Earlham Institute

SCIENTIFIC DATA MANAGEMENT POLICY

Date:	4 December 2012		
Operational Owner:	Head of Research e-Infrastructures	**Release:**	Draft V6
Approver:	EI's Board of Directors		
Document Number:	EI Policies Ref X		

Note: This document is only valid on the day it was printed

Revision History

Date of next revision:

Revision Date	Previous Revision Date	Summary of Changes	Changes Marked	Version
29/11/12	n/a	Draft for Circulation to Executive Team for comment	No	V1
3/12/12	29/11/12	Revisions from Sarah Cossey	No	V2
4/12/12	3/12/12	Revisions from Mario Caccamo	No	V3
6/12/12	4/12/12	Revisions from Rob Davey. Draft ready for circulation to Board.	No	V4
15/2/15	6/12/12	Revisions from Tim Stitt	No	V5
03/07/18	15/2/15	Brand update by Christine Fosker	No	V6

CONTENTS

1. PRINCIPLES OF SCIENTIFIC DATA MANAGEMENT

1.1 Scientific data is one of the main outputs of EI's activities; therefore the policy and processes around the management of these assets are central to the day-to-day running of the Centre. The principles of data management underpin the generation, preservation, storage and manipulation of data generated at EI. The mechanisms required to implement a data management policy should be:

(a) Cost effective

(b) Focused on data quality

(c) Led by the needs of the scientific community

(d) Driven by science

(e) Honest and open

(f) Secure

1.2 This scientific data management policy covers all data types generated by experiments and analysis carried out at EI within the remit of publically funded research. Henceforth the term "data" in this policy will refer explicitly to "scientific data".

1.3 Data types range from the raw sequences generated by the sequencing instruments to the technical information obtained by downstream bioinformatics analysis, including the development of software (see 1.4 below). The policy is focused on the following aspects of scientific data manipulation:

(a) Collection

(b) Data integrity

(c) Standards

(d) Retention

(e) Availability

(f) Recovery

1.4 This data policy also applies to the generation of software developed at EI under the open-source philosophy. The mechanism for the dissemination and long-term archiving of software will differ from the other data types outlined in this document but the principles are essentially the same.

2. Data Generation, Collection and Backups.

2.1 Data generation refers to the outputs from the sequencing and mapping instruments hosted at EI.

2.2 It also includes the recording and archiving of related metadata[1] referring to the description of the sample sources and preparations, experimental parameters, and configuration of the instruments. These details are tracked and stored in EI's laboratory information management system (LIMS). SOP documents available to EI staff describe the use and operation of the LIMS system and ancillary strategies (such as the use of document management and issue tracking productivity tools).

2.3 At all times, the data should be stored on a device that is regularly backed up. Given the demands of current data volumes two levels of redundancy have been implemented to support data backup:

(a) Replication of the entire primary storage filesystem onto a secondary ancillary device that is physically hosted in different premises i.e. "mirroring". Mirroring runs several times per day and is managed by NBIP on EI's behalf.

(b) Data "snapshots" of the primary storage filesystem are taken several times per day, allowing point in time recovery at file or folder level by end users or system administrators. Snapshots are retained daily, weekly, monthly, and quarterly to allow recovery to previous time points, at decreasing granularity.

2.4 Other supporting infrastructure such as Virtual Machine images etc. is backed up from dedicated storage to an enterprise grade tape library.

2.5 All storage systems are located in physically secure datacentres, with power supplies protected by UPS and generator backup. The datacentres are managed by NBIP on behalf of EI.

2.6 The Scientific Computing team will review EI's backup strategy, managed by NBIP on EI's behalf, annually. It will also be tested for reliability and response times, internally every quarter, and externally every two years. Performances reports will be provided to the Finance, Resources and Audit Committee on an annual basis.

[1] Metadata refers to information that is recorded about data such as when it was collected and by whom etc.

3. DATA INTEGRITY

3.1 The preservation of the integrity of the data refers to maintaining and assuring its accuracy and consistency throughout its lifecycle within EI. Strategies for data integrity should monitor and correct accidental data corruption due to hardware malfunctioning as well as actions, unintended or otherwise, from member of staff or external users.

3.2 Data signatures called ‘checksums’ or ‘hashsums’ will be generated for each dataset. These will be based on encryption algorithms and provide a near unique (within a small collision probability) signature that identifies a dataset.

3.3 Signatures will be checked when data is transferred over a link or moved or copied internally, when the risk of error increases.

3.4 Primary raw data and their signatures will be stored as read-only files, and will be versioned to reflect the data lifecycle. The original ‘checksums’ can be reviewed at any point to check data integrity.

4. DATA STANDARDS

4.1 Data standards provide the framework to implement unambiguous and machine-readable descriptions of the data. EI will subscribe to accepted data standards and will also engage with the community to promote evolution of these standards as well as to keep staff trained and aware of novel developments.

4.2 Compliance with data standards will be regularly checked by health-check strategies implemented by the Sequencing Informatics teams, ensuring that the data made available to collaborators are formatted and presented in accordance with the suitable standards. Examples of data standards used by EI are:

(a) FASTA/FASTQ/BAM/SAM for sequences

(b) VCF/BCF for variation and allele information

(c) AGP for sequence assemblies

(d) GFF for gene annotation

(e) SRA metadata schemas for data submission

5. DATA RETENTION AND ARCHIVING

5.1 Policies for data retention and archiving refer to the strategies and mechanisms to ensure that once data are generated, they are readily available and accessible.

5.2 The principles of good practice for data retention are based on latency (i.e. access time) and compliance with third party contractual agreements. One example is the obligation imposed by funding bodies linked to the generation of the data.

5.3 EI will use the best available compression tools to ensure that data storage is performed at reasonable cost without impacting access beyond reasonable or expected time frames.

5.4 Data will be stored on the most suitable hardware unit that befits the retention of the data. New data will be stored on faster hardware to facilitate secondary analysis, whereas older, less frequently accessed data will be on slower cheaper (£/storage size unit) hardware. Data can be moved back and forth between these areas as necessary by designated staff.

5.5 After completion of a project, EI will deposit the sequences and analysis data in the public repositories hosted by the European Bioinformatics Institute (EBI) with the understanding that this represents a reliable long-term solution. EI will regularly monitor the viability and suitability of this solution and will consider hosting the data elsewhere if deemed necessary.

6. DATA AVAILABILITY AND DISSEMINATION

6.1 EI supports 3rd party repositories, namely those at the European Bioinformatics Institute, as its primary dissemination focus. These repositories provide the long-term standards and retention capability that promotes the widest-possible efficient and timely data availability.

6.2 Where appropriate, deposition of data will be made in agreement with relevant collaborators to allow sufficient time for publication of any findings, whilst ensuring any confidentiality or ethical considerations.

EI software development projects will release source code to 3rd party repositories, such as Github[2], to promote community engagement and reproducibility of any analysis undertaken.

[2] Git is an extremely fast, efficient, distributed version control system ideal for the collaborative development of software. GitHub is an online community portal that interacts with the Git software, allowing easy dissemination of a code base, thus promoting collaboration with others on software projects.