View in browser

Global Alliance for Genomics & Health

GDPR Brief

Data Protection Implications of Publishing Metadata to Enable Discovery

June 2022

Adrian Thorogood

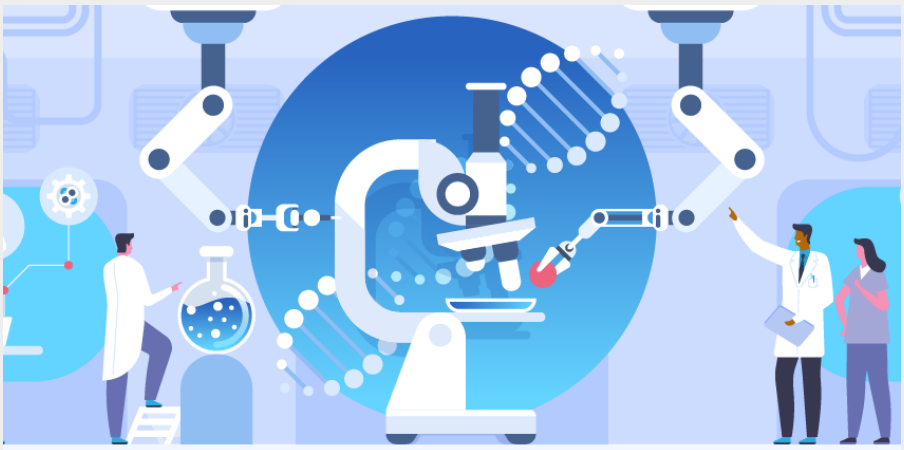


Image by Ico Maker via AdobeStock

Metadata is data that describes data or another resource (e.g., a biospecimen), and is a key part of what ensures data are findable, accessible, interoperable and re-useable (FAIR). Metadata may in some circumstances be considered personal data – sometimes to the surprise of organizations seeking to process or publish them. What counts as metadata, what types of metadata are needed to realize FAIRness, and whether or not metadata need to be published depend on the context.

Some metadata models describe datasets (i.e., collections or cohorts) as part of data catalogues. Data catalogues are typically published to enable the research community to discover relevant datasets for their projects. Dataset metadata may include the following:

- information about the nature of the data (e.g., documentation describing data models, dictionaries, and standards) and data quality (e.g., completeness of data fields).
- administrative information concerning the organization holding the dataset, and applicable access and use conditions.
- statistical descriptions of the dataset, such as the mean and distribution of different data fields.

Organizations publishing dataset metadata should keep a look out for certain risks of privacy breach. For example, some administrative metadata is personal data–such as

the names, email addresses, and affiliations of the responsible Principal Investigators. Permission may be required to publish these data. Statistical descriptions of datasets may leak personal data, e.g., where the datasets are small (e.g., in rare disease contexts) or incrementally updated. Indeed, where a small number of individuals are added to a dataset, the difference between the new statistical description and the old one may leak personal data. Publishing genomic summary statistics might also be subject to attacks that reveal if an individual is a member of the dataset (in turn potentially revealing sensitive attributes like disease status). Keep in mind also that *generating* statistics from underlying health and genetic data is a form of processing that requires adherence to applicable data protection regulations.

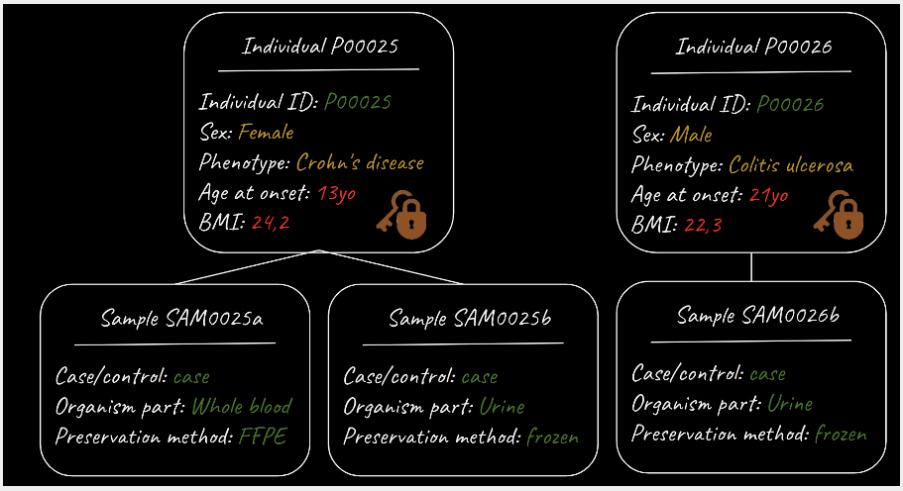
Bioinformaticians also commonly use the term metadata for data that describes *individual* research participants (e.g., donor ID, age, gender, primary phenotype), as well as samples extracted from an individual (e.g., tissue type, preparation), and sequencing experiments run on these samples. Individual-level metadata enable the comparison, linkage, and re-use of sequence data. A handful of individual-level metadata fields may also be published to facilitate data discovery, such as each subject's age, gender, and phenotype. The label "metadata" should not, however, distract from the fact that these data are in essence individual-level demographic and health data (albeit usually low-dimensional data).

