of incentives?

Incentives have been defined as an external influence that incites someone to act [4], as a benefit, reward, or cost that motivates an action [19]. Regarding RDM, usually two categories of incentives are brought up: incentives for research career development, and funding. In other words, tenure decisions, promotions, annual salary decisions on the one side, including hiring, and funding of research projects on the other side [12] [26].

Other incentives are mentioned less often, such as encouragement by the supervisor or employer [23], "giving attention" on university's website or in promotional events or awards (research prizes) [16], or a "temporary exclusive use of the data" which is sometimes awarded to the initial data-producing researcher or institution and which could be developed and formalized by the funding sources in co-operation with the research communities [17].

Considering data sharing, researchers are generally well aware of personal benefits, like career or performance advantages, even if the degree to which benefits arise varies in relation to disciplines and hierarchy levels [22]. Then again, nevertheless, a systematic review in the field of health and medical research revealed that there are in fact few evidence-based incentives for data sharing, with open data badges being the only tested incentive. "The irony is that we live in an evidence-based world, which is built upon the availability of raw data, but we hardly have any evidence to demonstrate what will motivate researchers to share data" [19].

Those badges are promoted by the Center for Open Science (COS) as incentives for researchers to share data, materials, or to

preregister; they signal to the reader that the content has been made available and certify its accessibility in a persistent location. So far, however, only very few journals offer open science badges to signal and reward when underlying data, materials, or preregistrations are available[1].

The COS badge initiative emphasizes the publisher's responsibility and the crucial role of academic journals for data sharing and reuse, especially in period of COVID-19 pandemic [3]. Journals should provide incentives for following open science practices and not only encourage, but make adherence mandatory [21]. Another paper asserts that this publisher's responsibility should include innovation and better infrastructures [11].

Politicians and research managers use to speak about carrots and sticks, a combination of reward and punishment to induce the desired open science behavior. Behavioral psychologists know that generally, reward is more efficient for positive reinforcement than punishment. In the field of RDM, the incentives most often put forward are not really positive awards. Funding, promotion, tenure decisions – all this is already in place and part of normal academic life. Requiring data curation, sharing and reuse as additional criteria for these decisions, does not provide new benefits but represents additional workload and makes it more difficult to get those funding, salary or jobs. In the past, applying for a research grant didn't involve the writing of a data management plan; now it does. This is more a potential threat, a stick "from above" (top down), than a carrot. It may work; but it is not helpful to promote RDM and open science as a positive value and objective.

5 CHANGING THE APPROACH

Open science has become the new paradigm for research. RDM, in particular data sharing and data reuse, is one of the main pillars of open science. The challenge is how to get there. The purpose of our paper is to comment rewarding and incentivizing data-related behaviors, which are in several papers, reports and conferences described (and requested) as the (a) best way to achieve the goal.

This approach may appear intuitive and self-evident, yet there are major issues:

- Rewarding and incentivizing require a clearly defined target behavior. However, so far, a comprehensive classification of data-related and contextualized behavior is missing.
- Most proposals of rewarding and incentivizing put the focus on the researcher. Yet, research data management is team work with essential contributions by data librarians, data engineers, data stewards and so on.
- Some of the most often suggested incentives (tenure, promotion, funding) are no real rewards – they already exist and the proposal would just add new conditions to get them. More sticks than carrots, in some way.

Because of these major issues, our position is to readjust the approach to rewarding research data management, in order to improve its efficiency.

Contextualization

Any initiative to help researchers and other staff doing "good RDM" must take into consideration the context and adopt a systemic approach. The multidimensional character of research data policy, management, sharing and data reuse, has been highlighted by Borgman [5][4]. Research data in social sciences and humanities, in particular, requires more attention and care than simplistic formula, if the purpose is better science [13]. This would include a finer differentiation of incentives, as it has been suggested for the peer-to-peer review model ("personal" v. "general" incentives, "immediate" v. "long term" incentives…) allowing, also, for greater scalability [14].

