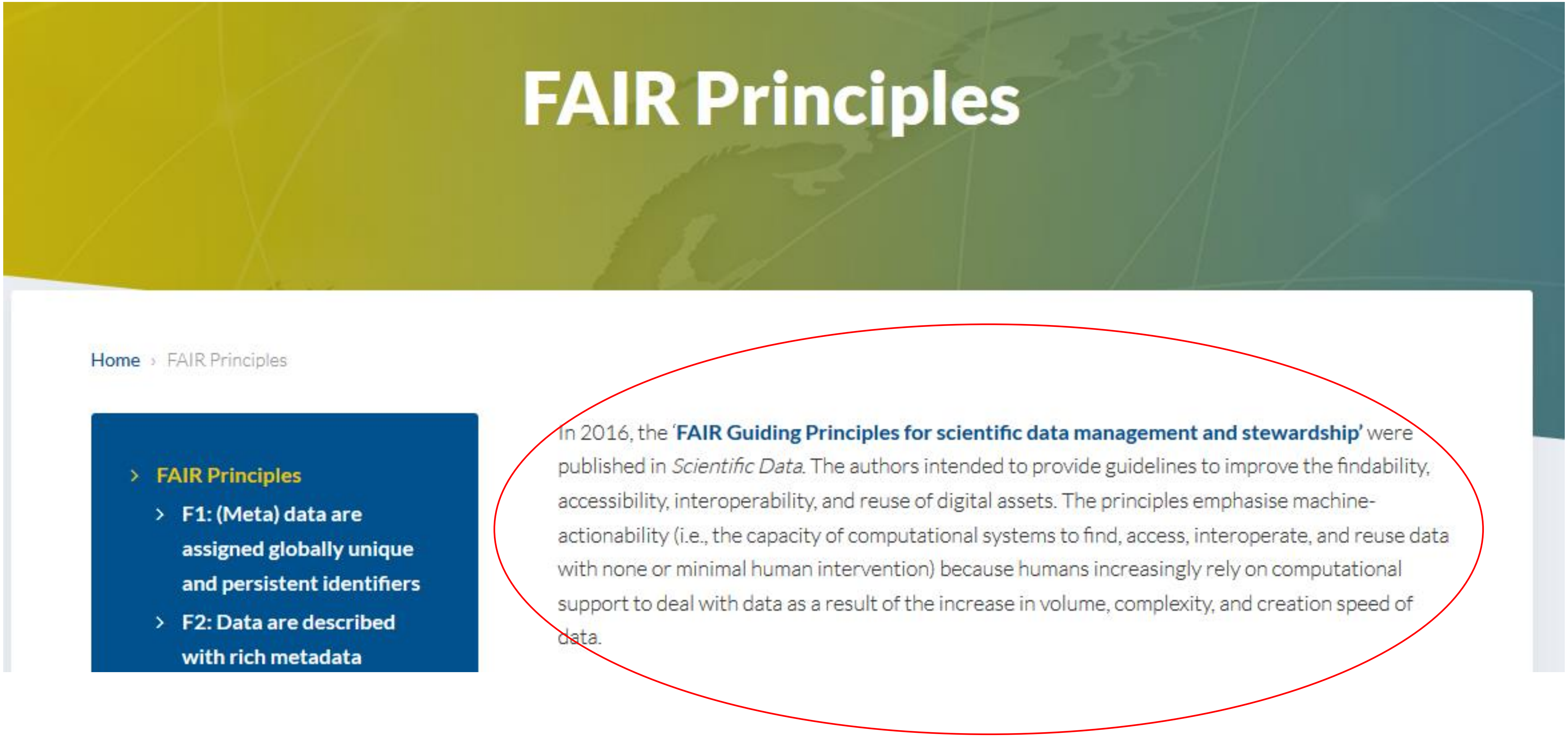
Open science

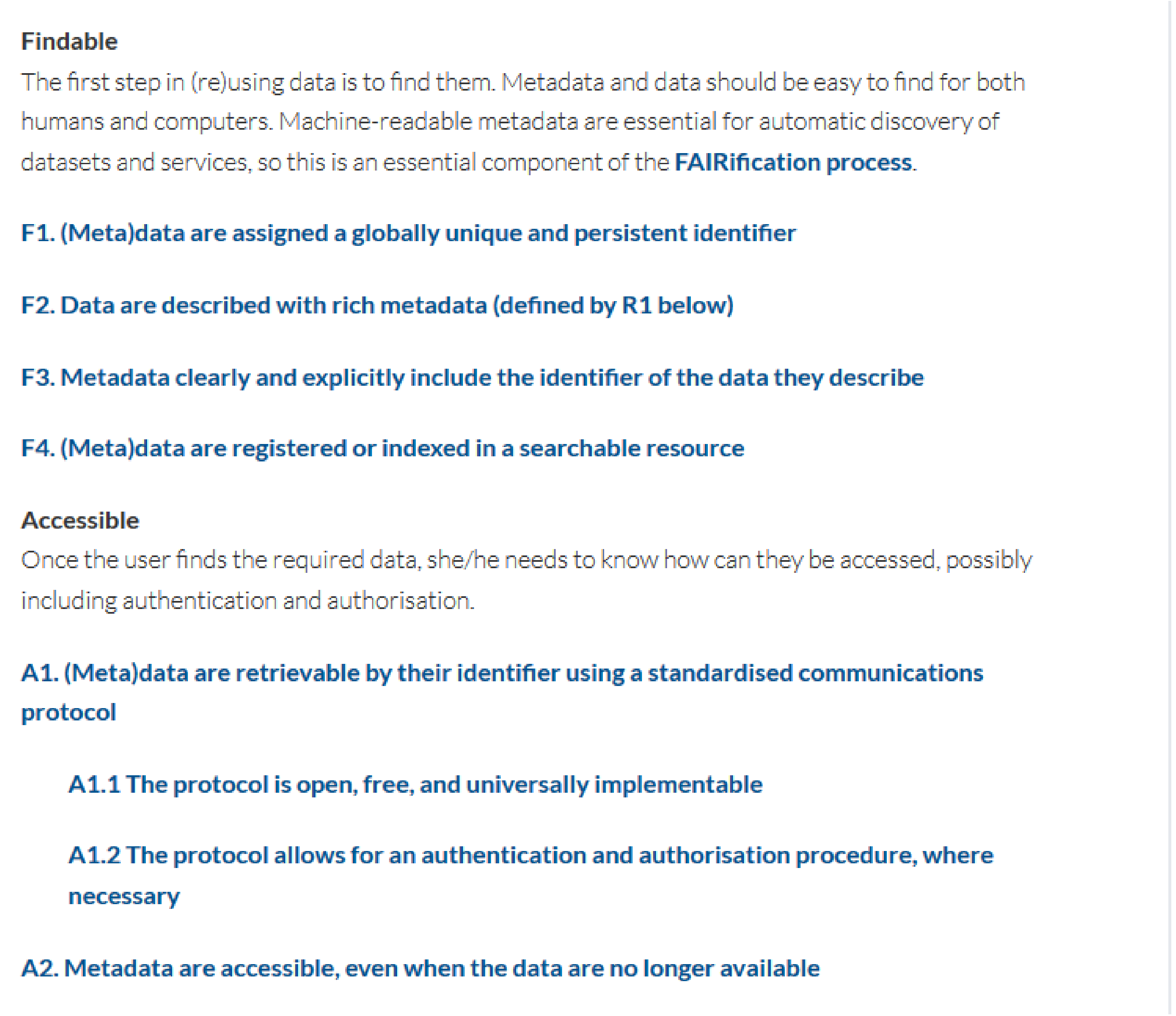
Use of data repositories



FAIR Principles

Home › FAIR Principles
› FAIR Principles
› F1: (Meta) data are assigned globally unique and persistent identifiers
› F2: Data are described with rich metadata

In 2016, the '**FAIR Guiding Principles for scientific data management and stewardship**' were published in *Scientific Data*. The authors intended to provide guidelines to improve the findability, accessibility, interoperability, and reuse of digital assets. The principles emphasise machine-actionability (i.e., the capacity of computational systems to find, access, interoperate, and reuse data with none or minimal human intervention) because humans increasingly rely on computational support to deal with data as a result of the increase in volume, complexity, and creation speed of data.



Findable

The first step in (re)using data is to find them. Metadata and data should be easy to find for both humans and computers. Machine-readable metadata are essential for automatic discovery of datasets and services, so this is an essential component of the **FAIRification process**.

F1. (Meta)data are assigned a globally unique and persistent identifier

F2. Data are described with rich metadata (defined by R1 below)

F3. Metadata clearly and explicitly include the identifier of the data they describe

F4. (Meta)data are registered or indexed in a searchable resource

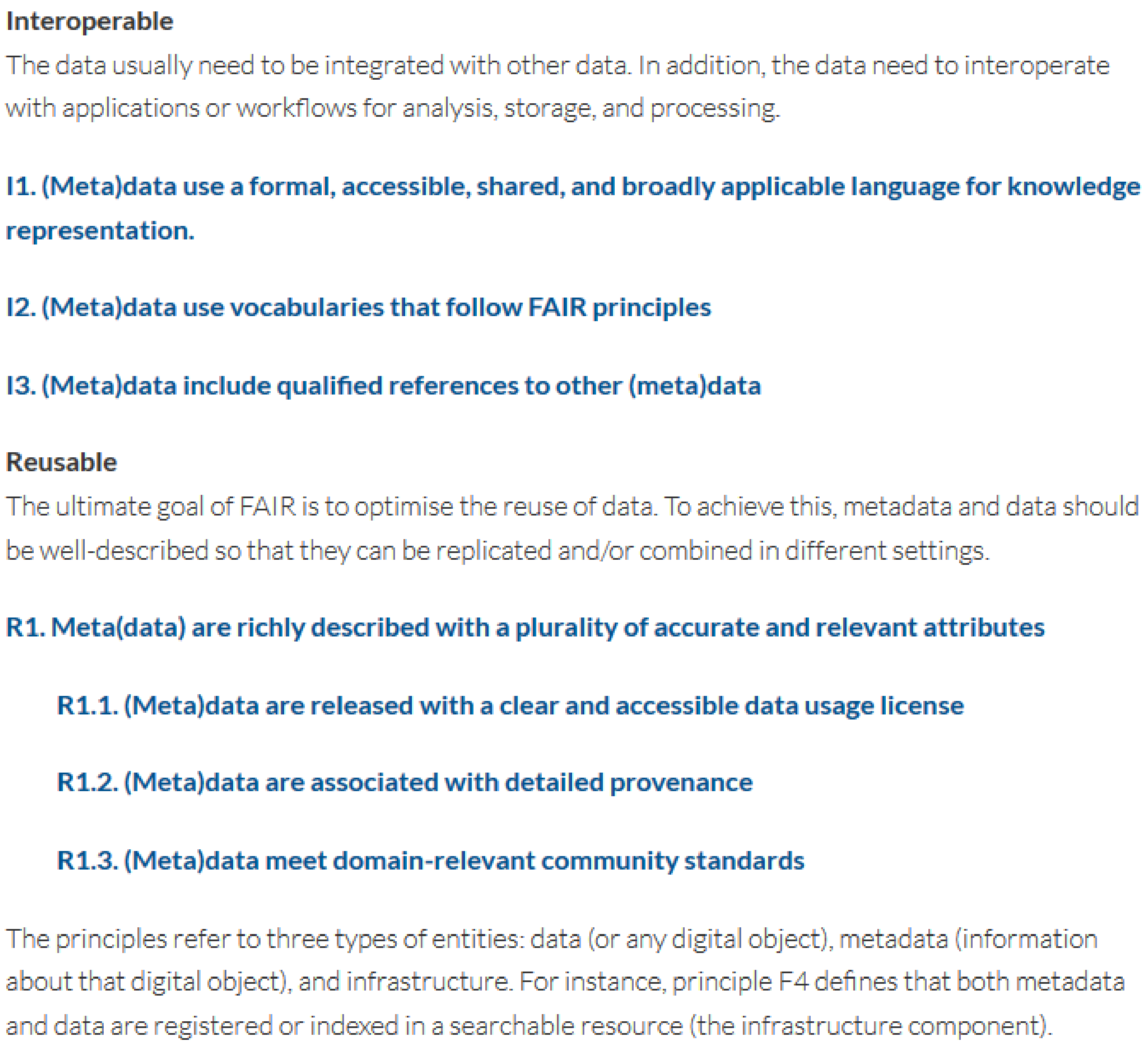
Accessible

Once the user finds the required data, she/he needs to know how can they be accessed, possibly including authentication and authorisation.

A1. (Meta)data are retrievable by their identifier using a standardised communications protocol

> **A1.1 The protocol is open, free, and universally implementable**
>
> **A1.2 The protocol allows for an authentication and authorisation procedure, where necessary**

A2. Metadata are accessible, even when the data are no longer available



Interoperable

The data usually need to be integrated with other data. In addition, the data need to interoperate with applications or workflows for analysis, storage, and processing.

I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

I2. (Meta)data use vocabularies that follow FAIR principles

I3. (Meta)data include qualified references to other (meta)data

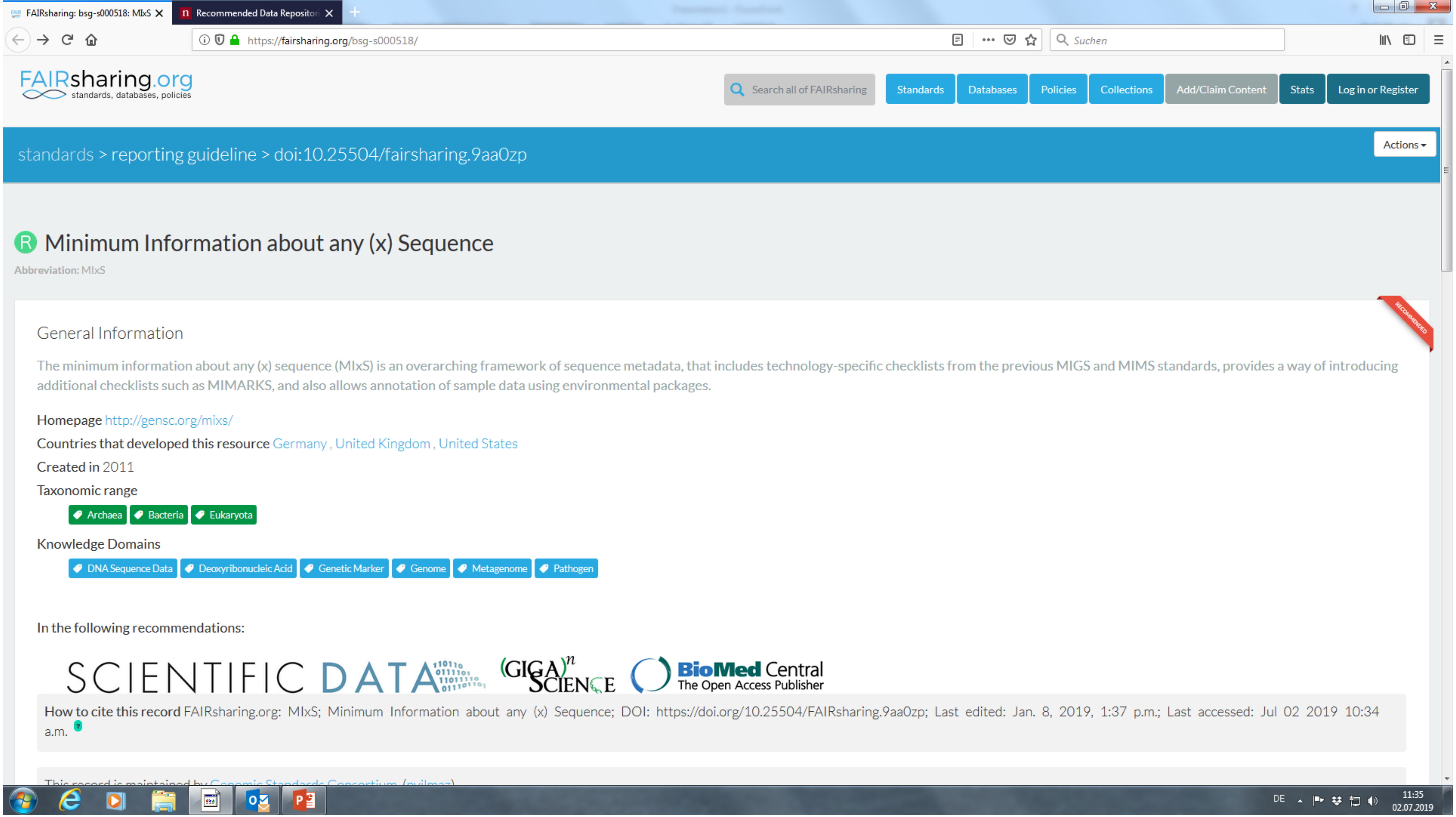
Reusable

The ultimate goal of FAIR is to optimise the reuse of data. To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings.

