

Wellcome Trust Sanger Institute
DATA SHARING GUIDELINES
July 2010

Rapid and open data release has been a core component of WTSI's strategy and success. It maximises the value and scientific impact of the data, and ensures transparency and equity in exploitation of the opportunities created.¹

These guidelines develop the terms of the Wellcome Trust Sanger Institute (WTSI) data sharing policy² and are intended to provide researchers with practical guidance for implementing the policy.

The policy deals principally with *pre-publication* data release. It is expected that by publication all data will have been shared and will meet the standards specified in these guidelines.

These guidelines will be kept under review and further developed by the WTSI Data Sharing Committee. Please contact Stephanie Dyke (WTSI Policy Adviser, sd4@sanger.ac.uk) if you have any questions about the data sharing policy and guidelines.

Table of Contents

1. General	2
2. Managed Access	4
3. Data Types	7
4. Researchers' Rights and Responsibilities	10
5. Development of IP	11
6. Collaborations	11
-- List of Abbreviations and Glossary	12

¹ Adapted from the *WTSI Strategic Plan 2006-2011*

² WTSI's data sharing policy is available at:

http://www.sanger.ac.uk/datasharing/docs/wtsi_datasharing_policy.pdf

Access (1)

The Institute aims to provide rapid access to data sets of use to the research community and will place these in publicly accessible repositories when possible.

1. General

1.1 Access

1.1.1 Data generated at the Institute will be shared either publicly or *via* a managed access procedure, when this is necessary to protect confidentiality and the privacy of research participants, or to respect the terms of their consent.

1.1.2 There are three stages of data sharing:

- Primary data³ should be shared immediately after basic quality control (QC)⁴ and users clearly notified of the level of QC which has taken place.⁵ The submission specifications of the major public repositories⁶ should be followed to ensure that sufficient metadata is provided.
- From then on, the results of preliminary analyses of use to the research community (see 3.1.2 for examples) should be regularly released *via* WTSI or other suitable databases.
- Final data sets and the final results of analysis should be submitted to appropriate reference databases⁷ (usually just prior to publication). Choosing appropriate data and integration standards, and including all relevant metadata, are key to enabling further research on data. To be clear: all of the data (including covariates, e.g., anonymised phenotypic and demographic data) necessary to replicate the research study as well as the analyses performed should be shared.

³ "Primary data" refers to the lowest level of data that is considered useful to archive in public primary data repositories. Over time, what is considered appropriate by the community is likely to evolve.

⁴ To facilitate the implementation of this policy, the release of high throughput data should make use of automatic quality control pipelines to the greatest extent possible.

⁵ For an overview of where and by when primary data sets should be deposited, please see Table 1 - *Summary table for submission of open and managed primary data*, p9.

⁶ "Major public repositories" refers to stable, archival databases of biological information which provide accession numbers for submissions, are considered public resources, accept and provide data freely to all users. These are often run by either the EBI or the NCBI.

⁷ "Reference databases" refers to databases that are recognised by the community as reference sources of data of a particular type in a particular field and have as a core part of their mission collecting, preserving and disseminating this information. These may include the databases of the major public repositories (see definition in footnote 6 above) as well as databases such as COSMIC.

- 1.1.3 Care must be taken to ensure effective anonymisation of research data that relates to research participants.
- 1.1.4 Data sets should be labeled with standard identifiers wherever possible so that users can easily identify and compare separate analyses of samples.
- 1.1.5 Exceptions to the “immediacy” of the data sharing policy may be justified when it is necessary to seek intellectual property (IP) protection to ensure health benefits occur. Please see the guidelines in section 5 to delay data release on these grounds.
- 1.1.6 Every effort should be made to share data as widely and effectively as possible. Data and interoperability standards are outlined in section 3 of these guidelines. In meeting the requirements of the policy, researchers should link associated data sets deposited in various databases and link data sets to any publications based on the data.⁸ Researchers should also consider adding further documentation to enhance the long-term value of the data (e.g., results of unpublished analyses, etc.).

Ethical considerations must be taken into account when determining whether data can be made publicly accessible. Data from research participants will generally be under managed access (see 1.1.1).

- To share data with researchers *via* a managed access procedure, please follow the guidelines in sections 2-4.
- For data that can be made publicly accessible, please skip section 2 and follow the guidelines in sections 3 and 4.

