Whose Commons? Data Protection as a Legal Limit of Open Science

Mark Phillips and Bartha M. Knoppers

Powerful institutions across the globe have recently joined the ranks of those making substantive commitments to "open science." For example, the European Commission and the NIH National Cancer Institute are supporting large-scale collaborations, such as the Cancer Genome Collaboratory,[1] the European Open Science Cloud,[2] and the Genomic Data Commons,[3] with the aim of making giant stores of genomic and other data readily available for analysis by researchers.[4] In the field of neuroscience, the Montreal Neurological Institute is midway through a novel five-year project through which it plans to adopt open science across the full spectrum of its research.[5] The commitment is "to make publicly available all positive and negative data by the date of first publication, to open its biobank to registered researchers and, perhaps most significantly, to withdraw its support of patenting on any direct research outputs."[6] The resources and influence of these institutions seem to be tipping the scales, transforming open science from a longstanding aspirational ideal into an existing reality.

Although open science lacks any standard, accepted definition, one widely-cited model proposed by the Austria-based advocacy effort openscienceASAP describes it by reference to six principles: open methodology, open source, open data, open access, open peer review, and open educational resources.[7] The overarching principle is "the idea that scientific knowledge of all kinds should be openly shared as early as is practical in the discovery process." This article adopts this principle as a working definition of open science, with a particular emphasis on open sharing of human data.

As noted above, many of the institutions committed to open science use the word "commons" to describe their initiatives, and the two concepts are closely related. "Medical information commons" refers to "a networked environment in which diverse sources of health, medical, and genomic information on large populations become widely shared resources."[8] Commentators explicitly link the success of information commons and progress in the research and clinical realms to open science-based design principles such as data access and transparent analysis (i.e., sharing of information about methods and other metadata together with medical or health data).[9]

But what legal, as well as ethical and social, factors will ultimately shape the contours of open science?

Mark Phillips *is an Academic Associate at the Centre of Genomics and Policy at McGill University. He is also a practicing member of the Quebec Bar Association.* **Bartha M. Knoppers, Ph.D.**, *is the Director of the Centre of Genomics and Policy at McGill University.*

The Journal of Law, Medicine & Ethics, 47 (2019): 106-111. © 2019 The Author(s)
DOI: 10.1177/1073110519840489

Should all restrictions be fought, or should some be allowed to persist, and if so, in what form? Given that a commons is not a free-for-all, in that its governing rules shape its outcomes, how might we tailor law and policy to channel open science to fulfill its highest aspirations, such as universalizing practical access to scientific knowledge and its benefits, and avoid potential pitfalls?[10] This article primarily concerns research data, although passing reference is also made to the approach to the terms under which academic publications are available, which are subject to similar debates.

We start from the perspective that the ultimate goal of both the open science movement and information commons creation is to increase practical access to scientific knowledge and its benefits across human society, and to ensure that this access is distributed as evenly as possible. The potential pitfalls of open science include exacerbating existing inequalities, by supporting the development of expensive new diagnostics and treatments that are practically available only to the stratum of the population who can afford them, while putting already-disadvantaged individuals and groups at risk of harms, such as discrimination and stigmatization. Inequities in algorithmic decision making based on big data have indeed become a widespread focus of attention and research.[11] A related risk is the *de facto* privatization of personal data, by organizing data in a manner that benefits only those who possess sufficient resources to allow them to usefully analyze them, thus transforming public funding of open science into an indirect subsidy to private industry.[12]

This article primarily concerns research data, although passing reference is also made to the approach to the terms under which academic publications are available, which are subject to similar debates.

Both of these tendencies relate to data protection as it is evolving. Although data protection frameworks have long included automated decision making and profiling within their scope, it is only with the recent surge of interest in machine learning techniques that a corresponding increase in attention is emerging in delineating what protections, if any, should exist in practice.[13] A leading understanding of data protection as a field is that it refers to a "set of legal rules that aims to protect the rights, freedoms, and interests of individuals, whose personal data are collected, stored, processed, disseminated, destroyed, etc. The ultimate objective is to ensure 'fairness in the processing of data and, to some extent, fairness in the outcomes of such processing.'"[14] In this context, recent interest in applying data protection rules to the contemporary big data and machine learning contexts should come as no surprise.