R1. Meta(data) are richly described with a plurality of accurate and relevant attributes

> **R1.1. (Meta)data are released with a clear and accessible data usage license**
>
> **R1.2. (Meta)data are associated with detailed provenance**
>
> **R1.3. (Meta)data meet domain-relevant community standards**

The principles refer to three types of entities: data (or any digital object), metadata (information about that digital object), and infrastructure. For instance, principle F4 defines that both metadata and data are registered or indexed in a searchable resource (the infrastructure component).


FAIRsharing: bsg-s000518: MIxS
Recommended Data Repositor
https://fairsharing.org/bsg-s000518/
Suchen
FAIRsharing.org
standards, databases, policies
Search all of FAIRsharing
Standards
Databases
Policies
Collections
Add/Claim Content
Stats
Log in or Register
Actions
RECOMMENDED
DE
11:35
02.07.2019

standards > reporting guideline > doi:10.25504/fairsharing.9aa0zp

R Minimum Information about any (x) Sequence

Abbreviation: MIxS

General Information

The minimum information about any (x) sequence (MIxS) is an overarching framework of sequence metadata, that includes technology-specific checklists from the previous MIGS and MIMS standards, provides a way of introducing additional checklists such as MIMARKS, and also allows annotation of sample data using environmental packages.

Homepage http://gensc.org/mixs/

Countries that developed this resource Germany , United Kingdom , United States

Created in 2011

Taxonomic range

Archaea Bacteria Eukaryota

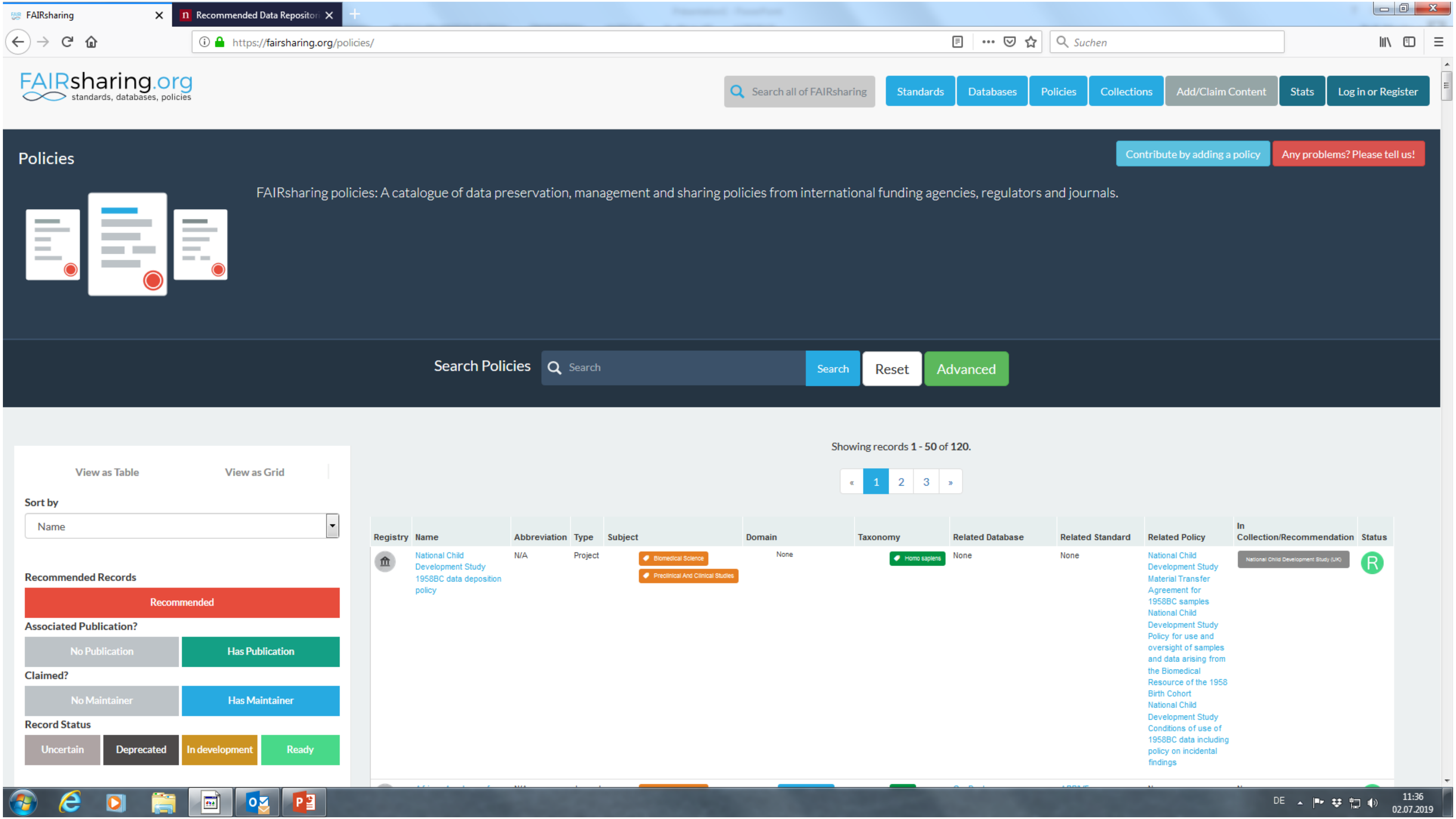
Knowledge Domains

DNA Sequence Data Deoxyribonucleic Acid Genetic Marker Genome Metagenome Pathogen

In the following recommendations:

How to cite this record FAIRsharing.org: MIxS; Minimum Information about any (x) Sequence; DOI: https://doi.org/10.25504/FAIRsharing.9aa0zp; Last edited: Jan. 8, 2019, 1:37 p.m.; Last accessed: Jul 02 2019 10:34 a.m. ?

This record is maintained by Genomic Standards Consortium (pyilmaz)



FAIRsharing
Recommended Data Repositori
https://fairsharing.org/policies/
Suchen
FAIRsharing.org
standards, databases, policies
Search all of FAIRsharing
Standards
Databases
Policies
Collections
Add/Claim Content
Stats
Log in or Register

Policies

FAIRsharing policies: A catalogue of data preservation, management and sharing policies from international funding agencies, regulators and journals.

Sort by: Name

Recommended Records: Recommended

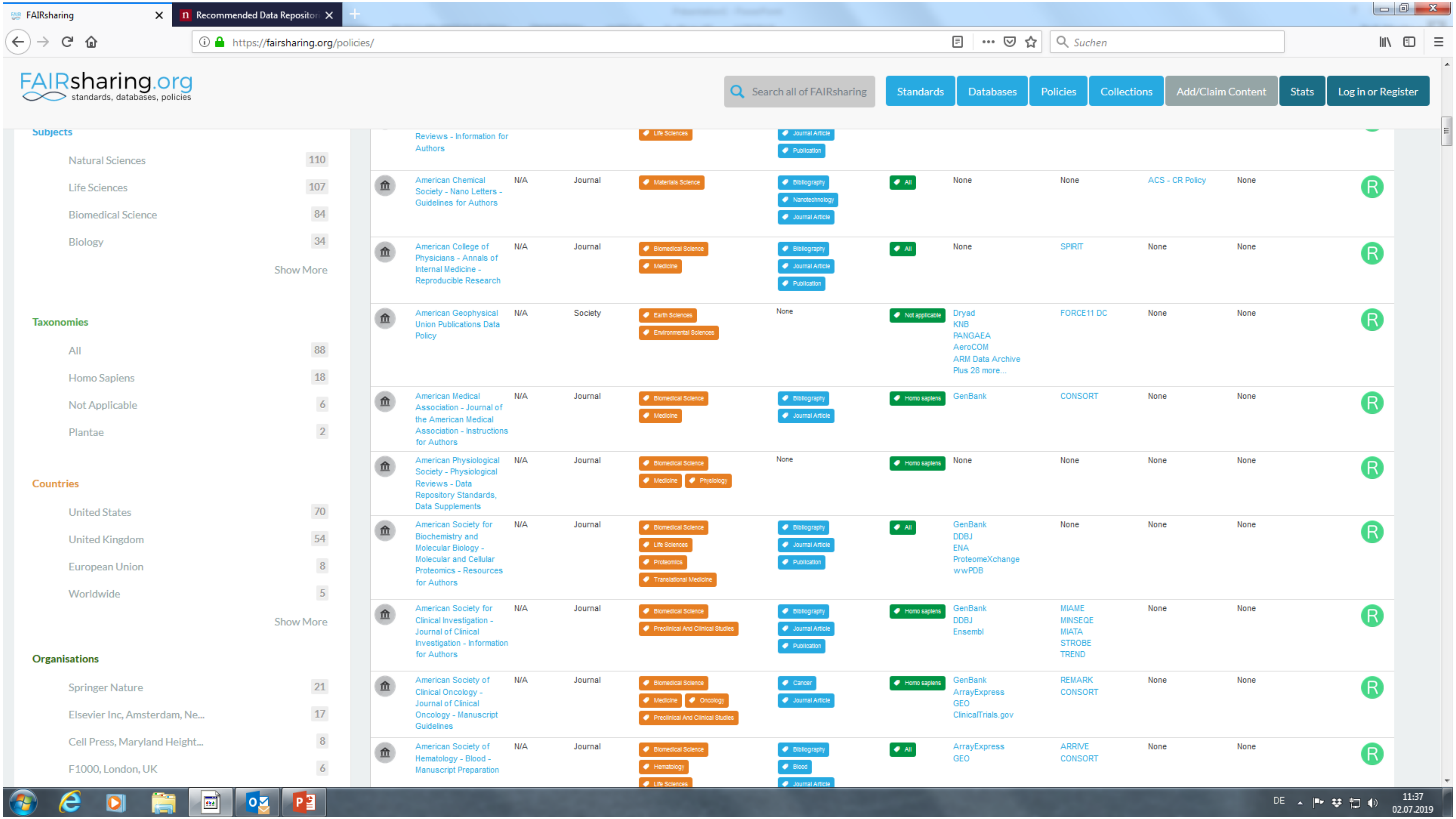
Associated Publication?: No Publication | Has Publication

Claimed?: No Maintainer | Has Maintainer

Record Status: Uncertain | Deprecated | In development | Ready

Showing records **1 - 50** of **120**.

Registry	Name	Abbreviation	Type	Subject	Domain	Taxonomy	Related Database	Related Standard	Related Policy	In Collection/Recommendation	Status
🏛	National Child Development Study 1958BC data deposition policy	N/A	Project	Biomedical Science Preclinical And Clinical Studies	None	Homo sapiens	None	None	National Child Development Study Material Transfer Agreement for 1958BC samples National Child Development Study Policy for use and oversight of samples and data arising from the Biomedical Resource of the 1958 Birth Cohort National Child Development Study Conditions of use of 1958BC data including policy on incidental findings	National Child Development Study (UK)	R



Subjects

Natural Sciences	110
Life Sciences	107
Biomedical Science	84
Biology	34

Show More

Taxonomies

All	88
Homo Sapiens	18
Not Applicable	6
Plantae	2

Countries

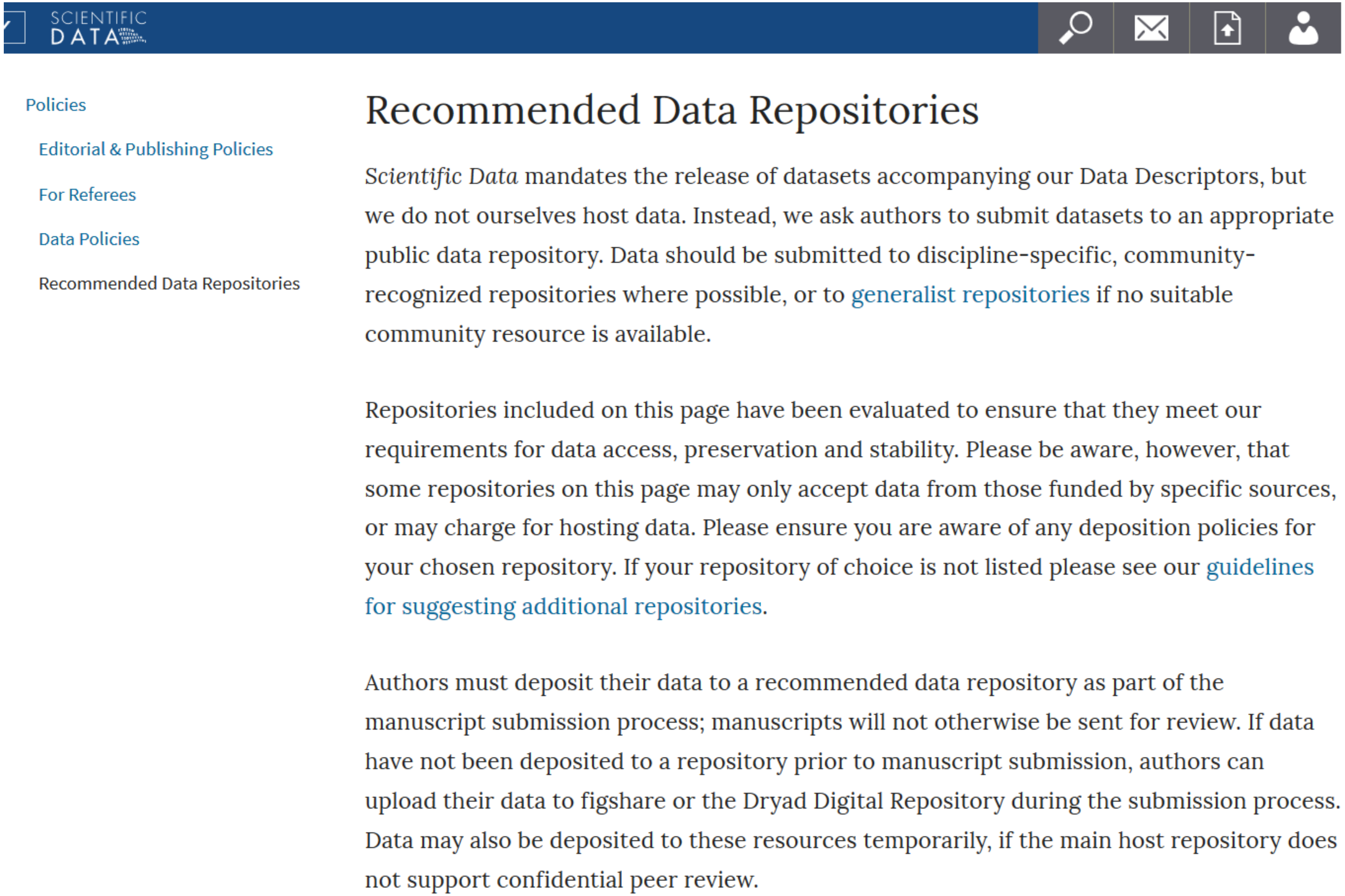
United States	70
United Kingdom	54
European Union	8
Worldwide	5

Show More

Organisations

Springer Nature	21
Elsevier Inc, Amsterdam, Ne...	17
Cell Press, Maryland Height...	8
F1000, London, UK	6

