BULLETIN OF THE POLISH ACADEMY OF SCIENCES
TECHNICAL SCIENCES, Vol. 69(1), 2021, Article number: e135401
DOI: 10.24425/bpasts.2020.135401

OECD Recommendation's draft concerning access to research data from public funding: A review

Lech MADEYSKI[1]*, Tomasz LEWOWSKI[1], and Barbara KITCHENHAM[2]

[1] Faculty of Computer Science and Management, Wroclaw University of Science and Technology, ul. Wybrzeze Wyspianskiego 27, 50-370 Wroclaw, Poland

[2] School of Computing and Mathematics, Keele University, Keele, Staffordshire, ST5 5BG, UK

Abstract. Sharing research data from public funding is an important topic, especially now, during times of global emergencies like the COVID-19 pandemic, when we need policies that enable rapid sharing of research data. Our aim is to discuss and review the revised Draft of the OECD Recommendation Concerning Access to Research Data from Public Funding. The Recommendation is based on ethical scientific practice, but in order to be able to apply it in real settings, we suggest several enhancements to make it more actionable. In particular, constant maintenance of provided software stipulated by the Recommendation is virtually impossible even for commercial software. Other major concerns are insufficient clarity regarding how to finance data repositories in joint private-public investments, inconsistencies between data security and user-friendliness of access, little focus on the reproducibility of submitted data, risks related to the mining of large data sets, and sensitive (particularly personal) data protection. In addition, we identify several risks and threats that need to be considered when designing and developing data platforms to implement the Recommendation (e.g., not only the descriptions of the data formats but also the data collection methods should be available). Furthermore, the non-even level of readiness of some countries for the practical implementation of the proposed Recommendation poses a risk of its delayed or incomplete implementation.

Key words: open data; open access; empirical research; data-driven research; data science.

1. Introduction

We present and review the the revised Draft of the OECD (Organisation for Economic Co-operation and Development) Recommendation Concerning Access to Research Data from Public Funding (referred to further as the Recommendation) aiming to replace the recommendation adopted by the OECD Council on 14 December 2006 [1]. The reviewed Recommendation document (dated June 25, 2020) was prepared by the OECD's Committee for Scientific and Technological Policy (CSTP) and sent for consultation to the Polish Academy of Sciences via the Ministry of Science and Higher Education in Poland[1]. The Recommendation provides a framework for sharing research data from public funding. It is an important topic in general, but even more now, during times of global emergencies, e.g., the COVID-19 pandemic, when we need policies that enable the sharing of research data to accelerate research in the fight against the disease. Thus, we present the overview of the most recent draft Recommendation sent for consultation in Section 2, review the Recommendation in Section 3, and formulate conclusions in Section 4.

[1] Someone interested in the Recommendation may contact Krzysztof Przemieniecki (e-mail: Krzysztof.Przemieniecki@mnisw.gov.pl) from the Ministry of Science and Higher Education in Poland who coordinates the consultation process.

*e-mail: lech.madeyski@pwr.edu.pl

Manuscript submitted 2020-09-01, revised 2020-09-01, initially accepted for publication 2020-10-21, published in February 2021

2. Overview of the Recommendation

The Recommendation consists of a preface and seven sections[2]. The earlier recommendation (from 2006) that is aimed to be replaced covers broadly the same topics. However, the original document was more focused on general principles, whereas the new version of the Recommendation concentrates more on the need for national and international infrastructure systems to store and to re-use the data. In this section, we present an overview of the new Recommendation sent for consultation.

The preface refers to existing OECD recommendations and guidelines closely related to the discussed Recommendation, the wider context (e.g., the rapid growth of data produced by and used in scientific research and innovation), as well as the definition of basic terms used, e.g., research data and research-relevant digital objects from public funding, algorithms, code, software and workflows used to generate research results.

Section 1 (*Data governance for trust*) describes rules related to organizing data access—the goal is to make the data available to as wide an audience as possible (for both individuals and organizations, including international data sharing), except for sensitive data, which, according to the Recommendation, should be provided on a need-to-know basis. The Recommendation also acknowledges the need for data anonymization and limiting consent to specific usages.

Section 2 (*Technical standards and practices*) describes some of the technical aspects that Adherents (states) should follow

[2] One section in the reviewed draft is not numbered, probably by mistake.

managing publicly-funded data. The aspects mentioned in this section are focused mainly on making the data findable and durable—mentioned techniques include availability under immutable links such as Digital Object Identifiers (DOIs), consistency with domain conventions and metadata availability. This section also recommends providing non-data digital objects (such as code, algorithms or workflows) as a free and open source.

