DATA SHARING GUIDANCE FOR CRUK RESEARCHERS

INITIATIVES AND REPOSITORIES TO SUPPORT DISCOVERY RESEARCH (SCIENCE COMMITTEE) RESEARCHERS WITH DATA MANAGEMENT AND SHARING

CONTENTS

1. INTRODUCTION

In the following pages, we highlight some key initiatives and repositories which may provide tools and solutions for data sharing in fields within the remit of Science Committee. Some of these are not discipline specific and are described as “generalist” (Section 2); others are specific to fields relevant to Science Committee, such as basic cancer biology and omics research (Section 3). It is the responsibility of the investigators to ensure that any repositories/standards/tools they intend to use are appropriate for the nature of the research envisaged.

CONTACT US

We would be grateful for any comments or suggestions to help improve this guidance. Please get in touch with jamie.enoch@cancer.org.uk with any feedback.

2. GENERALIST DATA SHARING INITIATIVES AND REPOSITORIES

2.1. INITIATIVES

Listed below are generic initiatives supporting data sharing in science and health research generally, which may provide useful tools, resources and methods you can factor into your data sharing plan.

DataCite provides advice on creating digital object identifiers (DOIs) for datasets. It also offers a range of services including a Metadata search, which allows researchers to locate individual datasets through access to the relevant metadata, and a search tool to discover appropriate research data repositories worldwide through Re3Data.

The Digital Curation Centre provides expert advice and practical help to researchers to store, manage, protect and share digital research data. It maintains a range of resources including How-to Guides and checklists, case studies in research data management, and training programmes for researchers and data custodians in research data management and sharing.

The Expert Advisory Group on Data Access convenes leading researchers on issues of data access and sharing who advise the funders on technical and legal issues in data sharing. It has produced a number of reports which advise on good practice in data sharing policy and governance.

GigaScience is an open-access open-data journal for 'big data studies' from across the life/biomedical sciences, in collaboration with BioMed Central. In many ways it is half journal and half repository. It links a standard manuscript publication with an extensive database, the GigaDB, which hosts associated data and provides data analysis tools and cloud-computing resources.

Nature *Scientific Data* publishes peer-reviewed articles known as data descriptors, which focus on helping others reuse data by describing the dataset with structured, machine readable information. The articles can be descriptions of datasets of any size, and can link to datasets underpinning published research or describe standalone datasets. The journal mandates the release of datasets accompanying manuscripts, and links to datasets hosted on third-party repositories.

The Research Data Alliance works to build the social and technical bridges to facilitate data sharing and re-use. Its constituent working groups aim to tackle fundamental issues in data sharing; for example, one working group is seeking to rationalise databases, standards and funder policies in the biomedical sciences by working with the BioSharing platform.

2.2. REPOSITORIES

Where possible, we would encourage researchers to deposit their data in repositories specifically intended for their discipline. However for some types of data (e.g. unstructured datasets), these generalist repositories may be more appropriate:

REPOSITORY	TYPES OF DATA RESEARCHERS MAY SUBMIT
Dryad Digital Repository	Data underlying scientific or medical publications
Figshare	Various data types, including figures, datasets and images
GigaDB	Data and tools for *GigaScience* and other articles
Zenodo	Research outputs from all fields of science

3. INITIATIVES AND REPOSITORIES SPECIFIC TO DATA SHARING WITHIN SCIENCE COMMITTEE DISCIPLINES

3.1. INITIATIVES

A number of initiatives are ongoing in the space of data sharing from basic biomedical science, facilitating the sharing of data generated from bioinformatics, genomics, other omics, cell biology or translational research. You may consider engaging with these initiatives or consulting the resources they produce as you plan for data sharing in your project.

Some select examples relevant to cancer, ranging from small grassroots organisations to large international bodies, include:

BioSharing, a global portal of information resources in the biosciences. It is particularly useful for searching for databases, repositories, standards and policies which can facilitate biomedical data sharing.

The Centre for Therapeutic Target Validation provides a precompetitive space to generate evidence on the validity of therapeutic targets, drawing on large volumes of data from genomics, proteomics, chemistry and disease biology.

DNA Digest is a not-for-profit organisation, aiming to educate, facilitate and engage on issues regarding sharing of and access to genomic data.

DREAM Challenges are an open science effort seeking to incentivise the development of new computational biology methods and the innovative re-use of existing data from large scale international cancer research efforts.

The European Bioinformatics Institute is a leading centre for computational biology, but also offers comprehensive molecular databases, together with a Data Submission Wizard to guide researchers to the most appropriate data repository. It also offers an extensive user training programme.

eMedLab aims to "maximise the gains for patients and for medical research that will come from the explosion in human health data" and provide a shared computer cluster to integrate and share heterogeneous data from personal healthcare records, imaging, pharmacoinformatics and genomics.

The Genomic Standards Consortium promotes mechanisms that standardize the description, exchange and integration of genomic data. For example, the GSC has established the Minimum Information about any (X) Sequence (MIxS).

The Global Alliance for Genomics and Health aims to maximise the potential of genomic medicine through responsible data sharing through four thematic working groups. These groups produce good practice guidelines e.g. the Regulatory & Ethics group's *Framework for Responsible Sharing of Genomic and Health-Related Data.*

The International Cancer Genome Consortium aims to obtain a comprehensive description of genomic, transcriptomic and epigenomic changes in 50 different tumour types and sub-types. ICGC's main output is its compendiums of genomic alterations in many cancer sub-types, which it maintains on its data portal.