Reviews - Information for Authors			Life Sciences	Journal Article Publication						
American Chemical Society - Nano Letters - Guidelines for Authors	N/A	Journal	Materials Science	Bibliography Nanotechnology Journal Article	All	None	None	ACS - CR Policy	None	R
American College of Physicians - Annals of Internal Medicine - Reproducible Research	N/A	Journal	Biomedical Science Medicine	Bibliography Journal Article Publication	All	None	SPIRIT	None	None	R
American Geophysical Union Publications Data Policy	N/A	Society	Earth Sciences Environmental Sciences	None	Not applicable	Dryad KNB PANGAEA AeroCOM ARM Data Archive Plus 28 more...	FORCE11 DC	None	None	R
American Medical Association - Journal of the American Medical Association - Instructions for Authors	N/A	Journal	Biomedical Science Medicine	Bibliography Journal Article	Homo sapiens	GenBank	CONSORT	None	None	R
American Physiological Society - Physiological Reviews - Data Repository Standards, Data Supplements	N/A	Journal	Biomedical Science Medicine Physiology	None	Homo sapiens	None	None	None	None	R
American Society for Biochemistry and Molecular Biology - Molecular and Cellular Proteomics - Resources for Authors	N/A	Journal	Biomedical Science Life Sciences Proteomics Translational Medicine	Bibliography Journal Article Publication	All	GenBank DDBJ ENA ProteomeXchange wwPDB	None	None	None	R
American Society for Clinical Investigation - Journal of Clinical Investigation - Information for Authors	N/A	Journal	Biomedical Science Preclinical And Clinical Studies	Bibliography Journal Article Publication	Homo sapiens	GenBank DDBJ Ensembl	MIAME MINSEQE MIATA STROBE TREND	None	None	R
American Society of Clinical Oncology - Journal of Clinical Oncology - Manuscript Guidelines	N/A	Journal	Biomedical Science Medicine Oncology Preclinical And Clinical Studies	Cancer Journal Article	Homo sapiens	GenBank ArrayExpress GEO ClinicalTrials.gov	REMARK CONSORT	None	None	R
American Society of Hematology - Blood - Manuscript Preparation	N/A	Journal	Biomedical Science Hematology Life Sciences	Bibliography Blood Journal Article	All	ArrayExpress GEO	ARRIVE CONSORT	None	None	R

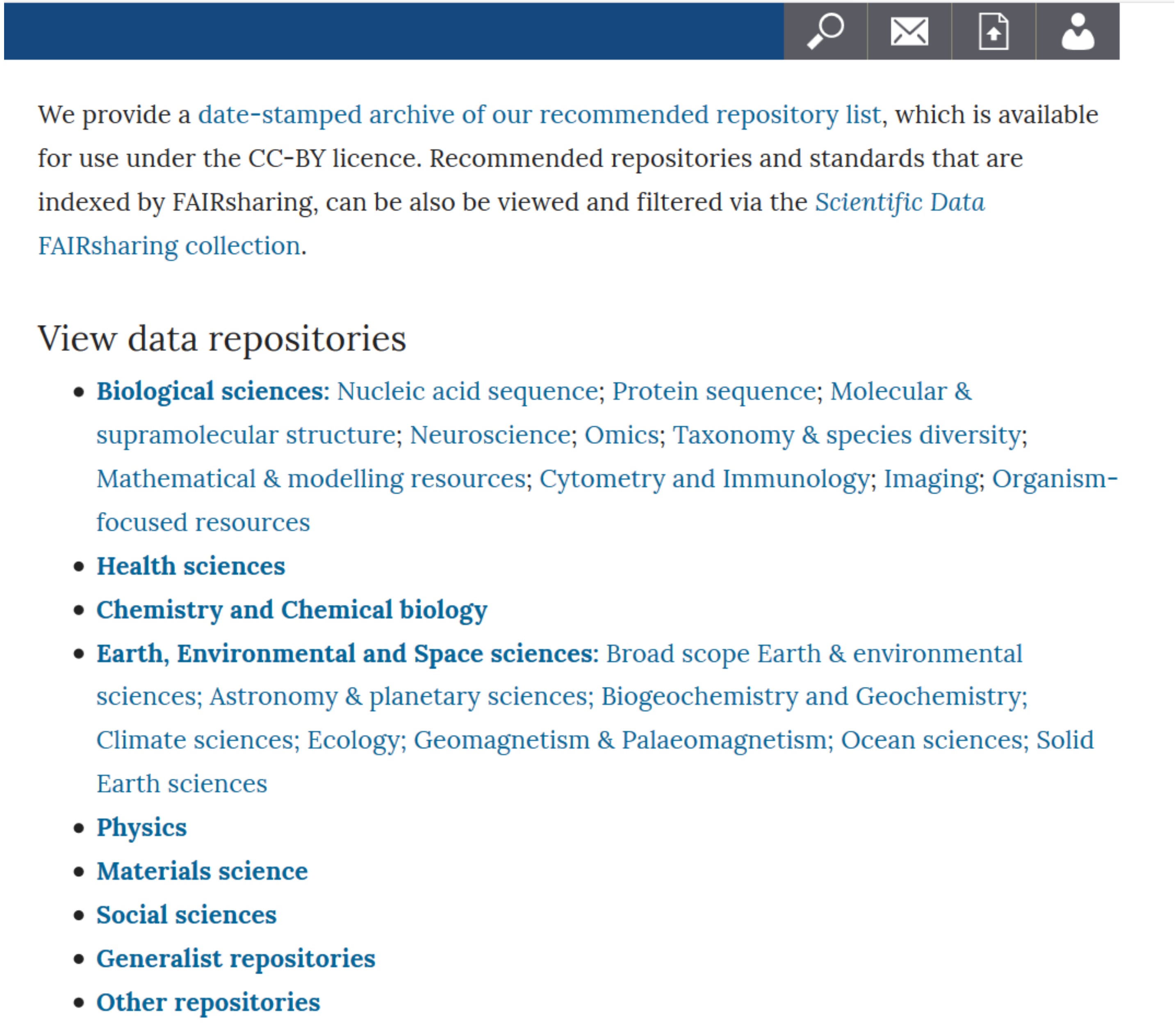

SCIENTIFIC DATA
Policies
Editorial & Publishing Policies
For Referees
Data Policies
Recommended Data Repositories

Recommended Data Repositories

Scientific Data mandates the release of datasets accompanying our Data Descriptors, but we do not ourselves host data. Instead, we ask authors to submit datasets to an appropriate public data repository. Data should be submitted to discipline-specific, community-recognized repositories where possible, or to generalist repositories if no suitable community resource is available.

Repositories included on this page have been evaluated to ensure that they meet our requirements for data access, preservation and stability. Please be aware, however, that some repositories on this page may only accept data from those funded by specific sources, or may charge for hosting data. Please ensure you are aware of any deposition policies for your chosen repository. If your repository of choice is not listed please see our guidelines for suggesting additional repositories.

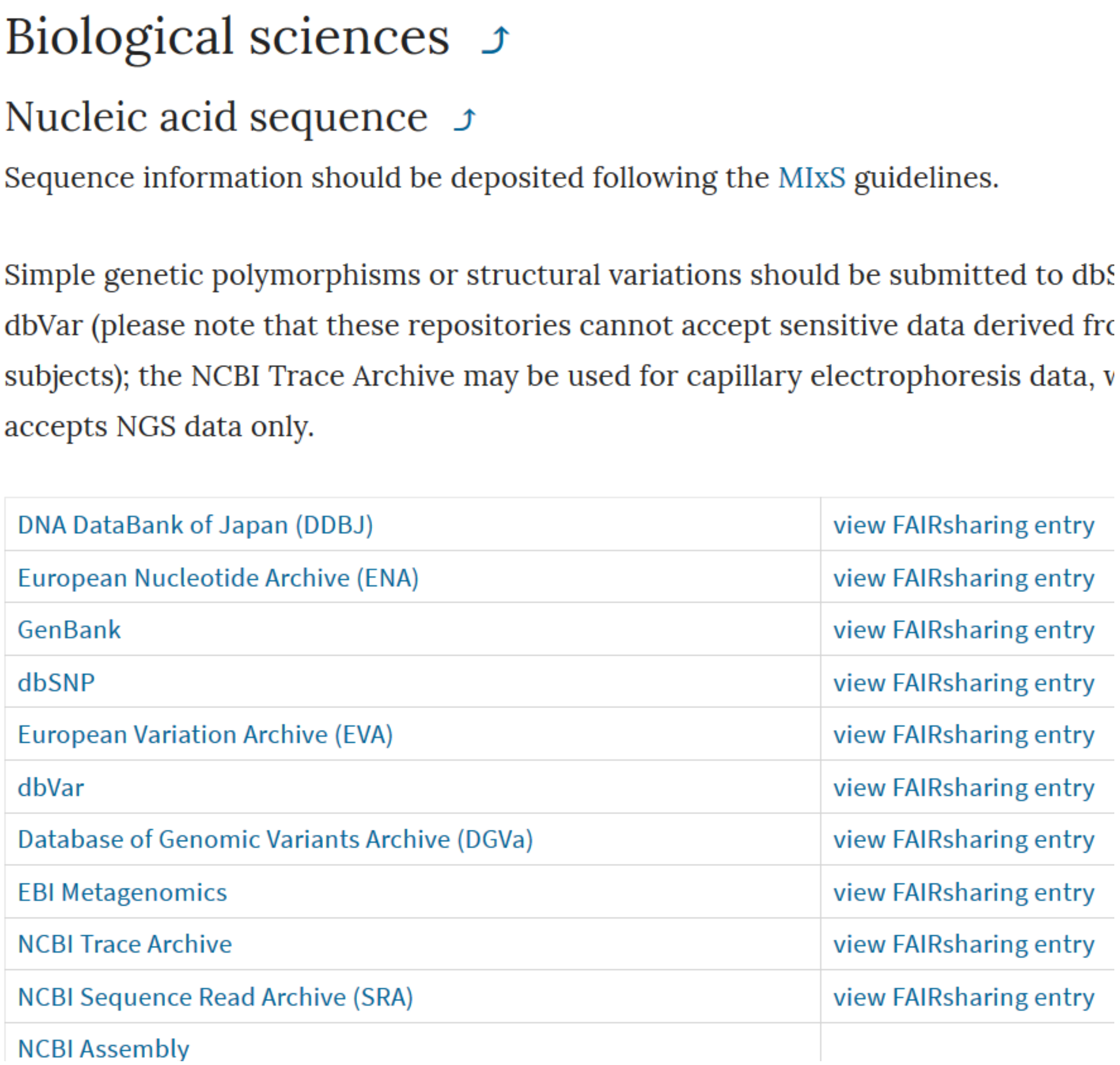
Authors must deposit their data to a recommended data repository as part of the manuscript submission process; manuscripts will not otherwise be sent for review. If data have not been deposited to a repository prior to manuscript submission, authors can upload their data to figshare or the Dryad Digital Repository during the submission process. Data may also be deposited to these resources temporarily, if the main host repository does not support confidential peer review.



We provide a date-stamped archive of our recommended repository list, which is available for use under the CC-BY licence. Recommended repositories and standards that are indexed by FAIRsharing, can be also be viewed and filtered via the *Scientific Data* FAIRsharing collection.

View data repositories

- **Biological sciences:** Nucleic acid sequence; Protein sequence; Molecular & supramolecular structure; Neuroscience; Omics; Taxonomy & species diversity; Mathematical & modelling resources; Cytometry and Immunology; Imaging; Organism-focused resources
- **Health sciences**
- **Chemistry and Chemical biology**
- **Earth, Environmental and Space sciences:** Broad scope Earth & environmental sciences; Astronomy & planetary sciences; Biogeochemistry and Geochemistry; Climate sciences; Ecology; Geomagnetism & Palaeomagnetism; Ocean sciences; Solid Earth sciences
- **Physics**
- **Materials science**
- **Social sciences**
- **Generalist repositories**
- **Other repositories**



Biological sciences

Nucleic acid sequence

Sequence information should be deposited following the MIxS guidelines.

Simple genetic polymorphisms or structural variations should be submitted to dbS dbVar (please note that these repositories cannot accept sensitive data derived fro subjects); the NCBI Trace Archive may be used for capillary electrophoresis data, w accepts NGS data only.

DNA DataBank of Japan (DDBJ)	view FAIRsharing entry
European Nucleotide Archive (ENA)	view FAIRsharing entry
GenBank	view FAIRsharing entry
dbSNP	view FAIRsharing entry
European Variation Archive (EVA)	view FAIRsharing entry
dbVar	view FAIRsharing entry
Database of Genomic Variants Archive (DGVa)	view FAIRsharing entry
EBI Metagenomics	view FAIRsharing entry
NCBI Trace Archive	view FAIRsharing entry
NCBI Sequence Read Archive (SRA)	view FAIRsharing entry
NCBI Assembly	

Omics

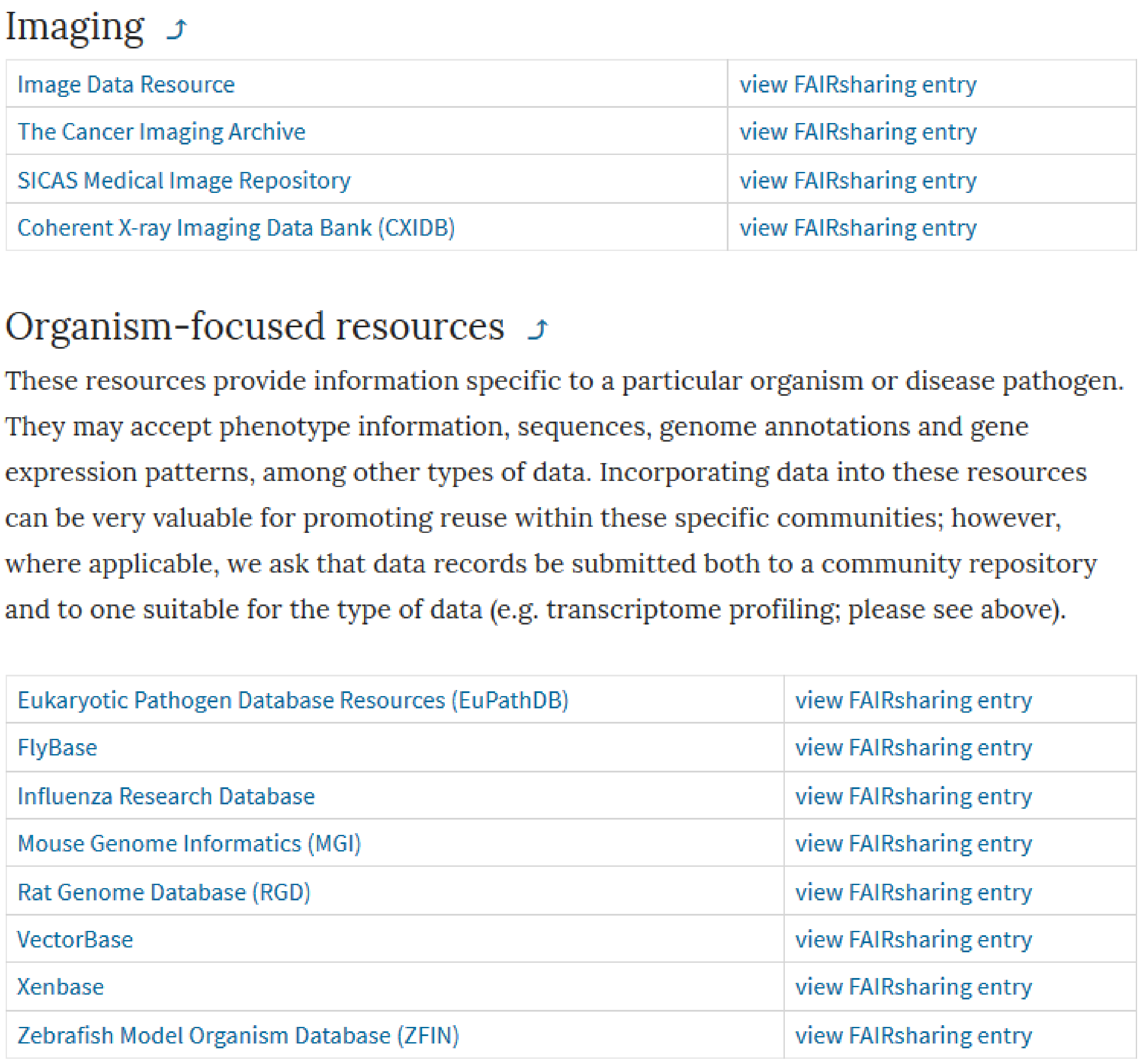
Functional genomics

Functional genomics is a broad experimental category, and *Scientific Data*'s recommendations in this discipline likewise bridge disparate research disciplines. Data should be deposited following the relevant community requirements where possible.