Data Protection Under the GDPR

The European Union's General Data Protection Regulation (GDPR) is now in full effect across the European Economic Area (EEA), encompassing 31 countries. It grants rights to "data subjects" (identified or identifiable natural persons), including the right to access, rectify, and object to the processing of their personal data. It also imposes duties on "data controllers" (natural or legal persons, public authorities, agencies, or other bodies that determine the purposes and means of processing "personal data," meaning information related to a data subject), for example, placing the burden on them to justify processing the personal data, to justify its transfer outside of the EEA, and to justify processing "special categories of data" (i.e., sensitive data, including data concerning health and genetic data), not only as having a lawful basis but also as falling within at least one of ten special category conditions (such as scientific research). Whereas the GDPR's precursor, the EU Data Protection Directive, had to be adopted into the national law of European Union member states, the Regulation is now directly legally applicable within the EEA (i.e., without the necessity of nation-level implementing legislation), as well as to many organizations located outside the EEA which process the personal data of individuals residing in the European Union. Despite the GDPR's aspirations to further harmonization, however, it still allows for member states to impose more stringent restrictions in certain areas it specifies, notably in the context of data the GDPR deems to be sensitive, such as data concerning health and genetic data.

Researchers' Evolving Response to the GDPR

The arrival of the European Union's General Data Protection Regulation (GDPR) set off a wave of unease in many data-intensive industries, including health research, sparking fears that it would effectively curtail their activities.[15] Although other areas of the GDPR, such as implementation of the new data subject right to data portability (which includes the right to have their personal data transmitted directly from one controller to another, where technically feasible) and the "right to be forgotten" (have a data controller erase personal data, cease further dissemination,

The Journal of Law, Medicine & Ethics, 47 (2019): 106-111. © 2019 The Author(s)

and potentially halt processing by third parties), and restrictions on data transfer to non-EU countries, have raised concerns in the research community, issues around consent have preoccupied them most, especially the specificity of consent. The GDPR's emphasis on specific consent initially alarmed translational researchers after an early draft version was released in 2016, but language favourable to broad consent when processing personal data for research purposes was ultimately incorporated in Recital 33 in the version enacted.[16] Whereas specific consent requires research participants to consent to each specific project they are participating in, and to be informed in advance of the concrete ways in which their data will be used, broad consent allows participants to consent to having their data used in multiple research projects based on a description of a broad type (or types) of research and the governance thereof. For example, participants might consent to the use of their data in any future project aimed at developing treatments for a particular disorder, facilitating sharing of data among investigators in particular fields of research and, potentially, contribution to a relevant data repository. They may also agree to future unspecified research within a field subject to proper oversight.

Although alarm in the health research sector continues to some degree, many in the field have now instead come to view the GDPR as sufficiently attuned to the nature of research, its processes and needs, and as "a well-drafted piece of legislation that raises the standards of data protection globally."[17] Indeed, consent is but one of several alternative bases upon which the final version of the GDPR allows personal data to be lawfully processed. In countries such as the UK, for example, regulators and scholars are currently actively discouraging reliance on consent to fulfill data protection duties in most cases when personal data processing is necessary for research or clinical purposes.[18] They instead suggest that it is generally preferable to rely on another legal basis, specifically that processing is being carried out for research purposes in the public interest (in the case of a public institution), or that such processing is necessary for pursuing legitimate interests (for private sector institutions).[19] This approach does not mean that research participants' consent will not be needed: indeed, irrespective of the GDPR, research ethics duties generally require this. The idea is not to rely on the consent, even though it must generally be obtained to satisfy research ethics duties, for the purpose of fulfilling GDPR obligations, and to instead rely on an alternative legal basis with respect to the GDPR. The impression of the cited writers in the UK, at least for now, is that alternative legal bases such as where processing is based on the "public interest" or the "legitimate interests" of the entity controlling the personal data mitigate the possible infringement and impact of the GDPR on open science principles.

Interpreting the GDPR

One of the difficulties in interpreting the GDPR with respect to health research is that its default lens tends to focus on relationships between private sector companies and their customers. For example, the important question of when the GDPR applies to an organization outside of the EU is determined in part based on whether that organization is "offering goods or services, irrespective of whether a payment ... is required" to people in the European Union (Article 3(2)(*a*)). The research context, including the degree of connection that a cohort of research participants would need to have to the European Union to satisfy this condition, is largely ignored by recent guidance on the interpretation of this article of the GDPR published by the European Data Protection Board.[20]

Because of this overarching lens, guidance from the European Commission on interpreting the GDPR in the health research context are to be welcomed. One recent attempt to raise awareness of ethics and data protection issues in the scientific community, which focuses on the GDPR, however, represents a missed opportunity in that a number of important questions were either ignored altogether or dealt with in too cursory a fashion. The document in question, "Ethics and data protection," was prepared at the request of and published by the Research and Innovation arm of the Commission, which oversees the Commission's Horizon 2020 research funding program.[21] The document's approach to consent is particularly notable:

> Whenever you collect personal data directly from research participants, you must seek their informed consent by means of a procedure that meets the minimum standards of the GDPR. This requires consent to be given by a clear affirmative act establishing a freely given, specific, informed and unambiguous indication of the subject's agreement to the processing of their personal data.[22]

Although this interpretation stays close to the GDPR's definition of consent in its Article 4(11), the approach appears to be at odds with the regulation in multiple other respects. First, as noted earlier, the regulation's Recital 33 makes an exception to the requirement that consent must always be specific insofar as "data subjects should be allowed to give their consent to certain areas of scientific research when in keeping with recog-

The Journal of Law, Medicine & Ethics, 47 (2019): 106-111. © 2019 The Author(s)

nised ethical standards for scientific research." It is odd for this exception to be entirely ignored here, despite being previously recognized by guidance endorsed by the Commission's European Data Protection Board.[23] Second, the GDPR allows personal data to be collected directly from data subjects on a basis other than consent (e.g. this is implicit in Article 13(1)(*d*)). Third, the distinction between GDPR consent and research ethics consent appears to be blurred. This distinction is important in that even if a basis other than consent is used to justify processing personal data with respect to the GDPR, consent will still generally be sought from research participants in order to comply with research ethics duties that apply independently. But the form of consent required in such circumstances will be defined by existing research ethics rules, not by the dictates of the GDPR.

In sum, although there is uncertainty because these novel elements of the GDPR remain to be tested (perhaps via a challenge by data subjects in real-life proceedings) the text of the GDPR appears to provide a number of additional routes through which research initiatives can satisfy their data protection obligations. The proviso is that they are attentive, as indeed they should be, to the rights and interests of those whose data they hold, and aim to adopt proportionate measures to safeguard them. These measures include ensuring the technical confidentiality of the data and integrity of the security of their systems, ensuring that the data's confidentiality will not be jeopardized in the hands of any third parties to whom it will be disclosed or transferred, and that participants' rights to access and rectify their data are ensured. Rather than continuing to wait for the courts and regulators to weigh in, an alternative to gaining clarity on some of these details would be to develop a data protection code of conduct for the health sector. As of now, BBMRI-ERIC is leading such a proposed initiative.[24] Once approved by EU data protection authorities according to a process set out in the GDPR itself, adherence to such a code would provide evidence of compliance with the GDPR, and adherence, when combined with binding and enforceable commitments to apply the appropriate safeguards, by an organization outside the EU wishing to receive personal data from an entity subject to the GDPR satisfies the regulation's restrictions on transfer.

Remaining Tensions Between Open Science and Data Protection

Despite the appearance that the GDPR strikes the proper balance between accommodating scientific research and securing individual rights and dignity, the tension between open science and data protection goes to the very core of the two movements.[25] When "open source" or "open data" have been given formal definitions, such as in the GNU Public License, these generally appear in an absolutist form and require that the information in question must be provided in its entirety and must be free to use, free to disseminate, and free to adapt with as few restrictions as feasible.

A key way in which the GDPR's restrictions are circumscribed is according to the purpose for which personal data were collected or are otherwise processed, which must be defined. For example, if researchers collect personal data, they must indicate the purpose of collection, such as to conduct a particular study. Data protection regimes have tended to prohibit any use of personal data for any purpose other than the one that was indicated at the time of collection. This contrasts sharply with the driving rationale behind the initiatives that constitute the open science movement, which instead emphasize the benefits that are possible through unforeseeable future uses to which any given information set might be put.[26] Open source software, for example, is made available to be reused for purposes that may have been unforeseen or even unforeseeable by its initial creator. This apparently fundamental tension is not entirely new: back in the 1970s and 1980s, a wave of new laws enacted around the world aimed to reconcile the protection of privacy with a new public right of access to government documents.[27] As the default position established was that government documents were presumed to be made freely available to the public unless some exception to access applied, this gave rise to a risk of violating the

Despite the appearance that the GDPR strikes the proper balance between accommodating scientific research and securing individual rights and dignity, the tension between open science and data protection goes to the very core of the two movements.

The Journal of Law, Medicine & Ethics, 47 (2019): 106-111. © 2019 The Author(s)

https://doi.org/10.1177/1073110519840489 Published online by Cambridge University Press

privacy of those whose personal information was contained in them. Through practices and frameworks established over time in the context of specific cases, the tension between these two objectives came to be workably smoothed out, often by redacting personal information, where appropriate.