⁸ NB: Research articles should be submitted to UKPMC within 6 months (maximum) of publication.

Ethical Considerations

Conducting genetic and genomic research carries responsibilities to protect confidentiality and the privacy of research participants. Access to certain data sets will therefore be carefully managed and granted in a transparent manner to all appropriately qualified researchers.

2. Managed Access

2.1 Providing access to the research community

- 2.1.1 Managed access data⁹ should be submitted to the European Genome-phenome Archive (EGA at EMBL-EBI) to be shared with the research community under the terms of a data access agreement based on WTSI's Research Community Access Policy (see 2.1.5).
- 2.1.2 Managed access data should be shared within the same timeframe and should meet the same standards as data that is publicly released, i.e., all relevant specifications in section 3 apply to managed access data sets.
- 2.1.3 Analyses yielding data which can be made publicly accessible should be submitted to the appropriate database(s).
- 2.1.4 Managed access data should remain open to the research community and the procedure for granting access should be transparent (i.e., clearly established and published on the project and/or database website, with reasons for any refusal of access to be made explicit).
- 2.1.5 Different projects may have different requirements and access provisions may need to be agreed with collaborators providing samples and the relevant ethics bodies for the study, but, insofar as is possible, these should be based on WTSI's research community access policy (see box below).¹⁰

⁹ "Managed access data" refers to data that cannot be made publicly accessible and is therefore shared with the research community *via* a carefully managed access procedure.

¹⁰ For retrospective studies: If there is any concern that sharing resulting data *via* a managed access procedure may be unlawful, the relevant ethics bodies for the study should be contacted for advice. For prospective studies: data sharing plans should be outlined in participant information leaflets and in any applications to relevant ethics bodies.

WTSI Research Community Access Policy

WTSI will share managed access data with all appropriately qualified researchers¹¹ from academia, charitable organisations and private companies, such as drug companies, whether based in the UK or abroad. Researchers accessing data will have to agree to the following conditions:

Conditions of Access

- 1- Agree not to re-identify data - including any eventual re-identification made possible by combining data sets - to preserve data confidentiality and protect research participants' privacy
- 2- Agree not to pass on data to others and to take appropriate security measures to protect data confidentiality
- 3- Respect WTSI's code of access (see box in section 4) and any agreements in line with the Ft Lauderdale data sharing principles, such as "first publication rights"¹²
- 4- Respect conditions relating to the Wellcome Trust (WT) IP policy
- 5- Acknowledge WTSI and any collaborators as data providers

2.1.6 The policy of open sharing with the research community may be restricted to protect research participants or to respect the terms of consent and other conditions of use.

2.1.7 Access to these data sets by the police or other law enforcement agencies will be acceded to only under court order, and WTSI will resist such access vigorously in all circumstances.

2.1.8 Fees may only be requested to recover the costs of making data available.

¹¹ We define an appropriately qualified researcher either as someone who has authored a relevant peer-reviewed article that we can locate on [PubMed](#), and who is still working in the field, or as a successful applicant to a relevant data access committee. We reserve the right to request further information.

¹² *Sharing Data from Large-scale Biological Research Projects: A System of Tripartite Responsibility*, Report of a meeting organized by the Wellcome Trust and held on 14–15 January 2003 at Fort Lauderdale, USA.

2.2 Procedure

- 2.2.1 The project's access policy should be based on the WTSI Research Community Access Policy (see box in 2.1.5) and any modifications to this template must be approved by the WTSI Data Sharing Committee.
- 2.2.2 The project's access policy should be published on the project (and database, if possible) web-page (see 2.1.4), and the procedure for application clearly explained.
- 2.2.3 The data access agreement used should be based on the 5 principles outlined under Conditions of Access in the WTSI Research Community Access Policy (see box in 2.1.5) and approved by the WTSI Data Sharing Committee.
- 2.2.4 A record should be kept of access requests, to include the following:
- data set(s) requested (ID)
 - date of application
 - minimum applicant information:
 - name, contact e-mail/address, and institutional affiliation
 - relevant publications
 - whether the applicant was approved or not
 - date of approval/denial of access
 - reason(s) for denial of access
 - details of any complaints about the application procedure

Please note that the WTSI Data Sharing Committee reserves the right to inspect these records.