Please refer to the MIAME standard for microarray data. Molecular interaction data should be deposited with a member of the International Molecular Exchange Consortium (IMEx), following the MIMIx recommendations.

For data linking genotyping and phenotyping information in human subjects, we strongly recommend submission to dbGAP, EGA or JGA, which have mechanisms in place to handle sensitive data.

ArrayExpress	view FAIRsharing entry
Gene Expression Omnibus (GEO)	view FAIRsharing entry
GenomeRNAi	view FAIRsharing entry
dbGAP	view FAIRsharing entry
The European Genome-phenome Archive (EGA)	view FAIRsharing entry
Database of Interacting Proteins (DIP)	view FAIRsharing entry
IntAct	view FAIRsharing entry
Japanese Genotype-phenotype Archive (JGA)	view FAIRsharing entry
Biological General Repository for Interaction Datasets *	view FAIRsharing entry
NCBI PubChem BioAssay	view FAIRsharing entry
Genomic Expression Archive (GEA)	view FAIRsharing entry



Imaging ⤴

Image Data Resource	view FAIRsharing entry
The Cancer Imaging Archive	view FAIRsharing entry
SICAS Medical Image Repository	view FAIRsharing entry
Coherent X-ray Imaging Data Bank (CXIDB)	view FAIRsharing entry

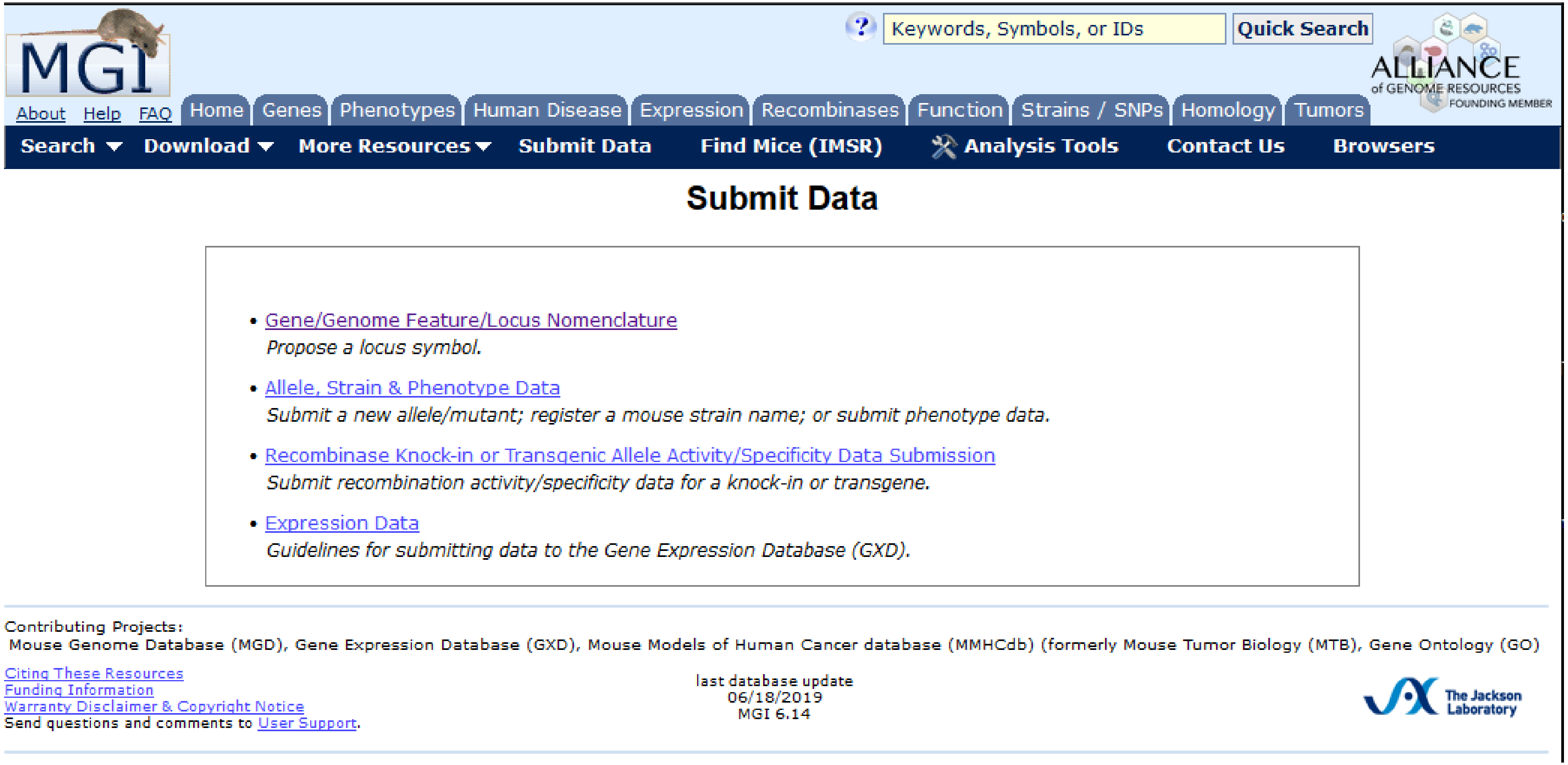
Organism-focused resources ⤴

These resources provide information specific to a particular organism or disease pathogen. They may accept phenotype information, sequences, genome annotations and gene expression patterns, among other types of data. Incorporating data into these resources can be very valuable for promoting reuse within these specific communities; however, where applicable, we ask that data records be submitted both to a community repository and to one suitable for the type of data (e.g. transcriptome profiling; please see above).

Eukaryotic Pathogen Database Resources (EuPathDB)	view FAIRsharing entry
FlyBase	view FAIRsharing entry
Influenza Research Database	view FAIRsharing entry
Mouse Genome Informatics (MGI)	view FAIRsharing entry
Rat Genome Database (RGD)	view FAIRsharing entry
VectorBase	view FAIRsharing entry
Xenbase	view FAIRsharing entry
Zebrafish Model Organism Database (ZFIN)	view FAIRsharing entry

JAX:

http://www.informatics.jax.org/submit.shtml



Submit Data

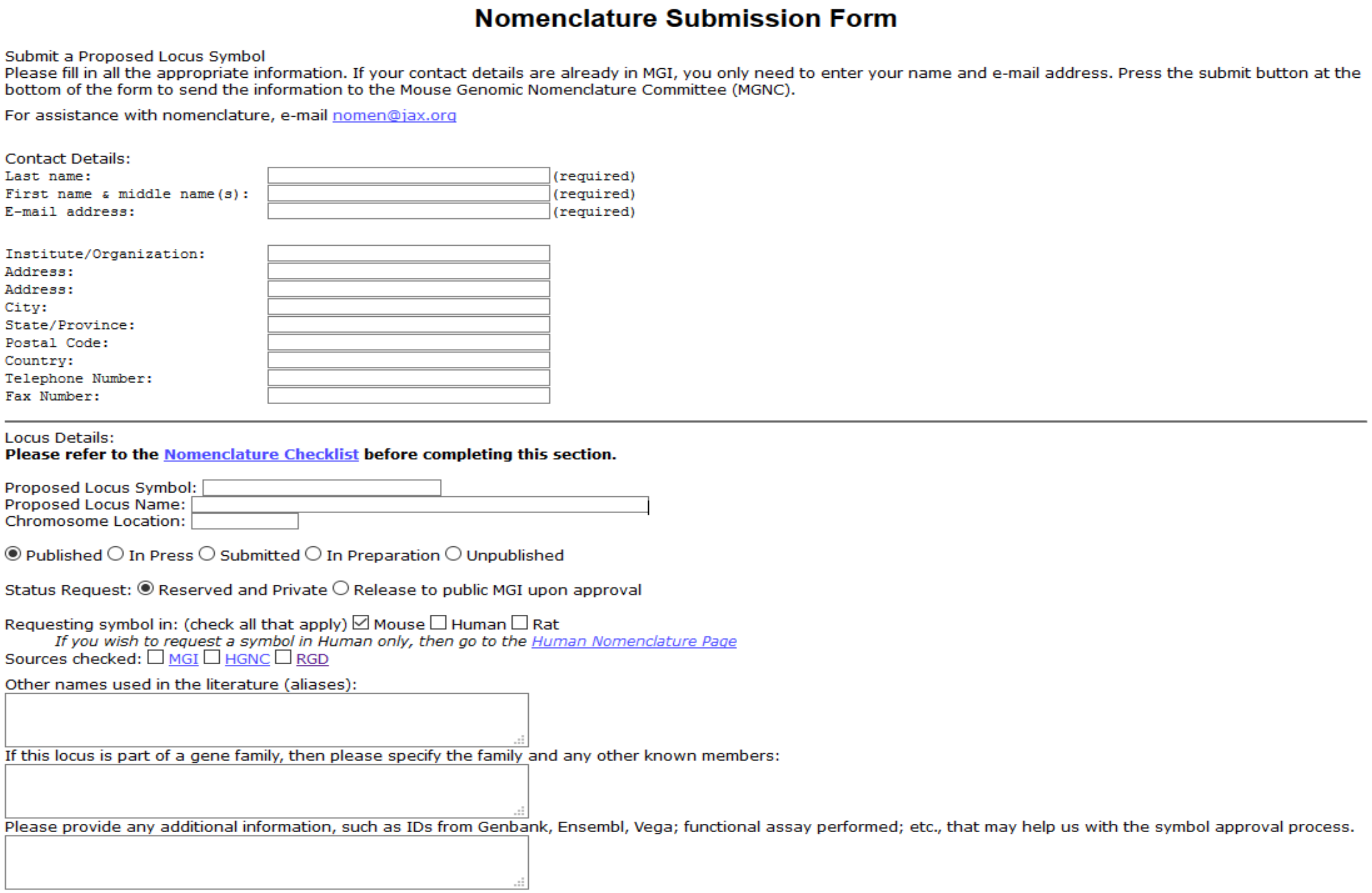
- Gene/Genome Feature/Locus Nomenclature
  Propose a locus symbol.
- Allele, Strain & Phenotype Data
  Submit a new allele/mutant; register a mouse strain name; or submit phenotype data.
- Recombinase Knock-in or Transgenic Allele Activity/Specificity Data Submission
  Submit recombination activity/specificity data for a knock-in or transgene.
- Expression Data
  Guidelines for submitting data to the Gene Expression Database (GXD).

Contributing Projects:
Mouse Genome Database (MGD), Gene Expression Database (GXD), Mouse Models of Human Cancer database (MMHCdb) (formerly Mouse Tumor Biology (MTB), Gene Ontology (GO)

Citing These Resources
Funding Information
Warranty Disclaimer & Copyright Notice
Send questions and comments to User Support.

last database update
06/18/2019
MGI 6.14

Upload Gene/Genome Feature/Locus Nomenclature:



Nomenclature Submission Form

Submit a Proposed Locus Symbol
Please fill in all the appropriate information. If your contact details are already in MGI, you only need to enter your name and e-mail address. Press the submit button at the bottom of the form to send the information to the Mouse Genomic Nomenclature Committee (MGNC).

For assistance with nomenclature, e-mail nomen@jax.org

Contact Details:
Last name: (required)
First name & middle name(s): (required)
E-mail address: (required)

Institute/Organization:
Address:
Address:
City:
State/Province:
Postal Code:
Country:
Telephone Number:
Fax Number:

Locus Details:
Please refer to the Nomenclature Checklist before completing this section.

Proposed Locus Symbol:
Proposed Locus Name:
Chromosome Location:

● Published ○ In Press ○ Submitted ○ In Preparation ○ Unpublished

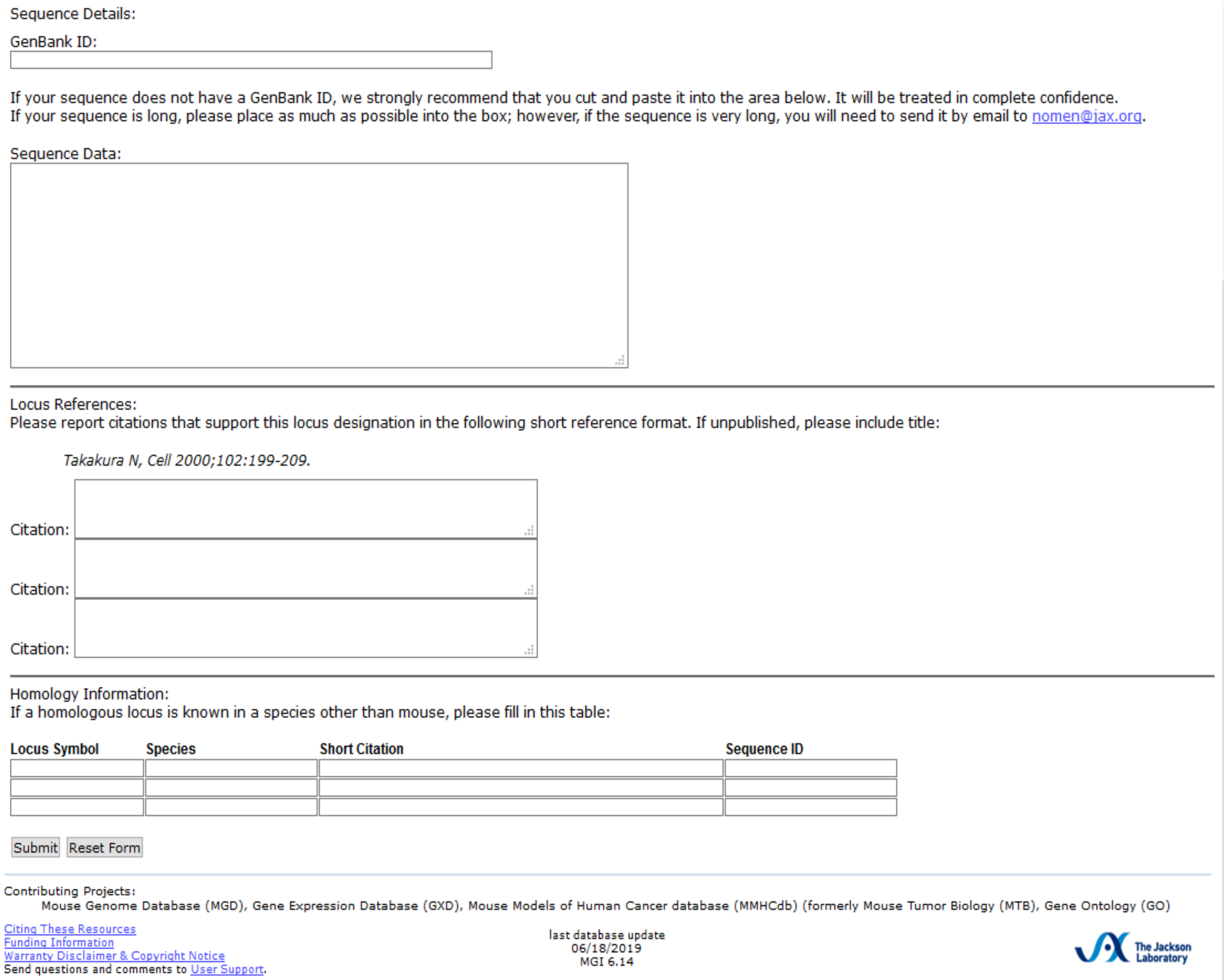
Status Request: ● Reserved and Private ○ Release to public MGI upon approval

Requesting symbol in: (check all that apply) ☑ Mouse □ Human □ Rat
If you wish to request a symbol in Human only, then go to the Human Nomenclature Page
Sources checked: □ MGI □ HGNC □ RGD

Other names used in the literature (aliases):

If this locus is part of a gene family, then please specify the family and any other known members:

Please provide any additional information, such as IDs from Genbank, Ensembl, Vega; functional assay performed; etc., that may help us with the symbol approval process.