The GDPR still has little or no judicial interpretation, apart from that portion of the jurisprudence of European courts, especially the European Court of Justice, regarding the previous Data Protection Directive that remains relevant to it. Further, experimentation with large-scale open science is only just beginning. As a body of legal interpretations develops, and as experience with open science increases, a similar process of incremental adjustment should be encouraged. To make the transition as smooth and steady as possible, more interplay is needed between the two, currently siloed, fields of study: advocates for open science should ensure that the courts' coming interpretations of the GDPR carefully weigh possible effects on the information commons, while also seeking to ensure that debates around open science incorporate careful consideration of the rights and concerns of data subjects and lead to steps to recognize and guarantee personal data protection.

As open science becomes institutionalized, we are in a key moment in which to establish the rules that will shape it, taking account of legitimate concerns about data protection and data sharing. Additional high-quality social science to shed light on the specific policy calibrations that will maximize open science while also giving data protection, and other extrinsic policy considerations, their due would be invaluable in this respect.

As an example of potential interplay between the two fields, the developers of a data protection code of conduct for health research, as discussed earlier, might explicitly incorporate a provision stating that the principle of open science or the medical information commons are to be viewed as important principles to guide the analysis. The general principle might be made concrete within a code of conduct in the specific context of health research, for example, by setting out best practices when establishing a data access committee, when necessary, and ensuring the approach is aimed at realizing the twin goals of data protection and open science. This approach is indeed in harmony with the underlying goal of data protection to promote the free movement of personal data so long as its processing appropriately protects and realizes the hopes of the people to whom it relates.

Conclusion

As open science becomes institutionalized, we are in a key moment in which to establish the rules that will shape it, taking account of legitimate concerns about data protection and data sharing. Additional high-quality social science to shed light on the specific policy calibrations that will maximize open science while also giving data protection, and other extrinsic policy considerations, their due would be invaluable in this respect. Even though the legislative process of the new EU General Data Protection Legislation is over, its substance will continue to rapidly evolve and take shape through the interpretations of courts and data protection authorities in specific cases and through any ancillary national legislation. The research community should watch for opportunities not only to have its voice heard in those processes, but to draw on past insights from access to public information laws to formulate interpretations that balance the promotion of self-determination for individuals with the promotion of data sharing and creation of an efficient open science information commons to support new discoveries.

Acknowledgements

The authors gratefully acknowledge funding from the Tannenbaum Open Science Institute (TOSI).

References

1. Cancer Genome Collaboratory, "Cloud Computing for BIG DATA Genomics," *available at* <https://cancercollaboratory.org> (last visited January 18, 2019).
2. European Commission, "European Open Science Cloud (EOSC)," *available at* <https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud> (last visited January 18, 2019).
3. NIH National Cancer Institute Genomic Data Commons, "The Next Generation Cancer Knowledge Network," *available at* <https://gdc.cancer.gov> (last visited January 18, 2019).
4. M. Phillips et al., "Of Clouds and Data Protection," (Forthcoming 2018); J. L. Contreras and B.M. Knoppers, "The Genomic Commons," *Annual Review of Genomics and Human Genetics* 19 (2018): 429–453.
5. E.R. Gold, "Accelerating Translational Research through Open Science: The Neuro Experiment," *PLoS Biology* 14, no. 12 (2016): e2001259, *available at* <https://doi.org/10.1371/journal.pbio.2001259> (last visited January 18, 2019).

The Journal of Law, Medicine & Ethics, 47 (2019): 106-111. © 2019 The Author(s)