Access (2)

The Institute will support data and interoperability standards to maximise access and ensure ease of integration with other global resources.

3. Data Types

Data type-specific information and requirements are provided in the following sections.

3.1 Genomic Sequencing (whole genome, exome, etc.)

3.1.1 Primary sequencing data should be submitted to the European Nucleotide Archive (ENA at EMBL-EBI) within 2 months (maximum) of generation.

Standards (Sequence Read Archive): *raw sequence with qualities (fastq) and traces (SRF or BAM)*
Time limit: *2 months*

3.1.2 Preliminary analyses may include alignments, assembled contigs, data charting the progress of assemblies (e.g., maps of clones), and analyses of genetic variation. Genome assemblies will be submitted to major public repositories within one year of completion¹³ even if they have not been published by then.

Standards (for next generation data): *alignments (BAM), variation (VCF)*

3.1.3 Final sequence data (i.e., refined to the maximum extent intended) should be submitted to EMBL-EBI and final sequence analysis data (EST, STS, SNP, verified mutations, CNV, etc.) should be submitted to appropriate reference databases (dbEST, UniSTS, dbSNP, COSMIC, etc.). Final sequence entries should contain any annotation generated during the course of the study. Where appropriate, tracks for standard browsers should be put up.

Standards/Metadata: *defined by public repository submission requirements*

3.2 Genotyping and Cytogenetics

3.2.1 Primary genotyping and cytogenetics data should be submitted to the European Genome-phenome Archive (EGA at EMBL-EBI) within 3 months (maximum) of generation.

Time limit: *3 months*

3.2.2 The final results of analysis should be submitted to dbSNP, etc.

¹³ "Completion" refers to the time at which internal analysis begins.

3.3 Functional Analysis Assays

(Transcriptomics, ChIP-seq, other sequencing-based assays, array data, etc.)

- 3.3.1 Primary data sets of use to the research community should be submitted to ArrayExpress (EMBL-EBI). For transcriptomics experiments, this should be done within 3 months (maximum) of generation.

Time limit (for transcriptomics): *3 months*

3.4 Mass Spectrometry

- 3.4.1 Primary mass spectrometry data should be submitted to PRIDE (EMBL-EBI).

3.5 Annotation Data

- 3.5.1 Annotation data should be made available as it is generated *via* an appropriate browser (e.g. ensembl, Vega, GeneDB) which allows users to both browse and export annotation to flat files. Where appropriate, annotation should be continuously available *via* Distributed Annotation System (DAS) sources, registered in the DAS registry, so they can be displayed by any genome annotation application or website that is a DAS client.
- 3.5.2 For annotation of sequence data generated at WTSI: the annotation should be included in the final sequence entry submitted to EMBL (see 3.1.3)

3.6 Other Biological/Biochemical Assay Data

- 3.6.1 Other biological/biochemical assay data, such as the results of receptor-ligand interaction studies, images of histological assays, etc., should also be shared *via* WTSI or other suitable databases (e.g., IntAct at EMBL-EBI for protein interaction data).

3.7 Model Organism Phenotypic Information

- 3.7.1 A list of genes under investigation and regular reports on progress should be displayed on the project webpage.¹⁴
- 3.7.2 Molecular phenotype data should be released as in sections 3.1-3.6 above, and links to the data (accession numbers) should be displayed on the project webpage.

¹⁴ Resources available, such as lines of model organisms for distribution, should also be listed on the project webpage or in an appropriate database (e.g., Zebrafish International Resource Centre).

3.7.3 Morphological and other phenotypic data should be submitted to a WTSI database or appropriate other (e.g., Zebrafish Model Organism Database).

3.8 Methods

3.8.1 The methods required to reuse data should be referenced or published with the data. Software should be shared using a free software license, as defined by the free software foundation.¹⁵

Table 1. Summary table for submission of open and managed primary data

Data type	time limit	where to submit primary data	
		open access	managed access
Genomic Sequencing	2 months	ENA	EGA
Genotyping + Cytogenetics	3 months	EGA	EGA
Functional Analysis Assays	3 months (for transcriptomics)	ArrayExpress	EGA

¹⁵ Please see: <http://www.fsf.org>

Rights of Data Providers

The Institute recognises the need for researchers to be appropriately credited for their scientific contribution and investment in data generation. It is therefore expected that all researchers both honour agreements in line with Fort Lauderdale’s data sharing principles,¹⁶ and appropriately acknowledge the contributions of others.