Sequence Details:

GenBank ID:

If your sequence does not have a GenBank ID, we strongly recommend that you cut and paste it into the area below. It will be treated in complete confidence.
If your sequence is long, please place as much as possible into the box; however, if the sequence is very long, you will need to send it by email to nomen@jax.org.

Sequence Data:

Locus References:
Please report citations that support this locus designation in the following short reference format. If unpublished, please include title:

Takakura N, Cell 2000;102:199-209.

Citation:

Citation:

Citation:

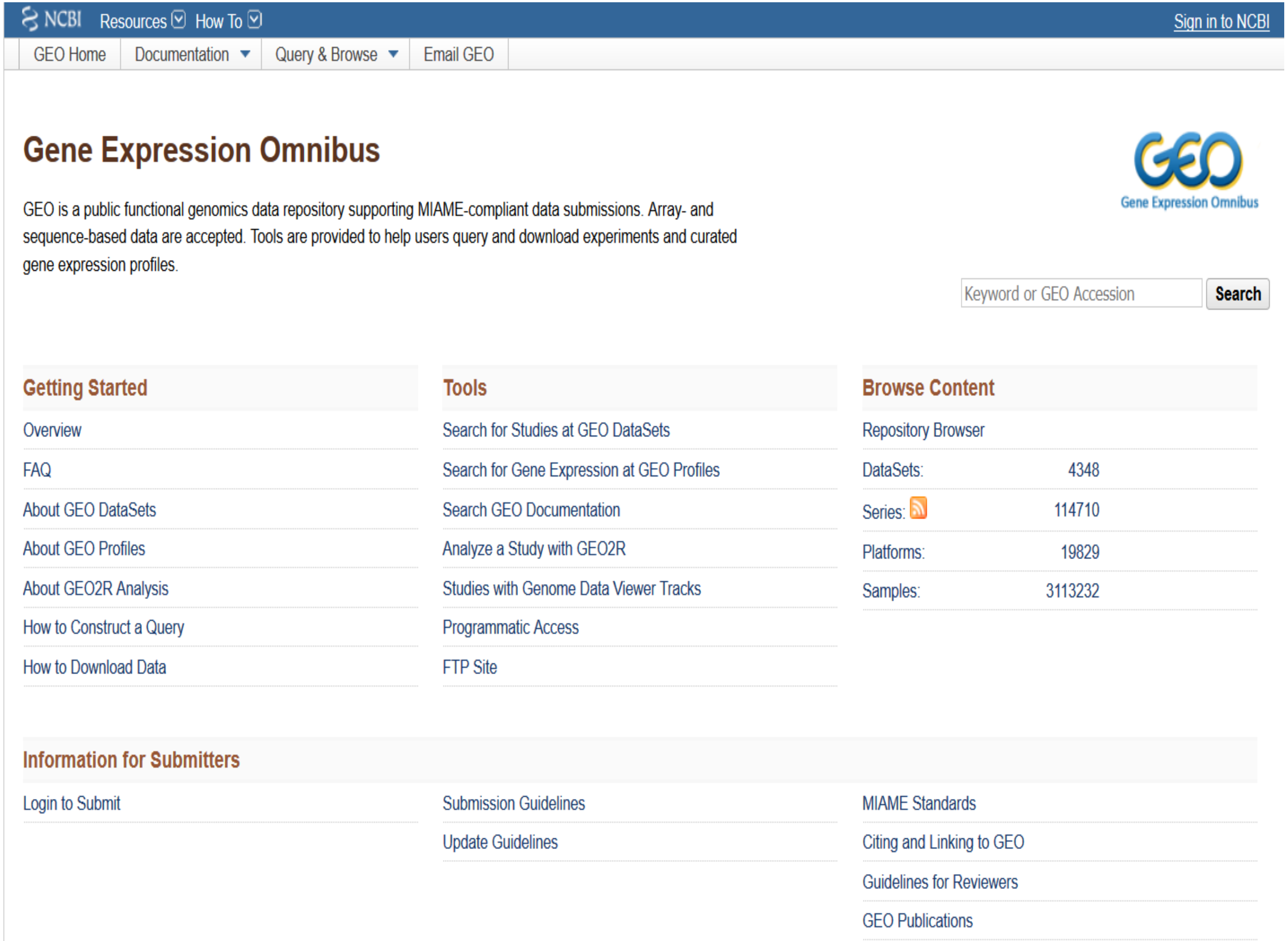
Homology Information:
If a homologous locus is known in a species other than mouse, please fill in this table:

Locus Symbol	Species	Short Citation	Sequence ID

Contributing Projects:
Mouse Genome Database (MGD), Gene Expression Database (GXD), Mouse Models of Human Cancer database (MMHCdb) (formerly Mouse Tumor Biology (MTB), Gene Ontology (GO)

Citing These Resources
Funding Information
Warranty Disclaimer & Copyright Notice
Send questions and comments to User Support.

last database update
06/18/2019
MGI 6.14



Gene Expression Omnibus

GEO is a public functional genomics data repository supporting MIAME-compliant data submissions. Array- and sequence-based data are accepted. Tools are provided to help users query and download experiments and curated gene expression profiles.

Getting Started

- Overview
- FAQ
- About GEO DataSets
- About GEO Profiles
- About GEO2R Analysis
- How to Construct a Query
- How to Download Data

Tools

- Search for Studies at GEO DataSets
- Search for Gene Expression at GEO Profiles
- Search GEO Documentation
- Analyze a Study with GEO2R
- Studies with Genome Data Viewer Tracks
- Programmatic Access
- FTP Site

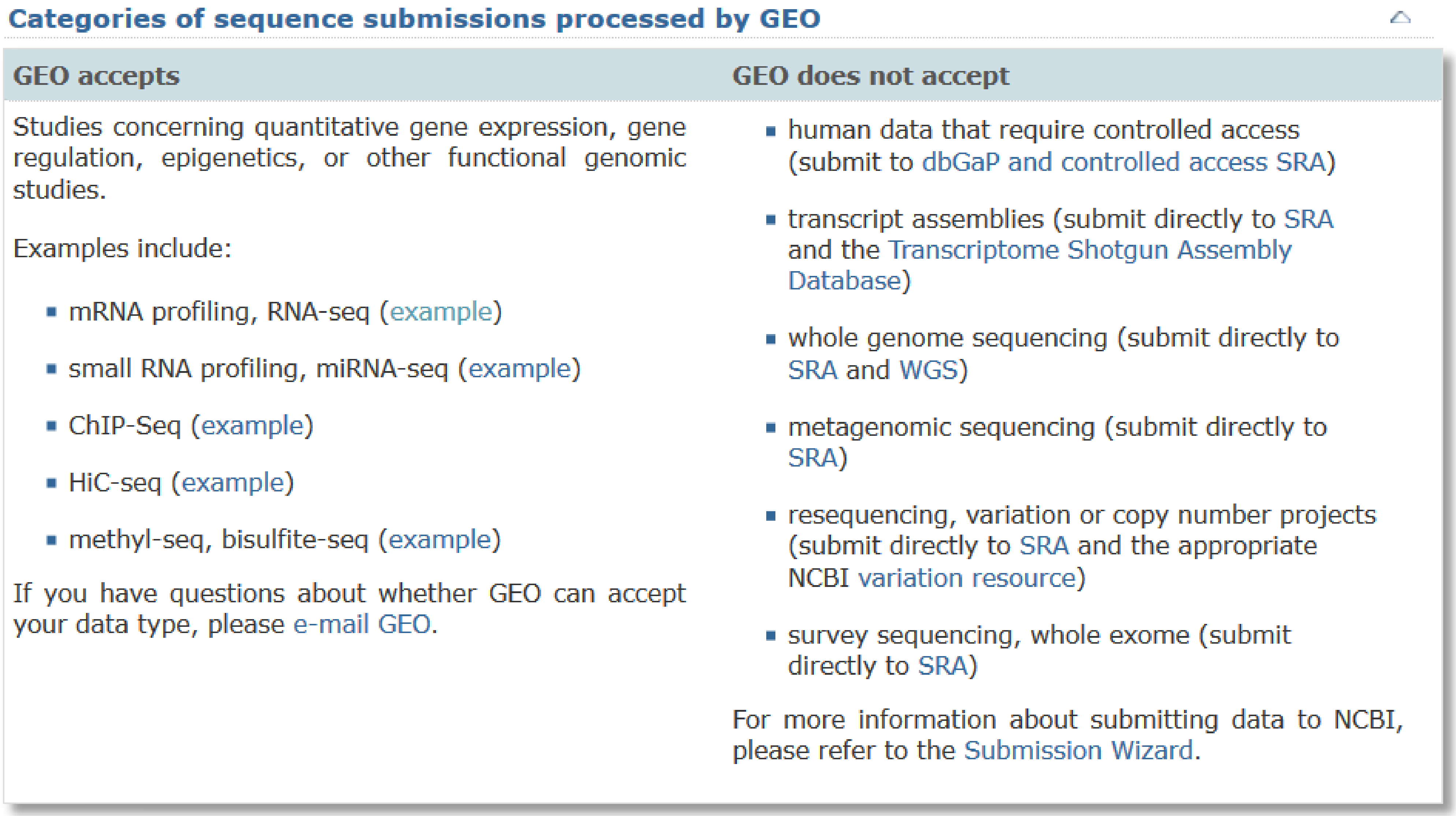
Browse Content

Repository Browser

DataSets:	4348
Series:	114710
Platforms:	19829
Samples:	3113232

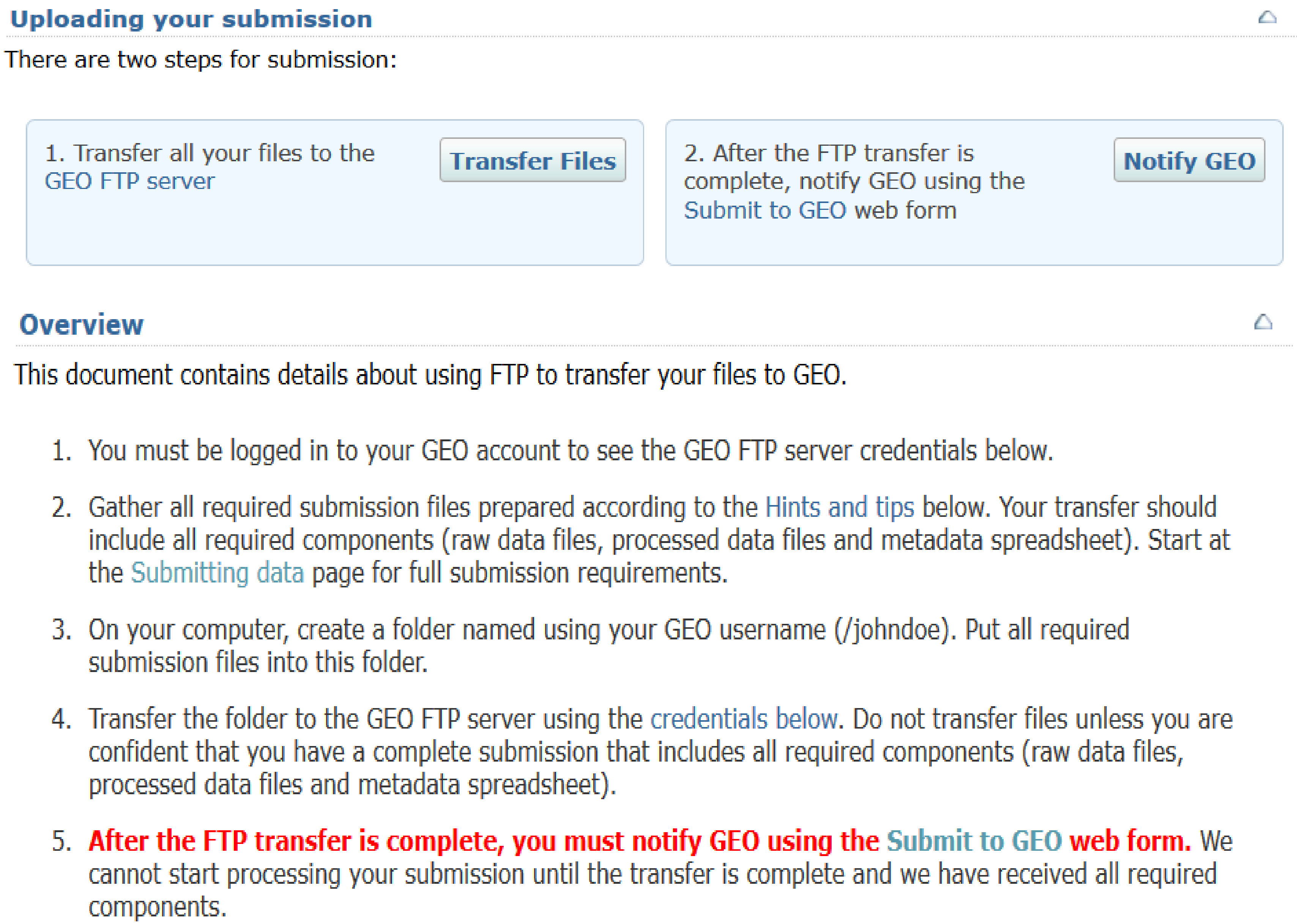
Information for Submitters

- Login to Submit
- Submission Guidelines
- Update Guidelines
- MIAME Standards
- Citing and Linking to GEO
- Guidelines for Reviewers
- GEO Publications



Categories of sequence submissions processed by GEO

GEO accepts	GEO does not accept
Studies concerning quantitative gene expression, gene regulation, epigenetics, or other functional genomic studies. Examples include: ▪ mRNA profiling, RNA-seq (example) ▪ small RNA profiling, miRNA-seq (example) ▪ ChIP-Seq (example) ▪ HiC-seq (example) ▪ methyl-seq, bisulfite-seq (example) If you have questions about whether GEO can accept your data type, please e-mail GEO.	▪ human data that require controlled access (submit to dbGaP and controlled access SRA) ▪ transcript assemblies (submit directly to SRA and the Transcriptome Shotgun Assembly Database) ▪ whole genome sequencing (submit directly to SRA and WGS) ▪ metagenomic sequencing (submit directly to SRA) ▪ resequencing, variation or copy number projects (submit directly to SRA and the appropriate NCBI variation resource) ▪ survey sequencing, whole exome (submit directly to SRA) For more information about submitting data to NCBI, please refer to the Submission Wizard.