6. S.E. Ali-Khan, L.W. Harris, and E. R. Gold, "Motivating Participation in Open Science by Examining Researcher Incentives," *eLife* 6 (2016): e29319, *available at* <10.7554/eLife.29319> (last visited January 18, 2019). A large part of the underlying motivation behind the endeavor rests on the hypothesis that the approach will generate further innovation and economic development in the local and regional geographic region: *see* Gold, (2016) (above). This assumption contrasts with the traditional emphasis on of "open" movements (open source, open data, etc.) on the borderlessness and placelessness of the collaborations and information exchanges they enable.
7. OpenscienceASAP website, *available at* <http://openscienceasap.org/open-science/> (last visited January 18, 2019).
8. R. Cook-Deegan and A.L. McGuire, "Moving Beyond Bermuda: Sharing Data to Build a Medical Information Commons," *Genome Research* 27, no. 8 (2017): 897–901.
9. *Id.*
10. *Id.*, at 899.
11. See e.g. V. Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor* (New York: St. Martin's Press, 2018); S. Barocas and A. D. Selbst, "Big Data's Disparate Impact," *California Law Review* 104, no. 3 (2016): 671–732.
12. A somewhat analogous real-life example, for example, occurred in Bangalore, where land titles were digitized and made publicly available, and yet the practical result was in fact to deepen inequalities, as wealthier people were the ones able to take the information provided and use that as the basis for instructions to land surveyors and lawyers and others to challenge titles, exploit gaps in titles, take advantage of mistakes in documentation, identify opportunities and targets for bribery, among others. They were able to directly translate their enhanced access to information along with their already available access to capital and professional skills into unequal contests around land titles, court actions, and offers of purchase for self–benefit and to further marginalize those already marginalized. M. Gurstein, "Open Data: Empowering the Empowered or Effective Data Use for Everyone?" (2011), *available at* <http://firstmonday.org/article/view/3316/2764> (last visited January 28, 2019). In developing a framework for open science, it would be wise to invest effort into envisioning how to avoid such pitfalls.
13. See e.g. Article 29 Data Protection Working Party, "Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679," (last revised February 6, 2018), *available at* <https://ec.europa.eu/newsroom/article29/document.cfm?action=display&doc_id=49826> (last visited January 28, 2019); L. Edwards, "Enslaving the Algorithm: From a 'Right to be Forgotten' to a 'Right to Better Decisions'?" *IEEE Security & Privacy* 16, no. 3 (2018): 46–54, doi:10.1109/MSP.2018.2701152.
14. M. Tzanou, "Data Protection as a Fundamental Right Next to Privacy? 'Reconstructing' a Not So New Right," *Data Protection Law* 3, no. 2 (2013): 88–94, at 89 (footnotes omitted), quoting the seminal work L. Bygrave, *Data Protection Law: Approaching Its Rationale, Logic, and Limits* (The Hague: Kluwer Law International, 2002): at 168.
15. E.S. Dove, D. Townend, and B.M. Knoppers, "Data Protection and Consent to Biomedical Research: A Step Forward?" *The Lancet* 384, no. 9946 (2014): P855, *available at* <https://doi.org/10.1016/S0140-6736(14)61488-4> (last visited January 28, 2019).
16. *Id.*
17. E.S. Dove, "The EU General Data Protection Regulation: Implications for International Scientific Research in the Digital Era," *Journal of Law, Medicine & Ethics* 46, no. 4 (2018): 1013-1030.
18. M.J. Taylor, S.E. Wallace, and M. Prictor, "United Kingdom: Transfers of Genomic Data to Third Countries," *Human Genetics* 137, no. 8 (2018): 637–645 at 638.
19. *Id.*, at 639.
20. European Data Protection Board, "Guidelines 3/2018 on the territorial scope of the GDPR (Article 3) – Version for public consultation: Adopted on 16 November 2018," *available at* <https://edpb.europa.eu/sites/edpb/files/consultation/edpb_guidelines_3_2018_territorial_scope_en.pdf> (last visited January 28, 2019).
21. European Commission, "Ethics and Data Protection" (November 14, 2018), *available at* <http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/ethics/h2020_hi_ethics-data-protection_en.pdf> (last visited January 28, 2019).
22. *Id.*, at 10–11.
23. The EDPB endorsed, with respect to the GDPR, the following guidance published by its predecessor: Article 29 Data Protection Working Party, "Guidelines on consent under Regulation 2016/679," at 28–30, *available at* <https://ec.europa.eu/newsroom/article29/document.cfm?action=display&doc_id=51030> (last visited January 28, 2019).
24. J.-E. Litton, "We Must Urgently Clarify Data-Sharing Rules," *Nature* 541, no. 7638 (2017): 437.
25. See E. S. Dove, "Reflections on the Concept of Open Data," *Scripted* 12, no. 2 (2015): 154–166.
26. J. Boyle, "5 Mertonianism Unbound? Imagining Free, Decentralized Access to Most Cultural and Scientific Material," in C. Hess and E. Ostrom, eds., *Understanding Knowledge as a Commons: From Theory to Practice* (MIT Press, 2007): at 123-144.
27. NIH-DOE Joint Subcommittee, "NIH-DOE guidelines for access to mapping and sequencing data and material resources," (adopted December 7, 1992), *available at* <http://www.genome.gov/10000925> (last visited January 28, 2019).

The Journal of Law, Medicine & Ethics, 47 (2019): 106-111. © 2019 The Author(s)