The *Open Journal of Bioresources* is a data journal featuring peer-reviewed short papers helping researchers to locate and cite bioresources with high reuse potential.

The (US) National Center for Biotechnology Information provides a number of repositories comparable to those maintained by EMBL-EBI (with whom they exchange their sequence data on a daily basis).

3.2. REPOSITORIES

Below are some widely-used community repositories for different kinds of basic science data, categorised into:

1. biological materials
2. data specific to cancer
3. generic biology/biochemistry
4. model organisms
5. omics
6. sequencing data
7. structural databases.

If you are looking for a more specific disciplinary repository or one not covered in the lists below, BioSharing provides a searchable catalogue of information and database resources.

3.2.1. *Biological Materials*

REPOSITORY	TYPES OF DATA RESEARCHERS MAY SUBMIT
Addgene	Plasmids
American Type Culture Collection (ATCC)	Cell lines and microorganisms
European Mouse Mutant Archive (EMMA)	Mutant mouse strains
Jackson Laboratory	Mutant mouse strains
Knockout Mouse Project Repository	Mouse embryonic stem cells containing a null mutation in every gene in the mouse genome
Mutant Mouse Resource and Research Centres	Genetically engineered mouse strains and mouse ES cell lines
RIKEN BioResource Centre	Cell lines
UK Stem Cell Bank	Human stem cell lines

3.2.2. *Data specific to cancer*

REPOSITORY	TYPES OF DATA RESEARCHERS MAY SUBMIT
caNanoLab	Nanomaterials
Cancer Models Database	Models for human cancer, including genetic description, histopathology, derived cell lines, associated images, carcinogenic agents, and therapeutic trials
COSMIC	Somatic mutations in cancer
Mouse Tumour Biology	Mouse tumour biology data
NCI Mouse Repository	Mouse cancer models and associated strains
The Cancer Imaging Archive	Medical images of cancer (matched to the NIH Cancer Genome Atlas)

3.2.3. *Generic biology and biochemistry data*

REPOSITORY	TYPES OF DATA RESEARCHERS MAY SUBMIT
BioModels	Computational models of biological processes
BioSamples	Reference sample data
Kinetic Models of Biological Systems	Quantitative kinetic models from systems biology
PubChem	Chemical molecules and their activities against biological assays
Rhea	Reaction data & annotations

3.2.4. *Model organisms*

REPOSITORY	TYPES OF DATA RESEARCHERS MAY SUBMIT
BioGRID	Genetic and protein interaction data from model organisms and humans
FlyBase	Database of *Drosophila* genes and genomes
Mouse Genome Informatics	Integrated genetic, genomic, and biological data from laboratory mice
Mouse Phenome Database	Measured data on laboratory mouse strains and populations
Rat Genome Database	Genomic, genetic, functional, physiological, pathway and disease data for rat
WormBase	Biological and genomic data on the nematode model organism
XenBase	Genomic and biological data on Xenopus frogs
Zebrafish Model Organism Database	Zebrafish genetic, genomic and developmental data

3.2.5. *Omics*

REPOSITORY	TYPES OF DATA RESEARCHERS MAY SUBMIT
Array Express	Microarray-based gene-expression data, compliant with MIAME standard
ClinVar	Information about genomic variation and its relationship to human health
Database of Interacting Proteins	Experimentally determined interactions between proteins
dbGaP	Coded genotype, phenotype, exposure, and pedigree data from genome-wide association studies; also copy number variants and large-scale sequencing
EGA - European Genome-Phenome Archive	Repository for all types of sequence and genotype experiments
Gene Expression Omnibus	Array- and sequence-based data, compliant with MIAME standard
Genome RNAi	Phenotypes from RNA interference (RNAi) screens in Drosophila and humans
IntAct	Molecular interactions
MetaboLights	Metabolomics data
Peptide Atlas	Peptides identified in mass spectrometry proteomics experiments
PRIDE	Protein and peptide identification data
ProteomeXchange	Mass spectrometry proteomics data

3.2.6. *Sequencing*

REPOSITORY	TYPES OF DATA RESEARCHERS MAY SUBMIT
dbSNP	Single nucleotide polymorphisms (SNPs) and multiple small-scale variations
dbVar	Genomic structural variation
DGVa	Genomic structural variation
DNA Databank of Japan	Nucleotide sequence data
EBI Metagenomics	Raw sequence data and associated meta-data
ENA - European Nucleotide Archive	Nucleotide sequence data
EVA - European Variation Archive	Genetic variation data from all species
GenBank	An annotated collection of all publicly available DNA sequences
NCBI Sequence Read Archive	Raw sequence data from "next-generation" sequencing technologies
UniProtKB	Protein sequence and functional information

3.2.7. *Structural databases*

REPOSITORY	TYPES OF DATA RESEARCHERS MAY SUBMIT
Biological Magnetic Resonance Data Bank	Data from NMR Spectroscopy on Proteins, Peptides, Nucleic Acids, and other Biomolecules
Cambridge Crystallographic Data Centre	Crystallographic data for small molecules
Coherent X-ray Imaging Data Bank	Data from Coherent X-ray Imaging experiments
Crystallography Open Database	Crystal structures of organic, inorganic, metal-organic compounds and minerals
Electron Microscopy Data Bank	Electron microscopy density maps of macromolecular complexes and subcellular structures
FlowRepository	Flow cytometry experiments, compliant with MIFlowCyt standard
Protein Circular Dichroism Data Bank	Circular dichroism spectral and metadata
Worldwide Protein Data Bank	Information about the 3D structures of proteins, nucleic acids, and complex assemblies