**Governing Human-Derived Data**

By Teresa Scassa

Platforms harvest vast quantities of user data for a variety of commercial and operational purposes, including driving targeted advertising programs. These data may also be made available to third parties. Platform data are also scraped by a [variety of actors](https://doi.org/10.1108/OIR-02-2018-0053), including researchers, civil society actors, commercial competitors and data brokers. One notorious example of the scraping of personal information from platforms comes from [Clearview AI’s scraping](https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html) of online photographs to build its massive facial recognition database.

Perhaps the most highly sought-after platform data are those derived from humans and their activities. Issues around the legitimacy of collecting, using or sharing these data are often intertwined with considerations about whether the data are about identifiable individuals or are either deidentified at source or subsequently anonymized. This brief will argue that questions of appropriate data collection/use – in the platform context and elsewhere – can no longer be sufficiently addressed by asking whether the data are about an identifiable individual (i.e., whether they are personal data) or whether they are anonymized. While “personal data” and “anonymized data” remain important classifications under data protection law, we also need a new concept of “human-derived data” with a distinct governance framework.

Data protection (privacy) laws typically govern information about an *identifiable individual*. This is because such laws are built upon [privacy principles](https://www.oecd-ilibrary.org/science-and-technology/oecd-guidelines-on-the-protection-of-privacy-and-transborder-flows-of-personal-data_9789264196391-en) that protect the rights of individuals to control information about themselves. In such a framework, if information cannot be linked to a particular individual, then that individual’s privacy rights cannot be impacted by its processing. As a result, in a context in which data have become the highly sought-after fuel for data-driven innovation, including artificial intelligence (AI), it is unsurprising that there is considerable pressure to distinguish between personal data and anonymized data, and to set anonymized data outside the scope of data protection laws.

The distinction between personal and anonymized data frees up anonymized data for more widespread use. This is the normative position adopted in the EU’s [*General Data Protection Regulation*](https://gdpr-info.eu/) (GDPR). It is also evident in the *Consumer Privacy Protection Act* portion of Canada’s [Bill C-27](https://www.parl.ca/DocumentViewer/en/44-1/bill/C-27/first-reading), and in legislation such as Ontario’s [*Personal Health Information Protection Act*](https://www.canlii.org/en/on/laws/stat/so-2004-c-3-sch-a/197771/so-2004-c-3-sch-a.html). Typically, personal data are subject to regulation and governance; anonymized data are available for use without knowledge or consent. Any oversight of the use of anonymized data will relate to the process of anonymization (since if it is not done properly, the data can be linked to an identifiable individual and thus remain personal data). Increasingly, data protection laws also provide penalties for deliberate reidentification of anonymized data.

There are problems with this approach to anonymized data. First, a growing number of [scholars](http://www.uclalawreview.org/pdf/57-6-3.pdf) and privacy advocates have warned that with contemporary volumes of data and data analytics tools, it will always be possible to [reidentify individuals](https://www.nature.com/articles/s41467-019-10933-3/) from datasets, making anonymization a chimera. Yet, that is not necessarily a problem under data protection laws, which often define anonymization in relative terms. Thus, the issue of reidentification requires consideration of a [variety of factors](http://www.teresascassa.ca/index.php?option=com_k2&view=item&id=368:court-decision-explores-reidentification-risk-in-access-to-information-request&Itemid=80), including the sensitivity of the data, what other relevant data are available that might lead to reidentification, and how likely it is that an adversary will seek to reidentify one or more individuals from the anonymized dataset. One difficulty is that reidentification risk for anonymized datasets may change over time, as more and more data become available and as new analytical tools are developed.

Second, some [scholars](https://link.springer.com/chapter/10.1007/978-3-319-46608-8_1) argue that we need a concept of group or collective privacy that recognizes potential group interests in data that have been collected from humans, even if those data are anonymized. However, contemporary data protection laws do not recognize group privacy, even though these arguments have been taken up not just by scholars, but by privacy advocates in a range of contexts. Debates over AI governance also raise concerns that even anonymized data can adversely impact individuals and/or groups. The *Artificial Intelligence and Data Act* component of Canada’s Bill C-27, for example, would create stewardship obligations for anonymized data, particularly in relation to their potential to lead to biased AI output.

The distinction between personal data and anonymized data leaves an important governance gap. To address this gap, we need to recognize a new category of data which I call “human-derived data”. Human-derived data are data derived from humans or their activities and that are not personal data. While the normative basis for privacy law is the autonomy and dignity of individuals, the normative basis for the governance of human-derived data is fundamental human rights. Individuals and/or the communities/groups they belong to are the source of these data; these data can be used in ways that harm or exploit the collective or its members, and they require some form of governance to ensure that they are not used in harmful, discriminatory or rights-limiting ways.

Why is privacy not a sufficient normative basis for governance of human-derived data? One reason is that privacy law is premised on the individual and their right to control their personal data, and through their data, their [identity](https://books.google.ca/books/about/Monitoring_Laws.html?id=w-y2DwAAQBAJ&redir_esc=y). The group privacy concept addresses more communal rights in data and often advances the concerns of equity-deserving groups. Because of the potential for human-derived data to impact decisions made about both groups and individuals in ways that go beyond privacy, governing human derived data can address human rights other than privacy, such as the right to be free from discrimination.

A focus on human-derived data (rather than anonymized personal data) puts the *human* and not the *individual* at the heart of the analysis. It is a more explicitly human rights-based focus. Adding governance of anonymized personal data is insufficient. Although some human-derived data may begin as personal data prior to anonymization, other human-derived data are collected in contexts where they are never linked to identifiable individuals. For example, data about the presence of the COVID-19 virus and its variants in [wastewater](https://techreg.org/article/view/11192) is not collected in ways that lead to the identification of specific individuals. It is therefore never personal data, but it is human-derived data. Anonymized data are thus only a subset of the category of human-derived data.

Because it is distinct from the individual who is at the normative heart of privacy law, the governance of human-derived data can embrace a greater range of considerations. These include broader human rights concerns (such as the right to be free from discrimination) as well as ethical principles. The collection and use of human-derived data, for example, might require transparency and public engagement. It could also require open access to results of analyses or research, or the return of direct or indirect (although not remote) benefits to the community. These concepts have already emerged in discussions about [data or knowledge commons](https://global.oup.com/academic/product/governing-knowledge-commons-9780190225827?cc=ca&lang=en&), in [citizen science ethics](https://theoryandpractice.citizenscienceassociation.org/articles/10.5334/cstp.235/), as well as in [access-benefit sharing frameworks](https://www.cbd.int/business/bc/ABS.shtml#:~:text=Parties%2C%20indigenous%20people%20and%20local%20communities%2C%20businesses%20and,clarity%20and%20legal%20certainty%20of%20access%20and%20benefit-sharing.) for the use of genetic resources. The factors relevant to determining appropriate governance for human-derived data may depend on context. For example, the governance of human-derived data may consider the nature or composition of the communities from which such data are gathered, and the relationship of the data gathering to public or private infrastructure.

An argument for the governance of human-derived data is not an argument against data protection law (which remains necessary) nor is it an argument against the governance of anonymized data – particularly regarding protections to ensure it remains anonymous. Rather, it is an argument that the ubiquitous collection of data about human activities and its growing use to drive decision-making about communities and individuals across sectors and spheres of activities require an appropriate framework to ensure transparency and engagement and to protect human rights.

*Teresa Scassa is a Professor and Canada Research Chair in Information Law and Policy at the University of Ottawa. Many thanks to the participants of the February 2023 workshop on platform governance for their thoughtful comments on an earlier draft of this piece.*