Wellcome Sanger Institute

Data Sharing Policy

Policy

January 2025 v3

The Wellcome Sanger Institute is dedicated to advancing genetic and genomic science for the benefit of all. Open data sharing (i.e., making research data available for others) strategically supports our mission and is a cornerstone of an open research culture that values transparency, reproducibility, equity and collaboration.[1]

1. The Institute expects its researchers to maximise the availability and usefulness of the data they generate by adhering to the FAIR (Findable, Accessible, Interoperable, Reusable) principles.[2] This ensures that data are easily discoverable, well-described and can be re-used in diverse contexts.

- All data sets underpinning a research article should be shared in research repositories with a global unique persistent identifier (such as a Digital Object Identifier [DOI] or accession number) by the time of publication at the very latest. Sharing potentially useful data sets that are not part of a manuscript is strongly encouraged.

- Data sets should have good documentation and rich associated metadata that follows community standards. As a general rule, someone who is not familiar with the data should be able to understand what it is about and what files are present using only the metadata and documentation provided. Any limitations on the use of data should be made explicit.

- Sharing intermediate or processed data is strongly encouraged to facilitate data re-use. Commonly used (non-proprietary) data formats should be selected for sharing.

2. Researchers must always protect the privacy and confidentiality of research participants. Human data sets should usually be managed access and submitted to the European Genome-phenome Archive (EGA), although this may vary according to the consent given by participants. In case of any ambiguity or if further guidance is needed, please contact the Legal and Research Governance team.

3. The Institute follows the general principle that data should be 'as open as possible, as closed as necessary'. Data may be kept closed to protect intellectual property and aid the translation of insights into impact; to prevent harm; to meet other legal or ethical requirements; or when open data sharing requires a disproportionate amount of effort and resources.

4. The Institute recognises the need for researchers to be appropriately credited for their investment in data generation. It is expected that all researchers acknowledge the contributions of others by citing the relevant data and publications. We encourage our researchers to use ORCID identifiers, which allows them to claim and link their contributions. As a signatory of the San Francisco Declaration on Research Assessment (DORA), the Institute values a wide range of research outputs when assessing researchers, including the sharing of data sets.

[1] Wellcome Sanger Institute position on Open Science

[2] FAIR principles

Implementation Guidelines

January 2025 v16

These guidelines provide practical guidance for implementing the Data Sharing Policy.

Please contact datasharing@sanger.ac.uk if you have any questions

A. Metadata submission

Data generators should annotate their metadata as fully as possible, and not only submit the bare minimum information to research repositories (recognising that there can be limits to the amount of metadata that can realistically be submitted). It is best practice to use controlled ontologies where they exist; an extensive list of ontologies can be found in the EMBL-EBI Ontology Lookup service https://www.ebi.ac.uk/ols4.

Metadata should also include detailed methods information, links to publication or other resources, and any documentation that allows other researchers to assess whether they should re-use the data for their purposes.

For all human data that is submitted to the EGA (i.e., managed data access studies), the Institute should comply with the following minimum metadata deposition: Gender; Donor/subject ID; Phenotype (e.g., type of cancer sample; please use the Experimental Factor Ontology http://www.ebi.ac.uk/efo/).

For all data submitted to the ENA, the Institute should comply with established minimal metadata deposition, which is dependent on the sample; for further information please check http://www.ebi.ac.uk/ena/submit/mixs-checklists#MIxS_share. The ENA default checklist includes collection data and geographic location: https://www.ebi.ac.uk/ena/browser/view/ERC000011.

B. Data types and repositories

To aid discoverability, the following data types should be submitted to repositories:

DNA Sequencing data – The Sanger Institute will, where appropriate, release the following sequencing file types to ENA or EGA: raw CRAMs or BAMs; aligned CRAMs or BAMs for whole-genome sequencing or exome sequencing; improved BAMs where applicable, VCF files where applicable.

Genotyping data – The Sanger Institute will, where appropriate, release the following genotype file types to EMBL-EBI repositories: iDAT files and Ped/map files.

Reference Genomes – Where a genome sequence/assembly is the first for that species, or is intended to be the *de facto* reference genome for a species, the Sanger Institute will release that sequence(s) as quickly as possible, preferably to the ENA but at a minimum to an FTP site.

Functional genomics data – are data obtained from microarray or high-throughput sequencing studies to describe gene expression or genomic properties, functions and interactions (e.g. RNA-seq, ChIP-seq, ATAC-seq, multi-omic studies). Primary data sets that are of use to the research community should be submitted to a research repository as soon as possible after generation, and definitely by publication.

Other Biological/Biochemical Assay Data – such as mass spectrometry data, protein interaction data or imaging data should also be shared in suitable databases as soon as possible after generation, and definitely by publication (e.g. PRIDE for mass spectrometry data, IntAct for protein interaction data, BioImage Archive for imaging data). When imaging data is tied to sequencing data they should be released at the same time. If no community/discipline-specific repositories for the data exist, they should be shared in a general repository (e.g. Dryad, Figshare, Zenodo).