Uploading your submission

There are two steps for submission:

1. Transfer all your files to the GEO FTP server

Transfer Files

2. After the FTP transfer is complete, notify GEO using the Submit to GEO web form

Notify GEO

Overview

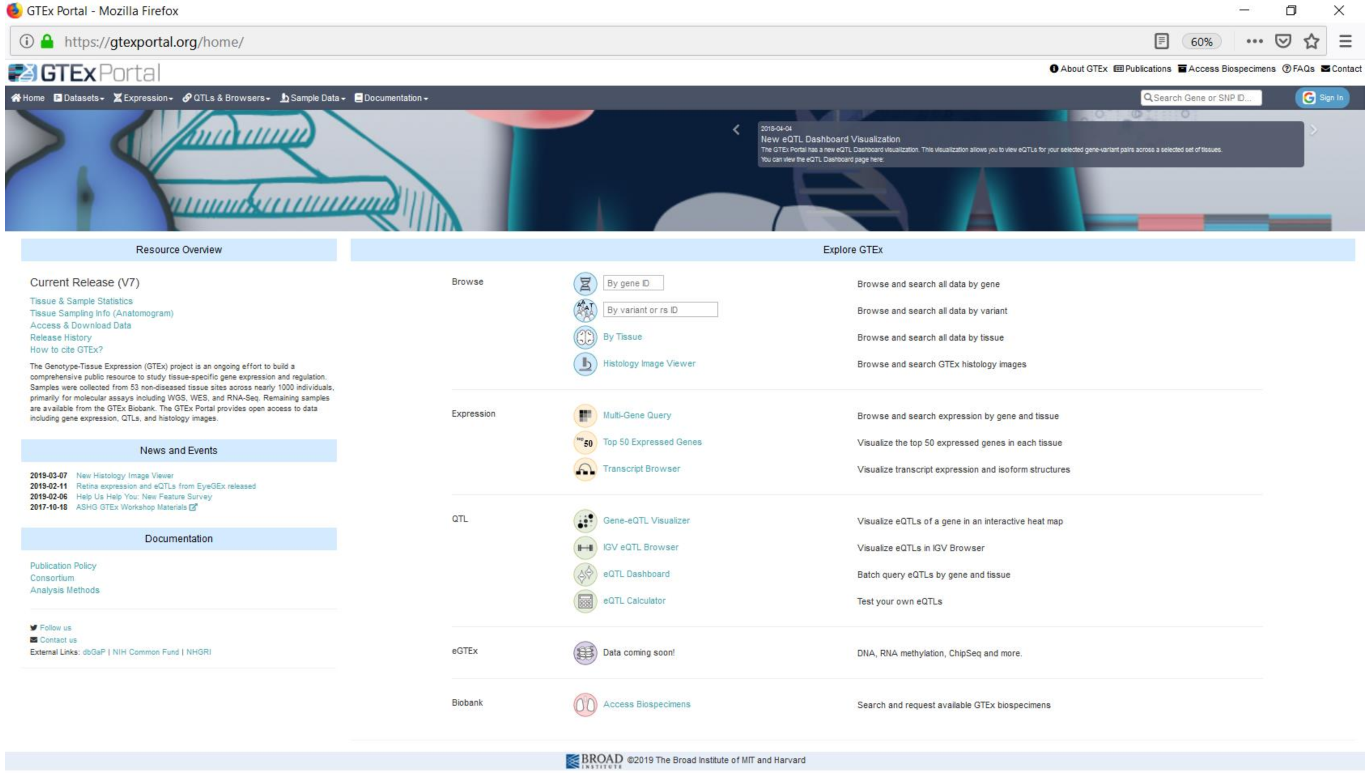
This document contains details about using FTP to transfer your files to GEO.

1. You must be logged in to your GEO account to see the GEO FTP server credentials below.
2. Gather all required submission files prepared according to the Hints and tips below. Your transfer should include all required components (raw data files, processed data files and metadata spreadsheet). Start at the Submitting data page for full submission requirements.
3. On your computer, create a folder named using your GEO username (/johndoe). Put all required submission files into this folder.
4. Transfer the folder to the GEO FTP server using the credentials below. Do not transfer files unless you are confident that you have a complete submission that includes all required components (raw data files, processed data files and metadata spreadsheet).
5. **After the FTP transfer is complete, you must notify GEO using the Submit to GEO web form.** We cannot start processing your submission until the transfer is complete and we have received all required components.

Genotype-Tissue Expression (GTEx) project

https://gtexportal.org/home/

The Genotype-Tissue Expression (GTEx) project is an ongoing effort to build a comprehensive public resource to study tissue-specific gene expression and regulation. Samples were collected from 53 non-diseased tissue sites across nearly 1000 individuals, primarily for molecular assays including WGS, WES, and RNA-Seq. Remaining samples are available from the GTEx Biobank. The GTEx Portal provides open access to data including gene expression, QTLs, and histology images.



2018-04-04
New eQTL Dashboard Visualization
The GTEx Portal has a new eQTL Dashboard visualization. This visualization allows you to view eQTLs for your selected gene-variant pairs across a selected set of tissues.
You can view the eQTL Dashboard page here:

Resource Overview

Current Release (V7)

Tissue & Sample Statistics
Tissue Sampling Info (Anatomogram)
Access & Download Data
Release History
How to cite GTEx?

The Genotype-Tissue Expression (GTEx) project is an ongoing effort to build a comprehensive public resource to study tissue-specific gene expression and regulation. Samples were collected from 53 non-diseased tissue sites across nearly 1000 individuals, primarily for molecular assays including WGS, WES, and RNA-Seq. Remaining samples are available from the GTEx Biobank. The GTEx Portal provides open access to data including gene expression, QTLs, and histology images.

News and Events

2019-03-07	New Histology Image Viewer
2019-02-11	Retina expression and eQTLs from EyeGEx released
2019-02-06	Help Us Help You: New Feature Survey
2017-10-18	ASHG GTEx Workshop Materials

Documentation

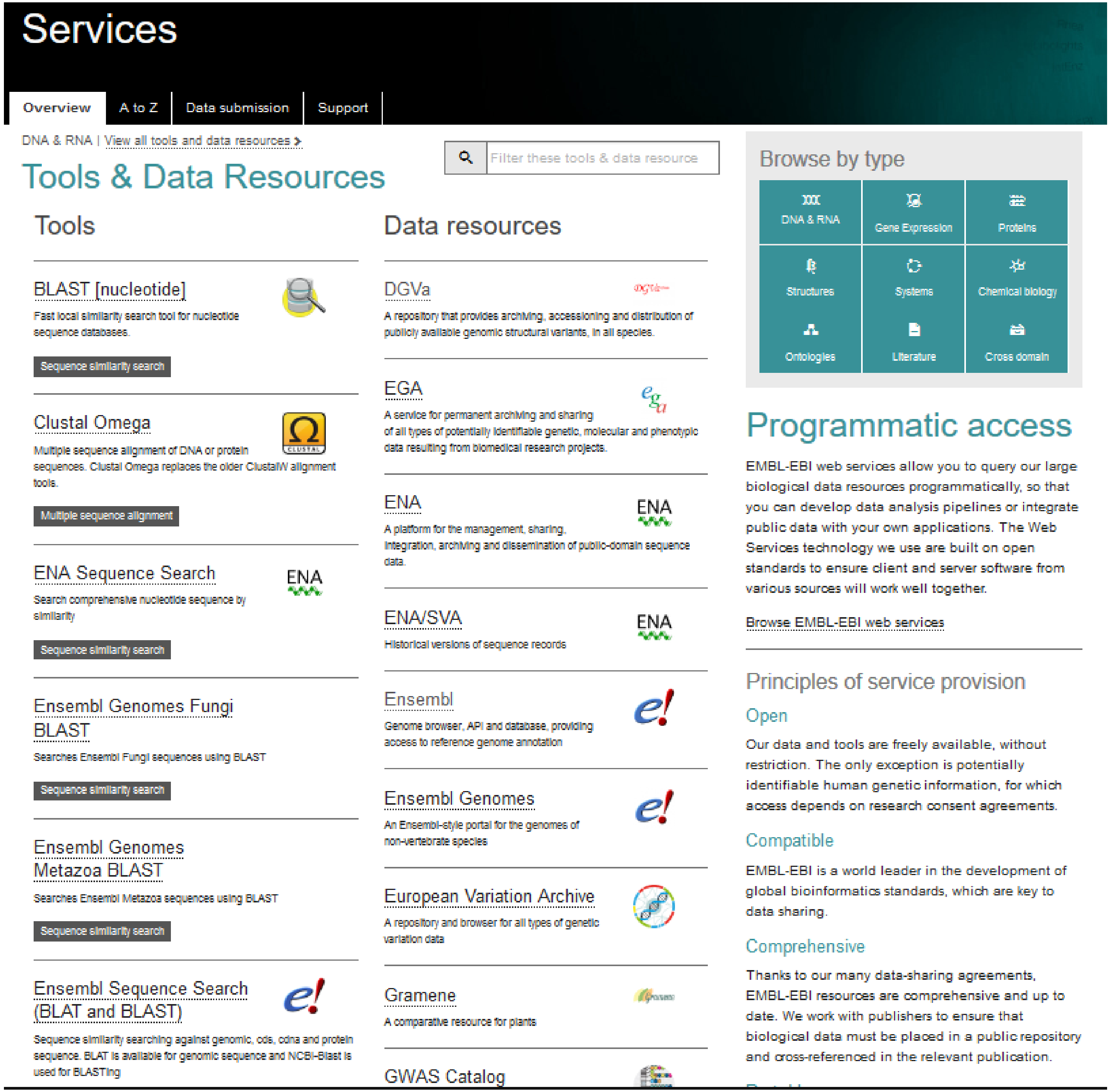
Publication Policy
Consortium
Analysis Methods

Follow us
Contact us
External Links: dbGaP | NIH Common Fund | NHGRI

Explore GTEx

Browse	By gene ID	Browse and search all data by gene
	By variant or rs ID	Browse and search all data by variant
	By Tissue	Browse and search all data by tissue
	Histology Image Viewer	Browse and search GTEx histology images
Expression	Multi-Gene Query	Browse and search expression by gene and tissue
	Top 50 Expressed Genes	Visualize the top 50 expressed genes in each tissue
	Transcript Browser	Visualize transcript expression and isoform structures
QTL	Gene-eQTL Visualizer	Visualize eQTLs of a gene in an interactive heat map
	IGV eQTL Browser	Visualize eQTLs in IGV Browser
	eQTL Dashboard	Batch query eQTLs by gene and tissue
	eQTL Calculator	Test your own eQTLs
eGTEx	Data coming soon!	DNA, RNA methylation, ChipSeq and more.
Biobank	Access Biospecimens	Search and request available GTEx biospecimens

BROAD INSTITUTE ©2019 The Broad Institute of MIT and Harvard



Services

DNA & RNA | View all tools and data resources ›

Tools & Data Resources

Tools

BLAST [nucleotide]

Fast local similarity search tool for nucleotide sequence databases.

Sequence similarity search

Clustal Omega

Multiple sequence alignment of DNA or protein sequences. Clustal Omega replaces the older ClustalW alignment tools.

Multiple sequence alignment

ENA Sequence Search

Search comprehensive nucleotide sequence by similarity

Sequence similarity search

Ensembl Genomes Fungi BLAST

Searches Ensembl Fungi sequences using BLAST

Sequence similarity search

Ensembl Genomes Metazoa BLAST

Searches Ensembl Metazoa sequences using BLAST

Sequence similarity search

Ensembl Sequence Search (BLAT and BLAST)

Sequence similarity searching against genomic, cds, cdna and protein sequence. BLAT is available for genomic sequence and NCBI-Blast is used for BLASTing

Data resources

DGVa

A repository that provides archiving, accessioning and distribution of publicly available genomic structural variants, in all species.

EGA

A service for permanent archiving and sharing of all types of potentially identifiable genetic, molecular and phenotypic data resulting from biomedical research projects.

ENA

A platform for the management, sharing, integration, archiving and dissemination of public-domain sequence data.

ENA/SVA

Historical versions of sequence records

Ensembl

Genome browser, API and database, providing access to reference genome annotation

Ensembl Genomes

An Ensembl-style portal for the genomes of non-vertebrate species

European Variation Archive

A repository and browser for all types of genetic variation data

Gramene

A comparative resource for plants

GWAS Catalog

Browse by type

Programmatic access

EMBL-EBI web services allow you to query our large biological data resources programmatically, so that you can develop data analysis pipelines or integrate public data with your own applications. The Web Services technology we use are built on open standards to ensure client and server software from various sources will work well together.

Browse EMBL-EBI web services

Principles of service provision

Open

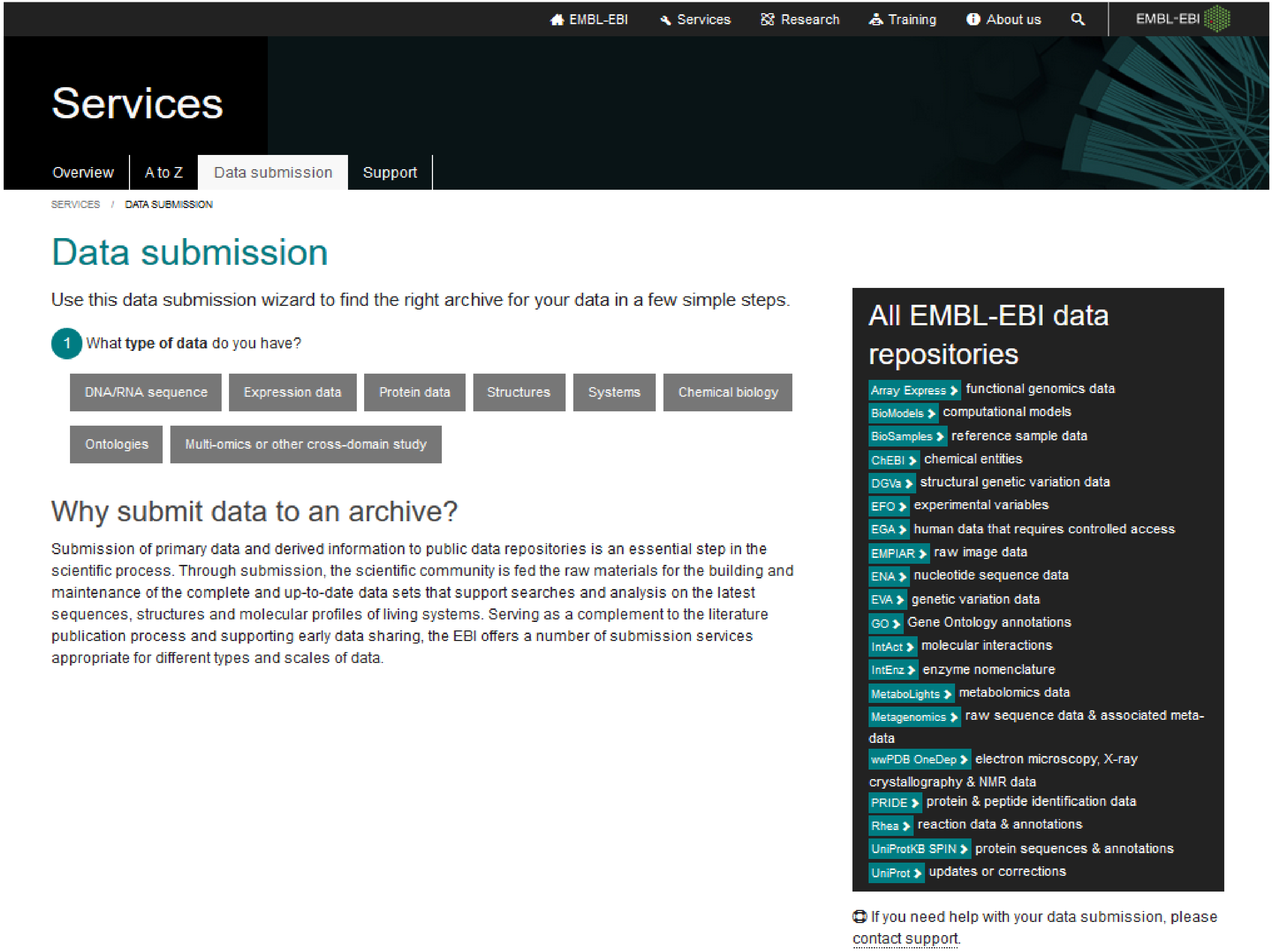
Our data and tools are freely available, without restriction. The only exception is potentially identifiable human genetic information, for which access depends on research consent agreements.

Compatible

EMBL-EBI is a world leader in the development of global bioinformatics standards, which are key to data sharing.

Comprehensive

Thanks to our many data-sharing agreements, EMBL-EBI resources are comprehensive and up to date. We work with publishers to ensure that biological data must be placed in a public repository and cross-referenced in the relevant publication.



Services

Data submission

Use this data submission wizard to find the right archive for your data in a few simple steps.