Section 3 (*Responsibility and ownership*) is a call for promoting good practices, data re-use and open licensing. It also recommends promoting access to research data resulting from public-private partnerships in ways ensuring that data collected with public funds is as open as possible.

Section 4 (*Incentives and rewards*) recommends that providing data should be properly recognized and rewarded. The document discusses this in a little more detail, pointing out that recognition requires developing data assessment criteria and indicators of impact, as well as contributor taxonomies. It also emphasizes promoting data and software citation in academic practice, as well as taking it into account while developing criteria for researcher recruitment, advancement, and grant review.

Section 5 (*Sustainable infrastructures*) describes some of the actions expected from Adherents, such as developing infrastructures, including data and software repositories and services, ensuring data prioritization, safeguarding and co-operating with the private sector for the public interest.

The unnumbered section entitled *Human capital* discusses basic needs in the area of training programs and providing career paths for data experts necessary for data-driven research, as well as basic literacy skills for the general population.

The final section (*International co-operation for access to research data*) recommends setting up international fora and including international initiatives as part of national strategy. In particular, it mentions developing internationally compatible procedures for assessing the sensitivity of data, allowing access to, and establishing secure remote access to such data.

3. Review of the Recommendation

In this section, we first discuss the infeasible guideline, as well as the proposed enhancement to the Recommendation to make it more actionable. Then we present ancillary risks and threats in terms of practical implementation issues, the risk of unintended consequences, as well as expected risks.

3.1. Issue and fix related to regular maintenance of software design. Section 1.1.III.1.b of the Recommendation urges the following:

> "Adherents should:
> *1. Foster and support open access by default to research data and other research-relevant digital objects from public funding, for both individuals and organizations, that, to the greatest extent possible, is:*
> …
> *b. supported by regular maintenance to prevent obsolescence of data format* ***and software design***;".

When phrased that way, one can read it as "*every software component created from public funding should be maintained and its design should be updated as long as the component is relevant*". While this would be an ideal situation, in practice it is simply not possible—software design decays [2], and even commercial products face the situation when their design becomes obsolete. Commercial products have, generally speaking, much more stable funding, so if maintaining software design in these is rare, it is unrealistic to expect such maintenance from a publicly funded project. The Recommendation seems to acknowledge this fact by adding "*to the greatest extent possible*".

In fact, the case is even worse—software nowadays rarely exists as a self-contained package, and often relies on external services (APIs) to complete its tasks. Since these services are managed by third parties, they cannot be a part of the archived package (as they are not created from public funding, but usually are commercial offerings). Most API providers have policies for versioning their interfaces, deprecating and removing functionalities. While these policies are generally not an issue for a maintained software product (as providers keep old versions running for at least a few months), they will definitely be an issue for an archived project in a few years (not to mention that used providers may be out of business by that time).

We are slightly concerned that a literal understanding of this section may be a disincentive to create and submit algorithms and code due to high maintenance effort. On the other hand, without literal understanding this part of the Recommendation is likely to be entirely ignored.

Possible fix would be to rephrase the sentence into "*b. supported by regular maintenance to preserve data format readability and keep the software executable*". While problems with APIs are inevitable, this at least would not hint at an obligation to follow recent design trends. It is likely that Adherents should be allowed to define various "Maintenance levels" so that most effort would be put into maintaining the most valuable data, code, designs and algorithms (as hinted is Section 1.5.VII of the Recommendation).

3.2. Issue related to reproducibility. Sections 1.1.III (*Data governance for trust*) and 1.2.IV (*Technical standards and practices*) of the Recommendation mention multiple principles that should be supported: timeliness, findability, user-friendliness, accessibility, interoperability, reusability, openness. What those sections omit, is reproducibility—it is briefly mentioned in the preface, but we believe it should receive more attention. Irreproducibility—which may either be planned or accidental—is a serious problem in many fields of science like medicine [3], biomedical studies [4], and software engineering [5] and, in our opinion, should definitely be included as a principle to be followed. On a practical level, this would include detailed research protocols, data collection methods and datasheets or even automated scripts ran to gather data [6]. Only by including all relevant details we can verify research claims and avoid problems like comparing results obtained from two data sets that were sampled from different populations—a problem especially visible during COVID-19 pandemic, where infection

data from various countries is incomparable (and sometimes is incomparable even within a country, when testing rules change).

In the subsequent section, we discuss risks and threats that can be seen as ancillary, but we think they are also worth addressing (to the greatest extent possible) in the reviewed Recommendation, as well as its future revisions.