As with any human data, controllers must assess the identifiability of metadata before proceeding with processing or publication. In doing so they must consider all means reasonably likely to be used by any person to re-identify the individual and objective factors including the costs and time required, available technology at the time of the processing and technological developments. A contextual, risk-based approach is recommended for identifiability assessments. The need to enable discovery by the research community offers a justification for publishing metadata, though publication also limits available safeguards and reduces contextual certainty over identifiability. Aggregate data or a handful of individual-level data fields are typically sufficient to enable data discovery. A data protection impact assessment can be performed to demonstrate that such data poses lower risks of both re-identification and disclosure of sensitive information than the underlying data. It is often possible to mathematically demonstrate a low or nil risk of re-identification for such data using frameworks like k-anonymity. Release of metadata can also be controlled through a Beacon-type query interface, limiting the amount of metadata released to any particular requestor. Differential privacy controls can be added to these interfaces that protect against a requestor making multiple different queries in the aim of re-identification. Finally, layered approaches can be adopted to balance discovery and protection, e.g., by releasing a subset of metadata publicly, while limiting access to richer metadata to registered requesters.

Data archives can also play a role in balancing discovery and protection. Data submission agreements should clarify who is responsible for ensuring the anonymity of metadata that will be published in a data catalogue. Data protection by design can inform the structure of metadata models. Metadata field descriptions should be clear and focused, and potentially constrained by controlled vocabularies, to avoid the risk of inadvertent entry and release of personal data. Data models can also intuitively segregate sensitive (e.g., individual disease status) and non-sensitive (e.g., sample type) attributes (see Figure 1).

From a policy perspective, it is important not to hyperfocus on the data protection risks of processing or publishing metadata. Metadata not only facilitates scientific activity; it also supports both data protection and sustainability. Ecosystems where researchers can effectively discover, access, and re-use *relevant* genomic and health data are both more ethical – avoiding unnecessary disclosures of *irrelevant* data – and more efficient – saving both researchers and data providers time and resources discovering, accessing, and re-using data.

Figure 1. Segregation of Sensitive and Non-sensitive Metadata



This figure shows a metadata model with fields describing individuals, as well as additional fields describing one or more samples taken from a given individual. The metadata model is nested – each sample can equally be described by the associated individual metadata. The metadata model also segregates "sensitive" individual level metadata (which is kept secure), from non-sensitive sample-level metadata. The figure is provided courtesy of Coline Thomas, European Molecular Biology Laboratory - European Bioinformatics Institute (EMBL-EBI).

View the brief online >

Adrian Thorogood is a legal researcher at the Luxembourg Centre for Systems Biomedicine, University of Luxembourg

---

Please note that GDPR Briefs neither constitute nor should be relied upon as legal advice. Briefs represent a consensus position among Forum Members regarding the current understanding of the GDPR and its implications for genomic and health-related research. As such, they are no substitute for legal advice from a licensed practitioner in your jurisdiction.

References

- Article 29 Working Party. Opinion 05/2014 on Anonymisation Techniques.
- Alexander Bernier, Hanshi Liu, and Bartha Maria Knoppers. "Computational Tools for Genomic Data De-identification: Facilitating Data

Protection Law Compliance." *Nature Communications* (2021).
- Alexander Bernier and Bartha Knoppers. "Biomedical Data Identifiability in Canada and the European Union: From Risk Qualification to Risk Quantification?" *SCRIPTed* 18 (2021).
- Groos, Daniel and Evert-Ben van Veen. "Anonymised Data and the Rule of Law" *European Data Protection Law Review* (2020).
- Freeberg, Mallory Ann Freeberg et al. "The European Genome-Phenome Archive in 2021" *Nucleic Acids Research* (2022).
- Hannes Ulrich et al. "Understanding the Nature of Metadata: Systematic Review" *Journal of Medical Internet Research* (2022).
- Mark D Wilkinson et al. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data* (2016).
- European Commission, Proposal for a Regulation - The European Health Data Space (2022).
- European Open Science Cloud, Glossary.
- GA4GH Beacon v2 Resources.

Relevant GDPR Provisions

- Article 4(1) – definition of personal data
- Recital 26 – GDPR not applicable to anonymous data

Past Briefs

Federated Analysis for responsible data sharing under the GDPR

April 2022



What constitutes an 'international data transfer?' Recent developments

February 2022



For a list of all previous briefs, please consult here.

Global Alliance for Genomics and Health, 661 University Avenue, Suite 510, Toronto, ON M5G 0A3, Canada

Unsubscribe Manage preferences