1 What **type of data** do you have?

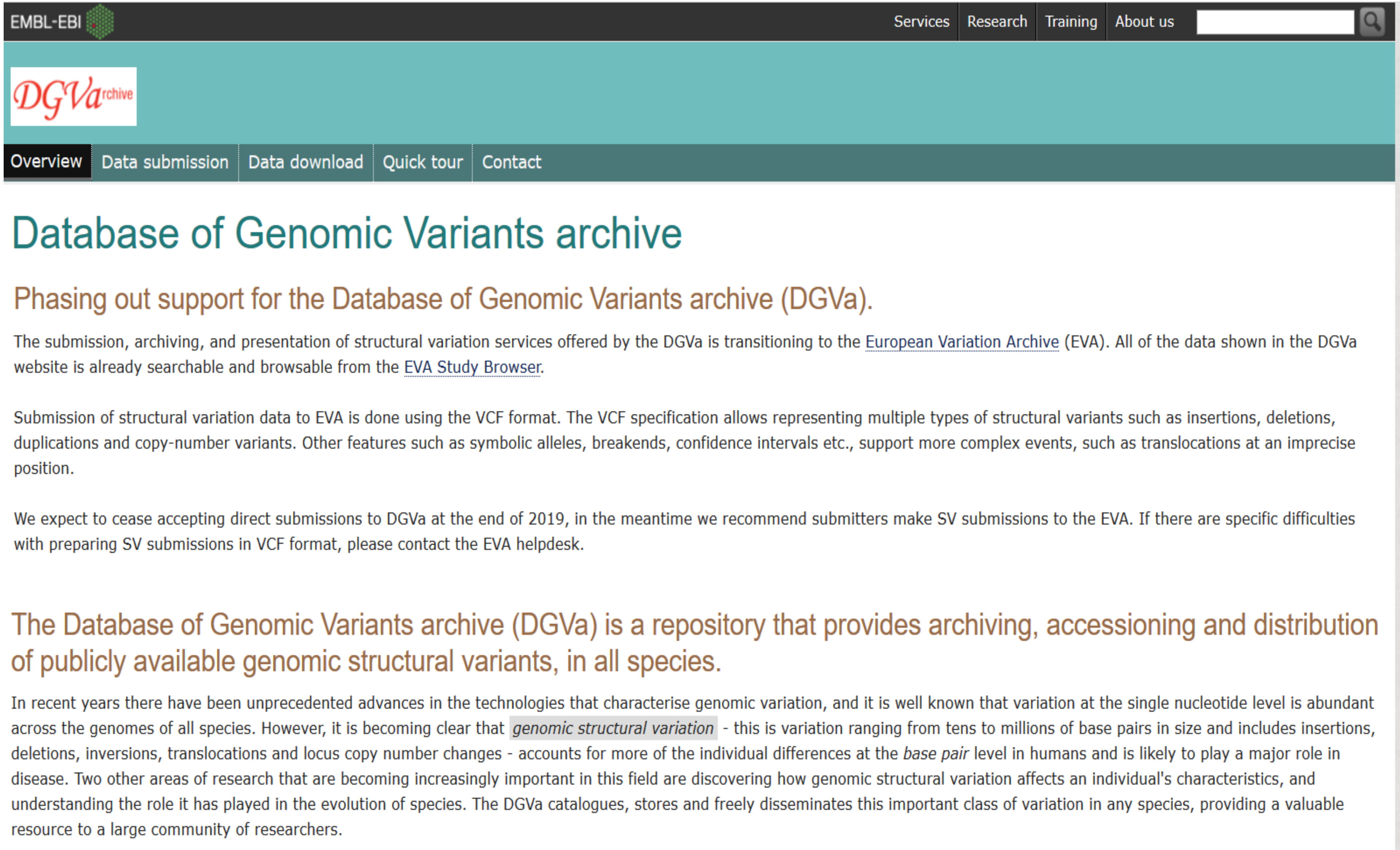
Why submit data to an archive?

Submission of primary data and derived information to public data repositories is an essential step in the scientific process. Through submission, the scientific community is fed the raw materials for the building and maintenance of the complete and up-to-date data sets that support searches and analysis on the latest sequences, structures and molecular profiles of living systems. Serving as a complement to the literature publication process and supporting early data sharing, the EBI offers a number of submission services appropriate for different types and scales of data.

All EMBL-EBI data repositories

- Array Express functional genomics data
- BioModels computational models
- BioSamples reference sample data
- ChEBI chemical entities
- DGVa structural genetic variation data
- EFO experimental variables
- EGA human data that requires controlled access
- EMPIAR raw image data
- ENA nucleotide sequence data
- EVA genetic variation data
- GO Gene Ontology annotations
- IntAct molecular interactions
- IntEnz enzyme nomenclature
- MetaboLights metabolomics data
- Metagenomics raw sequence data & associated meta-data
- wwPDB OneDep electron microscopy, X-ray crystallography & NMR data
- PRIDE protein & peptide identification data
- Rhea reaction data & annotations
- UniProtKB SPIN protein sequences & annotations
- UniProt updates or corrections

If you need help with your data submission, please contact support.



Database of Genomic Variants archive

Phasing out support for the Database of Genomic Variants archive (DGVa).

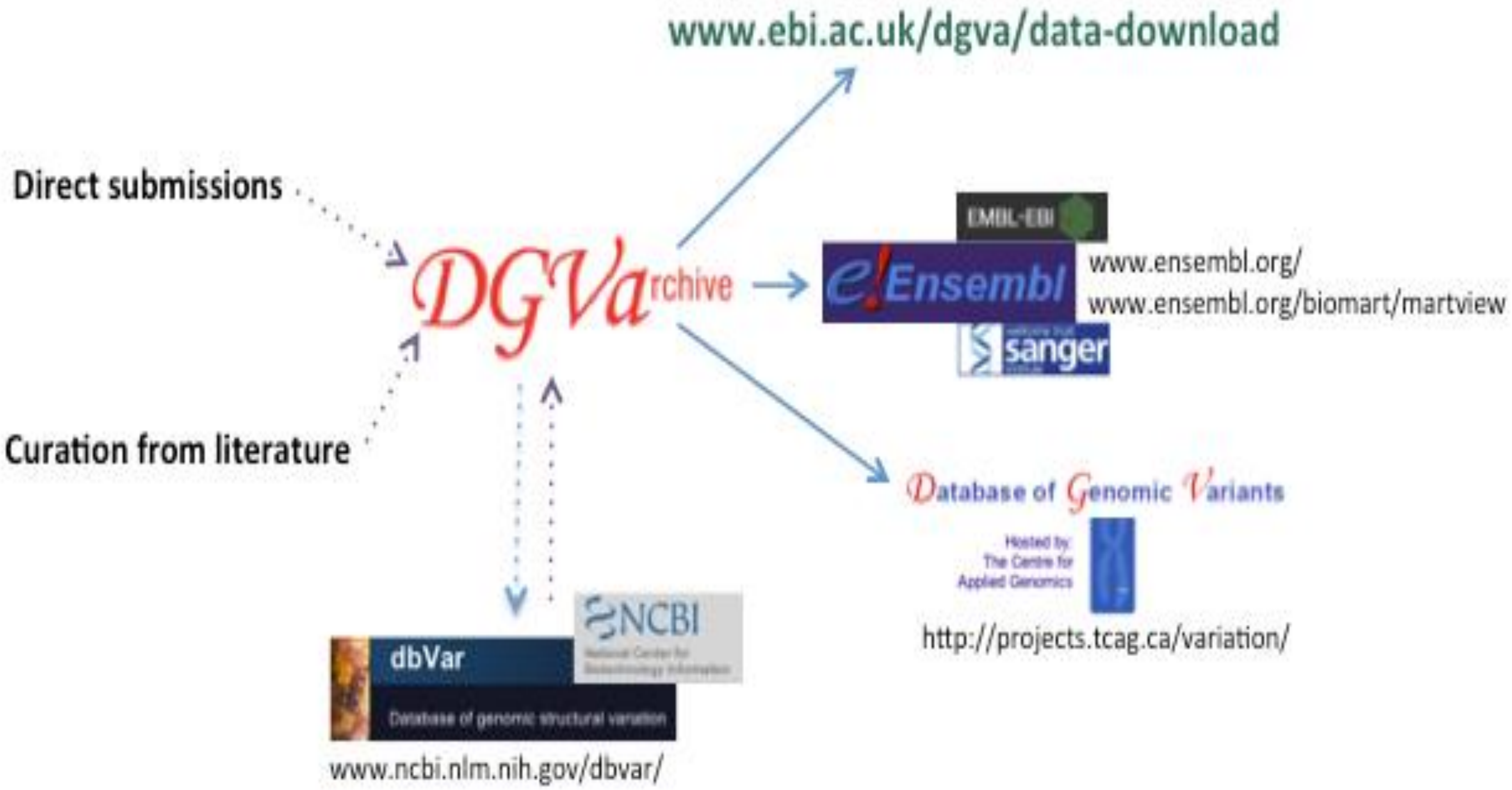
The submission, archiving, and presentation of structural variation services offered by the DGVa is transitioning to the European Variation Archive (EVA). All of the data shown in the DGVa website is already searchable and browsable from the EVA Study Browser.

Submission of structural variation data to EVA is done using the VCF format. The VCF specification allows representing multiple types of structural variants such as insertions, deletions, duplications and copy-number variants. Other features such as symbolic alleles, breakends, confidence intervals etc., support more complex events, such as translocations at an imprecise position.

We expect to cease accepting direct submissions to DGVa at the end of 2019, in the meantime we recommend submitters make SV submissions to the EVA. If there are specific difficulties with preparing SV submissions in VCF format, please contact the EVA helpdesk.

The Database of Genomic Variants archive (DGVa) is a repository that provides archiving, accessioning and distribution of publicly available genomic structural variants, in all species.

In recent years there have been unprecedented advances in the technologies that characterise genomic variation, and it is well known that variation at the single nucleotide level is abundant across the genomes of all species. However, it is becoming clear that *genomic structural variation* - this is variation ranging from tens to millions of base pairs in size and includes insertions, deletions, inversions, translocations and locus copy number changes - accounts for more of the individual differences at the *base pair* level in humans and is likely to play a major role in disease. Two other areas of research that are becoming increasingly important in this field are discovering how genomic structural variation affects an individual's characteristics, and understanding the role it has played in the evolution of species. The DGVa catalogues, stores and freely disseminates this important class of variation in any species, providing a valuable resource to a large community of researchers.



www.ebi.ac.uk/dgva/data-download
Direct submissions
DGVarchive
Curation from literature
EMBL-EBI
e!Ensembl
www.ensembl.org/
www.ensembl.org/biomart/martview
sanger
Database of Genomic Variants
Hosted by: The Centre for Applied Genomics
http://projects.tcag.ca/variation/
NCBI
National Center for Biotechnology Information
dbVar
Database of genomic structural variation
www.ncbi.nlm.nih.gov/dbvar/

Genomic repositories to upload data

- Gene Expression Omnibus GEO, NCBI
  https://www.ncbi.nlm.nih.gov/geo/

- European Bioinformatics Institute (EMBL-EBI)
  https://www.ebi.ac.uk/services

Why submit data to an archive?

Submission of primary data and derived information to public data repositories is an essential step in the scientific process.
Through submission, the scientific community is fed the raw materials for the building
and maintenance of the complete and up-to-date data sets that support searches and analysis
on the latest sequences, structures and molecular profiles of living systems.
Serving as a complement to the literature publication process and supporting early data sharing,
the EBI offers a number of submission services appropriate for different types and scales of data.

EMBL-EBI Data submission

Data submission

Use this data submission wizard to find the right archive for your data in a few simple steps.

1 You have **expression data**

2 Your data **do not** require controlled access

3 You have **microarray or RNA-seq gene expression** related data

You can submit your data to the following database:

Array Express

GEO Data submission I

Data types:

1. microarray
2. high-throughput sequencing
3. other (includes NanoString, RT-PCR, traditional SAGE).

If you are submitting human data, it is your responsibility to comply with Human Subject Guidelines.

GEO Data submission II

- GEO supports various submission formats:
    - GEOarchive spreadsheet submissions are recommended for most submitters.
    - If your data and metadata are already in a database, and you can generate and export data in SOFT plain text or MINiML XML, you can use the GEO Direct deposit form to submit data.
- GEO accession numbers are normally approved within 5 business days after completion of submission.
- Your GEO submissions can remain private until a manuscript citing the data is published.
- You can allow reviewers anonymous access to your private records.
- You can update or edit your existing GEO records at any time.

GEO Data submission III

Categories of sequence submissions processed by GEO

GEO accepts:

Studies concerning quantitative gene expression, gene regulation, epigenetics, or other functional genomic studies.

Examples include:

•mRNA profiling, RNA-seq

•small RNA profiling, miRNA-seq

•ChIP-Seq

•HiC-seq

•methyl-seq, bisulfite-seq

GEO does not accept:

•human data that require controlled access (submit to dbGaP and controlled access SRA)

•transcript assemblies (submit to SRA and the Transcriptome Shotgun Assembly Database)

•whole genome sequencing (submit to SRA and WGS)

•metagenomic sequencing (submit to SRA)

•resequencing, variation or copy number projects (submit to SRA and the appropriate NCBI variation resource)

•survey sequencing, whole exome (submit to SRA)

GEO Data submission IV

There are three required components for the spreadsheet-based submission method:

1. a metadata spreadsheet (excel sheet)

2. processed data files (e. g. read count files)

3. raw data files (e.g. fastq files)

GEO Data submission V

1. **a metadata spreadsheet**
   Metadata refers to descriptive information about the overall study, individual samples, all protocols, and references to processed and raw data file names.

2. **processed data files**
   Processed data are a required part of GEO submissions. The final processed data are defined as the data on which the conclusions in the related manuscript are based. We do not expect standard alignment files (e.g., BAM, SAM, BED) as processed data since conclusions are expected to be based on further-processed data.

   - quantitative data for features of interest (genes, transcripts, exons, miRNA)
     a) raw counts of sequencing reads for the features of interest, and/or
     b) normalized abundance measurements, e.g., output from Cuffdiff, DESeq, edgeR, etc
   - ChIP-Seq data might include peak files with quantitative data, tag density files, etc. (WIG, bedGraph)

GEO Data submission VI

3. **Raw data files**

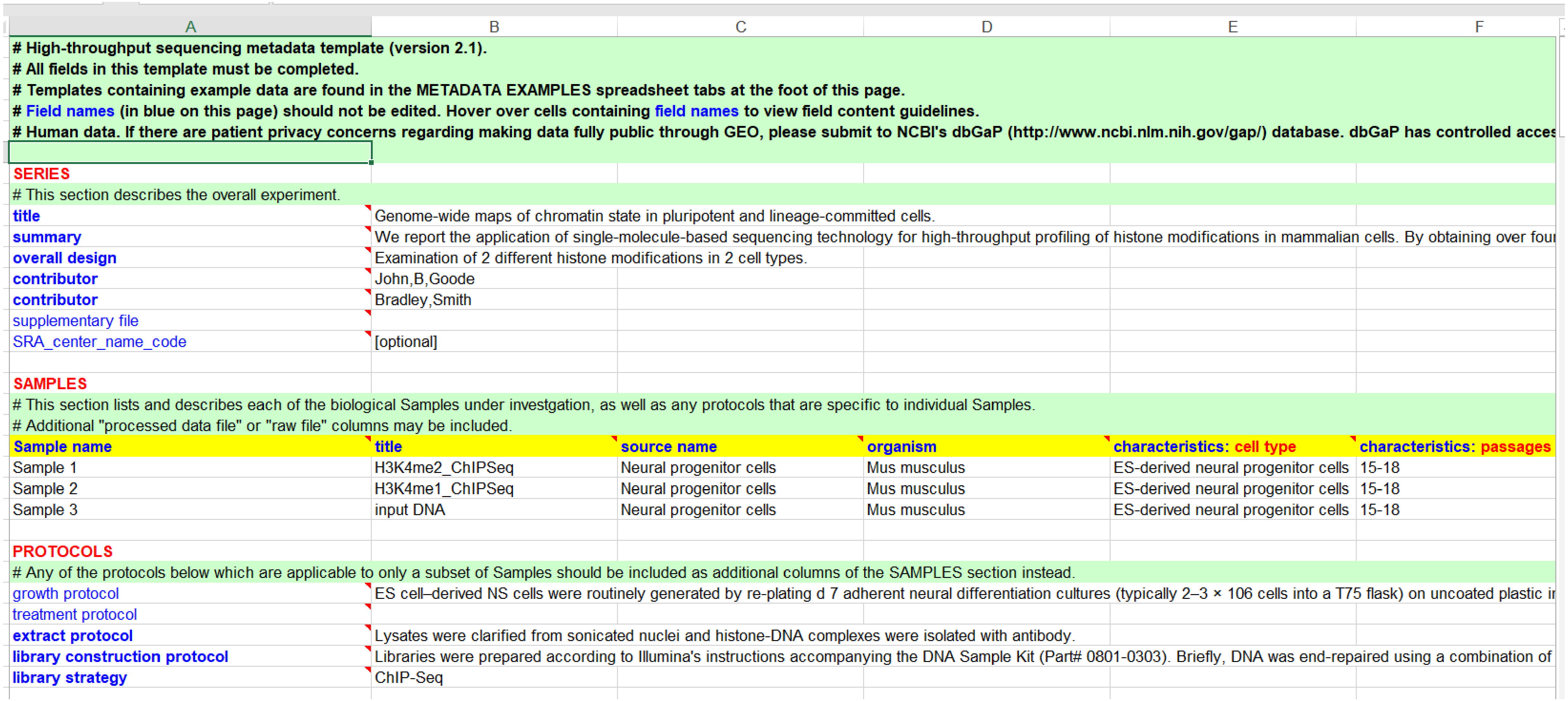
   The raw data files should be the original demultiplexed files containing reads and quality scores, as generated by the sequencing instrument so that each barcoded sample ends up with a dedicated run file (e. g. fastq-format).