3.3. Ancillary risks and threats. Our goal is to examine ancillary problems and risks of three types: practical implementation issues, unintended consequences and expected risks.

3.3.1. Practical implementation issues. We have identified two practical implementation issues:

1. Funding for infrastructure development and maintenance. The only issue directly mentioned in the Recommendation that might be considered controversial is the guideline to encourage private investment in research data infrastructures which some researchers may consider poses a threat to the independence of university research. Large scale research databases are a valuable resource which is likely to *increase* in value over time. Selling access to national resources to commercial organizations, perhaps with links to other countries, in return for providing infrastructure systems would require very careful negotiations with potential suppliers to ensure that the public benefits from the data, not just the infrastructure supplier. We believe that the resources of government-funded organizations are often far more limited than the resources available to private organizations when it comes to licensing, intellectual property and data ownership issues. The current discussion concerning medical data [7–10] might be a useful test case for identifying the risks and problems associated with funding database infrastructure and data maintenance. The critical issue is to balance the need for funding to support the infrastructure and the maintenance process required to implement the Recommendation, with its goal to gain additional benefit to the public from the outcomes of publicly-funded research.
2. Requirements for timely, findable, user-friendly, preferably Internet-based data storage facilities are in conflict with requirements for secure, access-controlled data. Meeting conflicting requirements is difficult and costly, and therefore adds more complexity to the basic infrastructure funding issue. It is critical that all non-functional requirements are fully specified and the decisions with respect to trade-offs have been discussed and agreed upon among the relevant stakeholders *prior* to awarding any contracts for software platform design and development.

3.3.2. The risk of unintended consequences. In the social sciences, unintended consequences (also referred to as unanticipated consequences or unforeseen consequences) are outcomes of a purposeful action that are not intended or foreseen. They can be grouped into three types:

1. Unexpected benefit: a positive unexpected benefit (also referred to as luck, serendipity or a windfall).
2. Unexpected drawback: an unexpected detriment occurring in addition to the desired effect of the policy.
3. Perverse result: a perverse effect contrary to what was originally intended (when an intended solution makes a problem worse).

In the context of open access, the rise of the predatory publishing industry might be considered an unexpected drawback. Drawing researchers from countries with few libraries and little access to high quality research into reliance on predatory publishing literature might be considered a perverse result. The increase in the numbers of high impact open access journals focused solely on open source software and algorithms (e.g., *Journal of Statistical Software*) might be considered an unexpected benefit of the rising popularity of open access journals.

It is hard to believe that there will be no unintended consequences of establishing national databases of publicly funded research data or that all unintended consequences will be positive.

3.3.3. Expected risks. Ideally we would like to be able to identify the novel risks and problems that might arise from the OECD Recommendation, but it might be useful to consider risks extrapolated from initial trends as a starting position for risk analysis.

1. Clearly, we can expect any database of potentially valuable research data to be a magnet for hacking attempts. This might be "just for fun" attacks from individual hackers, but could also involve industrial espionage sponsored by organizations or states who want to avoid paying licensing fees or obeying access restrictions. We should also be prepared for attempts to corrupt or destroy valuable data to undermine the work of legitimate research groups. This risk emphasizes the need for well-defined security requirements.
2. While the Recommendation acknowledges the need for seeking consent in the case of using personal data for new purposes, such action is extremely hard to enforce in practice, especially while maintaining as wide access to data as possible. After the introduction of GDPR and other data protection acts the concept of sensitive data is understood much better. However, the protection given by them to an individual may still be insufficient. As of now, proper risk assessment still requires experts in sensitive data handling for compliance. It is likely that Adherents will need to set up special repositories for sensitive data and data stewardship organizations to avoid abuse.
3. We second the idea of providing as many publicly funded digital objects as open source as possible (in fact, we develop the `reproducer` R package that includes open data sets from a range of empirical studies in software engineering, e.g., [11]). However, we would like to point out that "Open Source" is not exactly a well-defined term. There are at least several definitions (to name a few defining parties: Free Software Foundation, Open Source Initiative, Open Knowledge International), which are not fully compatible. This is a potential pitfall for Adherents that must be considered carefully, as the range of licenses acceptable as "Open Source" varies depending on chosen definition.
4. Organizations might try advanced data mining techniques to overcome anonymity and use the data in combination with

other data sources for potentially criminal purposes. The Cambridge Analytica scandal is a case in point. Cambridge Analytica Ltd (CA) was a British political consulting firm that combined misappropriation of digital assets, data mining, data brokerage, and data analysis with strategic communication during electoral processes. In 2016, CA worked for Donald Trump's presidential campaign as well as for Leave. EU (one of the organizations campaigning in the United Kingdom's referendum on European Union membership). CA's role in those campaigns is the subject of ongoing criminal investigations in both countries. This emphasizes the need for more well-defined, appropriate and audited access control to stored data.