GWAS Summary Statistics – Summary statistics for GWAS should be made openly available and preferably deposited in the GWAS Catalog (https://www.ebi.ac.uk/gwas/), which uses the GWAS-SSF standard, no later than the time of publication. Submission of polygenic risk scores to the PGS Catalog (https://www.pgscatalog.org/about/#submission) is encouraged.

Other Processed data – Sharing processed data in discipline-specific or general repositories, or dedicated websites/portals to aid data re-use is strongly encouraged.

C. Timing strategy for data submission and release

The Institute operates core pipelines to deposit sequencing data into EMBL-EBI repositories. DNA sequencing and functional genomic (e.g. transcriptomic) data will be submitted to ENA or EGA 12 months after data generation (from the point that data is archived in the file store iRODS). The process of data release differs between ENA and EGA. For ENA, data release occurs immediately and automatically after submission (at 12 months). For EGA, submission of data to EGA does not automatically result in its release, so researchers need to request dataset generation by emailing the Data Release Team (data-release@sanger.ac.uk), ideally as soon as all data underlying the work have been submitted to EGA and definitely before the time of publication.

It is strongly recommended that all data underpinning a study are released by the time of preprint posting (if applicable). If researchers wish to delay the release of their data to a time closer to journal publication – e.g. in the case of PhD studies, capacity building studies, intellectual property protection, or if more time is required to ensure data is FAIR – they can do so by actioning a delay in SequenceScape for data submission (ENA) or by delaying the request for dataset creation (EGA).

Where there is a significant public health benefit to sharing information rapidly, quality-assured interim and final data should be shared as rapidly and widely as possible, and in advance of journal publication; Heads of Programmes have the authority to determine what qualifies as a health crisis where Sanger should share data rapidly. For large-scale resource projects that span a long period of time, it is expected that data is shared continuously, rather than only at the time of publication.

The Data Release Team (DRT) can submit genotyping data as well as processed data (e.g. VCFs) to repositories upon request.

The following table provides an overview of the timelines for data submission and release to repositories.

Data Type	Data Submission	Data Release
Reference Genomes	As soon as possible or 12 months after data generation	Upon data submission
Non-sensitive DNA sequencing data	12 months after data generation (but can be delayed if necessary)	Upon data submission (12 months after data generation or by publication the very latest)
Sensitive DNA sequencing data (EGA)	12 months after data generation	Upon request to DRT; as soon as possible following submission but can be delayed until publication if necessary
Functional genomics data	12 months after data generation (can be delayed if necessary)	Can be delayed until publication if necessary
Genotyping data	Upon request to DRT; as soon as possible	Upon data submission (as soon as possible and no later than publication)
Processed data (e.g. GWAS summary statistics, VCFs)	Upon request to DRT; no later than publication	Upon data submission (no later than publication)

Exemptions to Data Release

Exemptions requiring permission

Where researchers wish to exempt potentially useful data from release, they should reach out to the Data Access Committee (DAC) datasharing@sanger.ac.uk . In these cases, no data release should be prevented without permission from the DAC, who will provide an approval number if permission is granted. Exemptions from data release will be considered by the DAC on a case-by-case basis and may be granted for the following:

- **Sensitive studies**: Studies where data are socio-politically sensitive or there is a significant risk of harm (particularly to individual participants if they are re-identified, even where the risk of re-identification itself is low) can be made exempt from data release.
- **Biosecurity**: Where the release of data, in particular from pathogens, could lead to potential misuse or presents a potential (bio)security threat, data can be exempt from data release.
- **Intellectual Property (IP)**: Data may be exempt from sharing if this is necessary to protect IP rights, including when this is necessary to optimise translation.
- **Other**: If you have any other reasons, please contact datasharing@sanger.ac.uk.

Exemptions NOT requiring permission

Data produced as part of optimisation and testing studies, or data that have no value for other researchers do not need to be released. Data from replication studies to validate initial findings also do not need to be submitted to repositories.

D. Collaborations

The Sanger Institute researchers are responsible for ensuring that collaborations respect the Institute's Data Sharing Policy and share data/results in a FAIR manner and timely fashion. Sanger researchers respect their collaborators and will not share data with other parties prior to depositing data to research repositories, unless otherwise agreed upon with all relevant parties. For collaborations between Sanger and partners based in low-and-middle-income countries (LMICs), the Institute strives to ensure principles of mutual benefit and equity are embedded in partnerships. When dealing with Indigenous data, researchers should follow the CARE principles for Indigenous Data Governance.

E. De-identification

In all but exceptional circumstances, the data that Sanger receives from collaborators and the research data that Sanger generates should never contain personal data that could lead to the identification of the research subject. We therefore require our collaborators to remove all identifiable information from the data they send us and to never share with us the "key" or "link" that allows the research subject to be identified.

Prior to receiving any data from collaborators or generating research datasets, Sanger researchers should always assess whether this data, either on its own or in combination with other publicly available information, could identify the research subject. Where it can, the Sanger researcher should first speak to a member of the Sanger Legal Team, sending the query to dataprotection@sanger.ac.uk.

F. Additional guidance and references

FAIR principles

CARE principles

Wellcome Sanger Institute position on Open Science

Embedding Equity in International Research Collaborations Guidance