Bottom-up, not top-down

In the terms of the Roundtable on Aligning Research Incentives for Open Science launched in 2019 by the US National Academies of Science, Engineering and Medicine (NASEM): the goal should not be to create a "monolithic set of 'one-size-fits-all' incentives" [12]. Incentives and rewards should be driven by the research and data communities themselves, in a multi-year perspective, reflecting faculty needs, and with institutional and political support if necessary: as bottom-up as possible, top-down only if necessary.

Technological, institutional and financial support

Rewarding and incentivizing "good data practice" has a proclivity to frame this issue as an individual behavioral problem. Yet, research data management is not an individual problem. The OECD report puts it very clearly: the main issues are not personal but technological, institutional and financial [17]. Following the international Confederation of Open Access Repositories (COAR), useful and efficient functionalities and tools and a supportive environment (advocacy, mandate, agreements with publishers…) are more important for sustainable practice than rewards and incentives [6].

Institutions and authorities should change the environment rather than the person and provide appropriate (FAIR) tools, services, infrastructures and funding rather than individual awards; enabling rather than rewarding. Researchers are not children, and research data management is worth more than some symbolic lollies or candies. In fact, we should drop research data activities from the evaluation of individual scholarship [18].

[1] COS, https://www.cos.io/initiatives/badges

REFERENCES

[1] Christopher Allen and David M. A. Mehler. 2019. Open science challenges, benefits and tips in early career and beyond. *PLOS Biology* 17, 5 (2019), e3000246. https://doi.org/10.1371/journal.pbio.3000246

[2] Otmane Azeroual and Nico Herbig. 2020. Mapping and semantic interoperability of the German RCD data model with the Europe-wide accepted CERIF. *Information Services and Use* 40, 1-2 (2020), 87–113. https://doi.org/10.3233/ISU-200076

[3] Lonni Besançon, Nathan Peiffer-Smadja, Corentin Segalas, Haiting Jiang, Paola Masuzzo, Cooper Smout, Eric Billy, View Maxime Deforet, and Clémence Leyrat. 2020. Open Science Saves Lives: Lessons from the COVID-19 Pandemic. *BioRxiv* 8, 13 (2020), 249847. https://doi.org/10.1101/2020.08.13.249847

[4] Christine L. Borgman. 2012. The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology* 63, 6 (2012), 1059–1078. https://doi.org/10.1002/asi.22634

[5] Christine L. Borgman. 2015. *Big data, little data, no data : scholarship in the networked world.* MIT Press, Cambridge MA. https://mitpress.mit.edu/books/big-data-little-data-no-data

[6] COAR. 2013. *Incentives, Integration, and Mediation: Sustainable Practices for Populating Repositories. Confederation of Open Access Repositories.* http://www.coar-repositories.org/files/Sustainable-best-practices_final.pdf

[7] Lisa Colledge. 2017. *Snowball Metrics Recipe Book: Standardised research metrics - by the sector, for the sector.* Snowball Metrics. https://www.snowballmetrics.com/metrics/snowball-metrics-recipe-book/

[8] European Commission. 2017. *Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020.* https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf

[9] DORA. 2012. *San Francisco Declaration on Research Assessment.* http://www.ascb.org/dora/

[10] Mark D. Wilkinson et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3, 1 (2016), 160018. https://doi.org/10.1038/sdata.2016.18

[11] Iain Hrynaszkiewicz. 2019. Publishers' Responsibilities in Promoting Data Quality and Reproducibility. In *Good Research Practice in Non-Clinical Pharmacology and Biomedicine. Handbook of Experimental Pharmacology*. Vol. 257. Springer, Cham, 319–348. https://doi.org/10.1007/164_2019_290

[12] Heather Joseph. 2021. Building Momentum to Realign Incentives to Support Open Science. *Data Intelligence* 3, 1 (2021), 71–78. https://doi.org/10.1162/dint_a_000795