5. There is a risk related to data mining using Artificial Intelligence (AI)/Machine Learning (ML) techniques. They use very large data sets to look for patterns among large numbers of variables but they can overfit, they can reflect the prejudices of their builders, and they assume that an ultra large data set has equivalent properties to a random data set. So we can expect more data mining studies based on analysis of public data sets, some providing extremely valuable scientific outcomes, but we can also expect a number of flawed and invalid analyses [12–14]. We fully recognise the value of data mining and machine learning, but we also believe that our scientific research and commercial exploitation must agree with the maxim "*first, must do no harm*".
6. The Recommendation includes incentives and rewards for data scientists which would encourage the production of meta-analytic studies. Unfortunately, this implies that research publishers can expect to be swamped with poor quality meta-analyses. In 2016, Ioannidis [15] identified that there is massive production of unnecessary, misleading, and contradictory systematic reviews and meta-analyses. He pointed out that instead of promoting evidence-based medicine and health care, they are either an easy way of increasing publication counts or marketing tools. He also pointed out that poor quality secondary studies are harmful "*given the major prestige and influence these types of studies have acquired*". He suggests that systematic reviews and meta-analyses studies need to better avoid biases and vested interests and to be better integrated with primary studies.
7. Another implication of the incentives and rewards is that we can expect to see more fake data (for an example of the prevalence of fake data see [16]). There are various techniques to detect fake data. Assuming the data set is available, one can use a range of strategies to figure out whether the numbers seem "*fabricated*" by the researcher. For example, it is known that humans have a preference for numbers ending in round values. Thus a χ-squared test can be used to see if these round numbers are too prevalent in the data set. There are also available statistical tools to detect data fabrication [17, 18]. Furthermore, one may use summary statistics to check for potential fraud or error, e.g., Brown and Heather's GRIM (Granularity-Related Inconsistency of Means) test [19] that can detect problems when the summary statistics have been fabricated, as well as related software tools, e.g., "*Sprite*" (Sample Parameter Reconstruction via Iterative Techniques) [20] (see also https://hackernoon.com/introducing-sprite-and-the-case-of-the-carthorse-child-58683c2bfeb). There are also known statistical methods for the detection of data fabrication in clinical trials (e.g., [21]). However, the database itself would be a source of real data sets that might be used to improve the construction and plausibility of fake data sets. Continual improvement of procedures to detect fake data integrated with appropriate auditing of submitted data are essential to protect the integrity of a repository of scientific data.

4. Conclusions

In this paper, we discuss and review the revised Draft of the OECD Recommendation Concerning Access to Research Data from Public Funding. In our opinion, the scientific principles embodied in the Recommendation are excellent and should be supported. However, as we point out in Section 3.1, the Recommendation must conform with the reality of software design and maintenance, which is clearly not the case, thus we have formulated a more actionable statement of the maintenance requirement.

As an extension to the Recommendation, in Section 3.2, we propose to include reproducibility as one of principles that should be considered when building data sets and supporting infrastructure with particular focus on data collection and sampling procedures.

Furthermore, the Recommendation seems to assume that all individual researchers, research groups, industry and nations are prepared to work together for the general public good. However, as discussed in Section 3.3, a world-wide repository of research data (even if it is distributed among different Adherents) is a potentially valuable target for both unethical and criminal behavior at the individual, organizational and national level. In particular, individual guidelines aimed at ensuring data can be easily reused are at odds with the need to prevent data from being corrupted or misused. To address the issues raised in Section 3.3, we suggest that OECD add the following two elements to Section 1.2 (*Technical standards and practices*) of the Recommendation:

1. *Put in place on-going risk management processes to identify, manage and monitor risks associated with maintaining the integrity and security of data*.
2. *Establish procedures to audit the quality of studies based on reusing the data*.

We also note that monitoring the use of data repositories will allow Adherents to publicise the benefits delivered from the reuse of publicly-funded data.

In addition, we suggest that that OECD add the following element to Section 1.1 (*Data governance for trust*) of the Recommendation: *Cooperate with the community to create rules and processes for retiring data and other digital objects developed under publicly-funded projects, which would take into account its relevance, integrity, requirement maintenance effort and other metrics*.

In conclusion, we believe that the Recommendation should be supported, but to apply it in real settings, we suggest the Recommendation be refined to address the issues we raise in Section 3.