   MD5 Checksums: We recommend that submitters provide MD5 checksums for their raw data files. The checksums are used to verify file integrity. Checksums can be calculated using the following methods: **Unix**: md5sum <file>
   OS X: md5 <file>
   Windows: Application required. Many are available for free download.

   Data File Compression: Individual files can be compressed to speed transfer, but this is not required. Acceptable compression formats are gzip and bzip2 (i.e. files ending with a .gz or .bz2 extension). Never compress binary files (e.g., BAM, bigWig, bigBed), and DO NOT upload ZIP archives (files with a .zip extension).

GEO Data submission VII

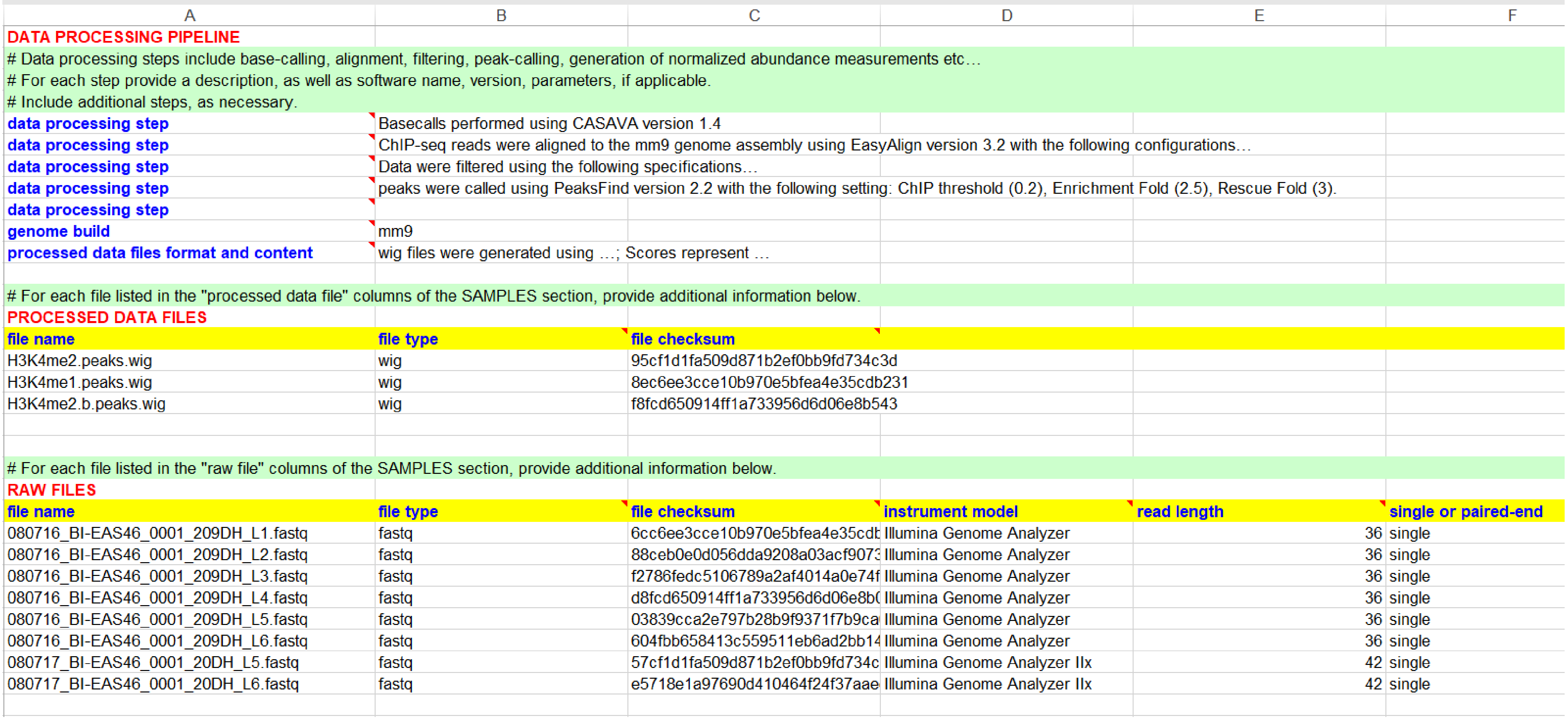
Metadata spreadsheet example I



A	B	C	D	E	F
# High-throughput sequencing metadata template (version 2.1).					
# All fields in this template must be completed.					
# Templates containing example data are found in the METADATA EXAMPLES spreadsheet tabs at the foot of this page.					
# Field names (in blue on this page) should not be edited. Hover over cells containing field names to view field content guidelines.					
# Human data. If there are patient privacy concerns regarding making data fully public through GEO, please submit to NCBI's dbGaP (http://www.ncbi.nlm.nih.gov/gap/) database. dbGaP has controlled acces					
SERIES					
# This section describes the overall experiment.					
title	Genome-wide maps of chromatin state in pluripotent and lineage-committed cells.				
summary	We report the application of single-molecule-based sequencing technology for high-throughput profiling of histone modifications in mammalian cells. By obtaining over four				
overall design	Examination of 2 different histone modifications in 2 cell types.				
contributor	John,B,Goode				
contributor	Bradley,Smith				
supplementary file					
SRA_center_name_code	[optional]				
SAMPLES					
# This section lists and describes each of the biological Samples under investgation, as well as any protocols that are specific to individual Samples.					
# Additional "processed data file" or "raw file" columns may be included.					
Sample name	title	source name	organism	characteristics: cell type	characteristics: passages
Sample 1	H3K4me2_ChIPSeq	Neural progenitor cells	Mus musculus	ES-derived neural progenitor cells	15-18
Sample 2	H3K4me1_ChIPSeq	Neural progenitor cells	Mus musculus	ES-derived neural progenitor cells	15-18
Sample 3	input DNA	Neural progenitor cells	Mus musculus	ES-derived neural progenitor cells	15-18
PROTOCOLS					
# Any of the protocols below which are applicable to only a subset of Samples should be included as additional columns of the SAMPLES section instead.					
growth protocol	ES cell–derived NS cells were routinely generated by re-plating d 7 adherent neural differentiation cultures (typically 2–3 × 106 cells into a T75 flask) on uncoated plastic ir				
treatment protocol					
extract protocol	Lysates were clarified from sonicated nuclei and histone-DNA complexes were isolated with antibody.				
library construction protocol	Libraries were prepared according to Illumina's instructions accompanying the DNA Sample Kit (Part# 0801-0303). Briefly, DNA was end-repaired using a combination of				
library strategy	ChIP-Seq				

GEO Data submission VIII

Metadata spreadsheet example II



A	B	C	D	E	F
DATA PROCESSING PIPELINE					
# Data processing steps include base-calling, alignment, filtering, peak-calling, generation of normalized abundance measurements etc…					
# For each step provide a description, as well as software name, version, parameters, if applicable.					
# Include additional steps, as necessary.					
data processing step	Basecalls performed using CASAVA version 1.4				
data processing step	ChIP-seq reads were aligned to the mm9 genome assembly using EasyAlign version 3.2 with the following configurations…				
data processing step	Data were filtered using the following specifications…				
data processing step	peaks were called using PeaksFind version 2.2 with the following setting: ChIP threshold (0.2), Enrichment Fold (2.5), Rescue Fold (3).				
data processing step					
genome build	mm9				
processed data files format and content	wig files were generated using …; Scores represent …				
# For each file listed in the "processed data file" columns of the SAMPLES section, provide additional information below.					
PROCESSED DATA FILES					
file name	**file type**	**file checksum**			
H3K4me2.peaks.wig	wig	95cf1d1fa509d871b2ef0bb9fd734c3d			
H3K4me1.peaks.wig	wig	8ec6ee3cce10b970e5bfea4e35cdb231			
H3K4me2.b.peaks.wig	wig	f8fcd650914ff1a733956d6d06e8b543			
# For each file listed in the "raw file" columns of the SAMPLES section, provide additional information below.					
RAW FILES					
file name	**file type**	**file checksum**	**instrument model**	**read length**	**single or paired-end**
080716_BI-EAS46_0001_209DH_L1.fastq	fastq	6cc6ee3cce10b970e5bfea4e35cdb	Illumina Genome Analyzer	36	single
080716_BI-EAS46_0001_209DH_L2.fastq	fastq	88ceb0e0d056dda9208a03acf9073	Illumina Genome Analyzer	36	single
080716_BI-EAS46_0001_209DH_L3.fastq	fastq	f2786fedc5106789a2af4014a0e74f	Illumina Genome Analyzer	36	single
080716_BI-EAS46_0001_209DH_L4.fastq	fastq	d8fcd650914ff1a733956d6d06e8b0	Illumina Genome Analyzer	36	single
080716_BI-EAS46_0001_209DH_L5.fastq	fastq	03839cca2e797b28b9f9371f7b9ca	Illumina Genome Analyzer	36	single
080716_BI-EAS46_0001_209DH_L6.fastq	fastq	604fbb658413c559511eb6ad2bb14	Illumina Genome Analyzer	36	single
080717_BI-EAS46_0001_20DH_L5.fastq	fastq	57cf1d1fa509d871b2ef0bb9fd734c	Illumina Genome Analyzer IIx	42	single
080717_BI-EAS46_0001_20DH_L6.fastq	fastq	e5718e1a97690d410464f24f37aae	Illumina Genome Analyzer IIx	42	single

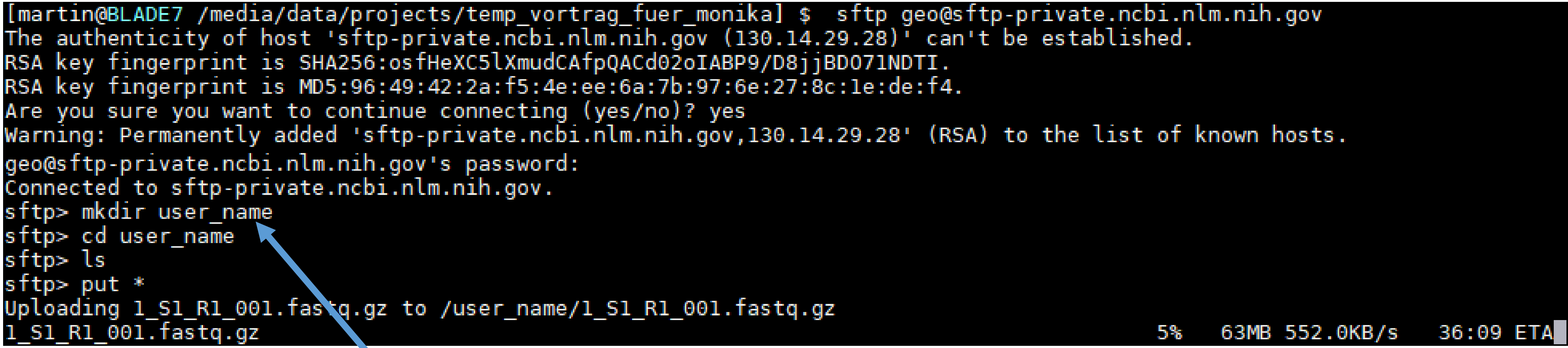
GEO Data submission IX

Transfer all your files to the GEO FTP server

- You must be logged in to your GEO account.
- Your transfer should include all required components (raw data files, processed data files and metadata spreadsheet).
- Start at the Submitting data page for full submission requirements.
- On your computer, create a folder named using your GEO username (/johndoe). Put all required submission files into this folder.
- Transfer the folder to the GEO FTP server using the credentials
- After the FTP transfer is complete, you must notify GEO using the Submit to GEO web form.

GEO Data submission X

connect to ftp server



```
[martin@BLADE7 /media/data/projects/temp_vortrag_fuer_monika] $  sftp geo@sftp-private.ncbi.nlm.nih.gov
The authenticity of host 'sftp-private.ncbi.nlm.nih.gov (130.14.29.28)' can't be established.
RSA key fingerprint is SHA256:osfHeXC5lXmudCAfpQACd02oIABP9/D8jjBDO71NDTI.
RSA key fingerprint is MD5:96:49:42:2a:f5:4e:ee:6a:7b:97:6e:27:8c:1e:de:f4.
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added 'sftp-private.ncbi.nlm.nih.gov,130.14.29.28' (RSA) to the list of known hosts.
geo@sftp-private.ncbi.nlm.nih.gov's password:
Connected to sftp-private.ncbi.nlm.nih.gov.
sftp> mkdir user_name
sftp> cd user_name
sftp> ls
sftp> put *
Uploading 1_S1_R1_001.fastq.gz to /user_name/1_S1_R1_001.fastq.gz
1_S1_R1_001.fastq.gz                                                                      5%   63MB 552.0KB/s   36:09 ETA
```

Create directory user_name: „mkdir user_name“
Change directory to user_name: „cd user_name“
List all files of the directory: „ls“
Upload all files of the directory:“put *“

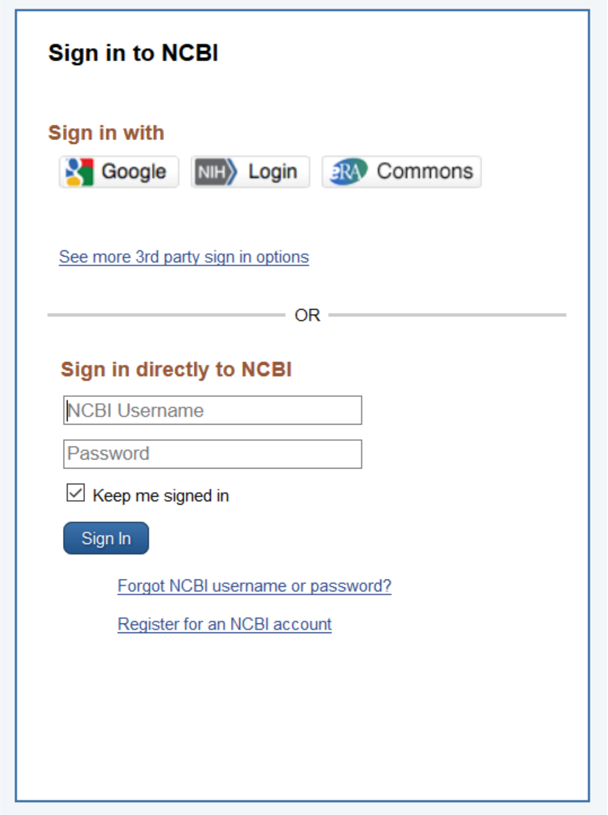
Upload speed

GEO Data submission XI

There are two steps for submission:

1. Transfer all your files to the GEO FTP server

2. After the FTP transfer is complete, notify GEO using the Submit to GEO web form



Sign in to NCBI

Sign in with

See more 3rd party sign in options

OR

Sign in directly to NCBI

NCBI Username

Password

☑ Keep me signed in

Forgot NCBI username or password?

Register for an NCBI account