[13] Marie-Laure Malingre, Morgane Mignon, Cécile Pierre, and Alexandre Serres. 2019. Construction(s) et contradictions des données de recherche en SHS. *Recherche d'information, Document et Web Sémantique* 2, 1 (2019), 1–21. https://doi.org/10.21494/ISTE.OP.2019.0336

[14] Chao man Chang and Roeland H. R. M. Aernoudts. 2010. Towards Scholarly Communication 2.0: Peer-to-Peer Review Ranking in Open Access Preprint Repositories. *Social Science Research Network* (2010). http://ssrn.com/abstract=1681478

[15] MESRI. 2018. *Plan national pour la science ouverte. Paris: Ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation.* https://www.ouvrirlascience.fr/plan-national-pour-la-science-ouverte/

[16] Conor O'Carroll, Bernard Rentier, Cecili Cabello Valdes, Fulvio Esposito, Eeva Kaunismaa, Katrien Maas, Janet Metcalfe, David McAllister, and Karen Vandevelde. 2017. *Evaluation of Research Careers fully acknowledging Open Science Practices: Rewards, incentives and/or recognition for researchers practicing Open Science.* https://doi.org/10.2777/75255

[17] OECD. 2007. *OECD Principles and Guidelines for Access to Research Data from Public Funding.* https://www.oecd.org/sti/inno/38500813.pdf

[18] Mario Pagliaro. 2021. Purposeful Evaluation of Scholarship in the Open Science Era. *Preprints* (2021), 2021010387. https://doi.org/10.20944/preprints202101.0387.v1

[19] Anisa Rowhani-Farid, Michelle Allen, and Adrian G. Barnett. 2017. What incentives increase data sharing in health and medical research? A systematic review. *Research Integrity and Peer Review* 2, 1 (2017), 4. https://doi.org/10.1186/s41073-017-0028-9

[20] Joachim Schöpfel, Hélène Prost, and Violane Rebouillat. 2017. Research Data in Current Research Information Systems. *Procedia Computer Science* 106 (2017), 305–320. https://doi.org/10.1016/j.procs.2017.03.030

[21] Manuel Spitschan, Marlene H. Schmidt, and Christine Blume. 2020. Transparency and open science principles in reporting guidelines in sleep research and chronobiology journals. *Wellcome Open Research* 5 (2020), 172. https://doi.org/10.12688/wellcomeopenres.16111.1

[22] Stefan Stieglitz, Konstantin Wilms, Milad Mirbabaie, Lennart Hofeditz, Bela Brenger, Ania López, and Stephanie Rehwald. 2020. When are researchers willing to share their data? – Impacts of values and uncertainty on open data in academia. *PLOS ONE* 15, 7 (2020), e0234172. https://doi.org/10.1371/journal.pone.0234172

[23] Bettina Suhr, Johanna Dungl, and Alexander Stocker. 2020. Search, reuse and sharing of research data in materials science and engineering—A qualitative interview study. *PLOS ONE* 15, 9 (2020), e0239216. https://doi.org/10.1371/journal.pone.0239216

[24] Clifford Tatum. 2017. What is the evaluative object of Open Science?. In *22nd Nordic Workshop on Bibliometrics and Research Policy, Helsinki, 09 November 2017.* https://figshare.com/articles/What_is_the_evaluative_object_of_Open_Science_/5624728

[25] Clifford Tatum and Sarah De Rijcke. 2018. *The (mis)alignment of Open Science and research evaluation: addressing complexity with existing resources and context-sensitive evaluation.* EuroCRIS Strategic Membership Meeting Autumn 2018 (Warsaw University of Technology, Warsaw, Poland, Nov 26-28, 2018). https://dspacecris.eurocris.org/handle/11366/719

[26] Ruben Vicente-Saez, Robin Gustafsson, and Lieve Van den Brande. 2020. The dawn of an open exploration era: Emergent principles and practices of open science and innovation of university research teams in a digital world. *Technological Forecasting and Social Change* 156 (2020), 120037. https://doi.org/10.1016/j.techfore.2020.120037