Acknowledgements. The authors thank the reviewers for their helpful comments. The authors were invited to prepare their joint review presented in this work by the Presidium of the Committee on Informatics of the Polish Academy of Sciences.

References

[1] OECD, Recommendation of the council concerning access to research data from public funding. [Online]. https://legalinstruments.oecd.org/en/instruments/OECDLEGAL-0347

[2] D.N. Le, A. Shahbazian, and N. Medvidovic, "An Empirical Study of Architectural Decay in Open-Source Software", *IEEE International Conference on Software Architecture (ICSA),* 2018.

[3] K.R. Sipido, "Irreproducible results in preclinical cardiovascular research: Opportunities in times of need", *Cardiovasc. Res.* 115 (3), E34–E36 (2019).

[4] D.A. Eisner, "Reproducibility of science: Fraud, impact factors and carelessness", *J. Mol. Cell. Cardiol.* 114, 364–368 (2018).

[5] L. Madeyski and B. Kitchenham, "Would wider adoption of reproducible research be beneficial for Empir. Softw. Eng. research?", *J. Intell. Fuzzy Syst.* 32(2), 1509–1521 (2017).

[6] T. Lewowski and L. Madeyski, "Creating Evolving Project Data Sets in Software Engineering", *Integr. Res. Pract. Softw. Eng.* 851, 1–14 (2020), doi: 10.1007/978-3-030-26574-8_1.

[7] T. Moberly, "Should we be worried about the NHS selling patient data?", *BMJ* 368, m113 (2020). doi: 10.1136/bmj.m113.

[8] C. Aicardi, L. Del Savio, E.S. Dove, F. Lucivero, N. Tempini, and B. Prainsack, "Emerging ethical issues regarding digital health data. On the World Medical Association Draft Declaration on Ethical Considerations Regarding Health Databases and Biobanks", *Croat. Med. J.* 57(2), 207–213 (2016), doi: 10.3325/cmj.2016.57.207.

[9] E. Mahase, "Government hands Amazon free access to NHS information", *BMJ* 367, l6901 (2019), doi: 10.1136/bmj.l6901.

[10] A. Ballantyne, "How should we think about clinical data ownership?", *J. Med. Ethics* 46(5), 289–294 (2020).

[11] B. Kitchenham, L. Madeyski, D. Budgen, J. Keung, P. Brereton, S. Charters, S. Gibbs, and A. Pohthong, "Robust Statistical Methods for Empir. Softw. Eng.", *Empir. Softw. Eng.* 22(2), 579–630 (2017).

[12] Ch. Edwards, "Malevolent machine learning", *Commun. ACM* 62(12), 13–15 (2019).

[13] S. Greengard, "An inability to reproduce", *Commun. ACM* 62(9), 13–15 (2019).

[14] F. Pasquale, "When machine learning is facially invalid", *Commun. ACM* 61(8), 25–27 (2018).

[15] J.P.A. Ioannidis: "The mass production of redundant, misleading, and conflicted systematic reviews and meta-analyses", *Milbank Q.* 94, 485–514 (2016).

[16] J.B. Carlisle, "Data fabrication and other reasons for nonrandom sampling in 5087 randomised, controlled trials in anaesthetic and general medical journals", *Anaesthesia* 72(8), 944–952, (2017)

[17] Ch.H.J. Hartgerink, J.M. Wicherts, and M.A. van Assen, "The value of statistical tools to detect data fabrication", *Res. Ideas Outcomes* 2, e8860 (2016).

[18] Ch.H.J. Hartgerink, J.G. Voelkel, J.M. Wicherts, and Marcel A.L.M. van Assen. "Detection of Data Fabrication Using Statistical Tools", PsyArXiv, 2019, doi: 10.31234/osf.io/jkws4.

[19] N.J.L. Brown and J.A.J. Heathers, "The grim test: A simple technique detects numerous anomalies in the reporting of results in psychology", *Soc. Psychol. Personal Sci.* 8(4), 363–369 (2017).

[20] J.A. Heathers, J. Anaya, T. van der Zee, and N. Brown "Recovering data from summary statistics: Sample Parameter Reconstruction via Iterative TEchniques (SPRITE)", *PeerJ Preprints*, e26968v1 (2018). doi: 10.7287/peerj.preprints.26968v1.

[21] S. Al-Marzouki, S. Evans, T. Marshall, and I. Roberts, "Are these data real? Statistical methods for the detection of data fabrication in clinical trials", *BMJ*, 331(7511), 267–270 (2005), doi: 10.1136/bmj.331.7511.267.