4. Researchers’ Rights and Responsibilities

4.1 Rights

4.1.1 The Institute’s experience indicates that the scientific credit gained from sharing data considerably outweighs the risks to researchers when the ability to publish is appropriately managed. In some cases, that are very visible, this can be done informally, but in others, explicit protection of “first publication rights” by a statement associated with data access¹⁷ or in conjunction with a data access agreement may be warranted. However, agreements such as “first publication rights,” currently referred to as “publication moratoria,” should not be open-ended (i.e., they should be of limited duration and limited to publishing the results of particular analyses), and projects will be monitored by the WTSI Data Sharing Committee to ensure that these do not disproportionately infringe upon other principles of the policy.

4.2 Responsibilities

4.2.1 Researchers are expected to abide by the WTSI code of access when using data that is publicly accessible or shared with the research community:

WTSI Code of Access

Data sharing policies will only be sustainable if researchers commit to:

- conducting appropriate research*
- protecting the confidentiality of managed access data sets*
- carefully communicating research results*
- respecting rights to first publication and to acknowledgment*
- sharing, in turn, resulting data and analysis with the research community*

¹⁶ *Sharing Data from Large-scale Biological Research Projects: A System of Tripartite Responsibility*, Report of a meeting organized by the Wellcome Trust and held on 14–15 January 2003 at Fort Lauderdale, USA.

¹⁷ by including a statement in WTSI’s Guidelines on the Use of Data in Publications, for example.

Optimising Translation

The Institute recognises that, in specific instances, the use of intellectual property protection and attendant potential delays to data sharing may be necessary to prevent inappropriately exclusive claims by others and to ensure health benefits occur.

5. Development of IP

5.1 IP protection

5.1.1 Data sharing may be delayed to seek IP protection when this is necessary to optimise translation.

5.2 Policy

5.2.1 In accordance with the WT policy on intellectual property and patenting,¹⁸ and the WT statement on genome data release.¹⁹

5.3 Procedure

5.3.1 Delays will require approval from the WTSI Data Sharing Committee.

6. Collaborations

6.1.1 WTSI researchers are responsible for ensuring that collaborations respect the Institute's data sharing policy.

¹⁸ available at <http://www.wellcome.ac.uk/About-us/Policy/Policy-and-position-statements/WTD002762.htm>

¹⁹ available at <http://www.wellcome.ac.uk/About-us/Policy/Policy-and-position-statements/WTD002751.htm>

List of Abbreviations

CNV	Copy number variation
COSMIC	The catalogue of somatic mutations in cancer database
dbEST	The expressed sequence tags database
dbSNP	The single nucleotide polymorphism database
dbSTS	The sequence tagged sites database
EBI	European Bioinformatics Institute
EGA	European Genome-phenome Archive
EMBL	European Molecular Biology Laboratory
ENA	European Nucleotide Archive
EST	Expressed sequence tag
IP	Intellectual property
NCBI	National Center for Biotechnology Information
PhD	Doctorate in philosophy
PRIDE	The proteomics identifications database
QC	Quality control
SNP	Single nucleotide polymorphism
SRA	Sequence Read Archive
STS	Sequence tagged sites
UKPMC	UK PubMedCentral
WT	Wellcome Trust
WTSI	Wellcome Trust Sanger Institute

Glossary

Primary data: refers to the lowest level of data that is considered useful to archive in public primary data repositories. Over time, what is considered appropriate by the community is likely to evolve.

Major public repositories: refers to stable, archival databases of biological information which provide accession numbers for submissions, are considered public resources, accept and provide data freely to all users. These are often run by either the EBI or the NCBI.

Reference databases: refers to databases that are recognised by the community as reference sources of data of a particular type in a particular field and have as a core part of their mission collecting, preserving and disseminating this information. These may include the databases of the major public repositories as well as databases such as COSMIC.

Managed access data: refers to data that cannot be made publicly accessible and is therefore shared with the research community *via* a carefully managed access procedure.

Appropriately qualified researcher: we define an appropriately qualified researcher either as someone who has authored a relevant peer-reviewed article that we can locate on [PubMed](#), and who is still working in the field, or as a successful applicant to a relevant data access committee.