

Choosing Discipline-Specific Research Data Repositories

In what cases cannot research data be deposited in RSU Dataverse?

The RSU Research Data Management Procedure foresees the deposit of datasets created by RSU researchers in the institutional repository RSU Dataverse, except in cases where:

- the terms of the project funding institution determine that data be deposited in another discipline-specific research data repository;
- it is generally accepted in the respective discipline to store data sets in another repository;
- RSU Dataverse is unable to ensure effective cooperation with any specific data processing or analytics tools;
- the size of the dataset is too large to be stored in RSU Dataverse (the size of one dataset may not exceed 50 GB);
- the project or other cooperation agreement stipulates otherwise.

What requirements must a repository meet?

In cases where data is not stored in RSU Dataverse, the requirements for research data repositories or data archives are as follows:

- ensures secure storage by performing regular data backups every 24 hours;
- offers open standards for data access and metadata standards;
- assigns a permanent unique identifier (e.g., DOI);
- is registered or certified (e.g., Re3Data, EOSC, CoreTrustSeal);
- grants licences (e.g., Creative Commons);
- clearly defines the access and duration of data and metadata storage;
- if applicable, provides options for storing sensitive information in restricted or closed access;
- provides automated versioning and description of changes in the event of revisions;
- data storage and processing take place in the EU or the European Economic Area (EEA). If processing takes place outside the EU/EEA, a data processing agreement must be concluded, which is coordinated with the Data Security and Management Unit and the Department of Information Technology.

Practical instructions for searching repositories

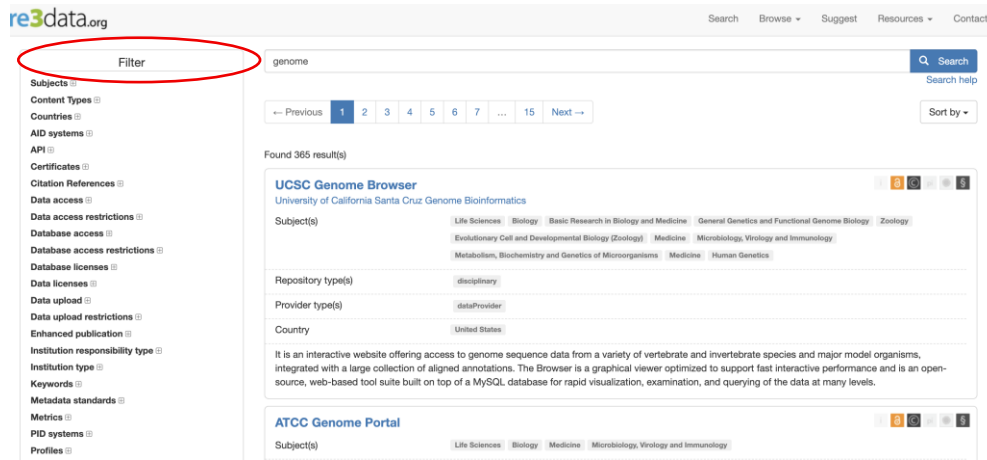
As mentioned above, one of the signs of a trusted repository is its registration on Open Science platforms such as Re3Data and EOSC. The presence of the repository on these platforms means that most of the basic requirements are met.

When looking for a discipline-specific repository, a good strategy is to use the Re3Data registry.

1. Visit the website <https://www.re3data.org/>.



2. Entering keywords for selecting repositories (e.g., genome), into the search engine will return 365 entries, which is a disproportionately large number for manual review. Therefore, it is recommended to use the filtering options available on the left side of the page.



3. By selecting the first criterion (Subject), you can significantly reduce the number of results. Below is an example of selecting the medical subject.

- Subjects** ▾
- Humanities and Social Sciences (8)
 - Humanities (1)
 - Ancient Cultures (1)
 - History (1)
 - Art History, Music, Theatre and Media Studies (1)
 - Social and Cultural Anthropology, Non-European Cultures, Jewish Studies and Religious Studies (1)
 - Theology (1)
 - Social and Behavioural Sciences (3)
 - Psychology (2)
 - Personality Psychology, Clinical and Medical Psychology, Methodology (1)
 - Social Sciences (1)
 - Life Sciences (211)
 - Biology (204)
 - Basic Research in Biology and Medicine (183)
 - Biochemistry (21)
 - Biophysics (3)
 - Cell Biology (27)
 - Structural Biology (10)
 - General Genetics and Functional Genome Biology (137)
 - Developmental Biology (2)
 - Bioinformatics and Theoretical Biology (48)
 - Plant Sciences (26)
 - Ecology and Biodiversity of Plants and Ecosystems (1)
 - Organismic Interactions, Chemical Ecology and Microbiomes of Plant Systems (1)
 - Plant Biochemistry and Biophysics (3)
 - Plant Genetics and Genomics (19)
 - Zoology (48)
 - Evolution, Anthropology (1)
 - Ecology and Biodiversity of Animals and Ecosystems, Organismic Interactions (2)
 - Sensory and Behavioural Biology (1)
 - Animal Physiology and Biochemistry (2)
 - Evolutionary Cell and Developmental Biology (Zooology) (42)
 - Medicine (211)**
 - Microbiology, Virology and Immunology (106)
 - Metabolism, Biochemistry and Genetics of Microorganisms (105)

← Previous **1** 2 3 4 5 6 7 8 9 Next →

Found 211 result(s)

UCSC Genome Browser
University of California Santa Cruz Genome Bioinformatics

Subject(s) Life Sciences Biology Basic Research

Evolutionary Cell and Developmental Biol
Metabolism, Biochemistry and Genetics of

Repository type(s) disciplinary

Provider type(s) dataProvider

Country United States

It is an interactive website offering access to genome sequence data from a integrated with a large collection of aligned annotations. The Browser is a g source, web-based tool suite built on top of a MySQL database for rapid vis

ATCC Genome Portal

Subject(s) Life Sciences Biology Medicine M

Repository type(s) institutional

Provider type(s) dataProvider

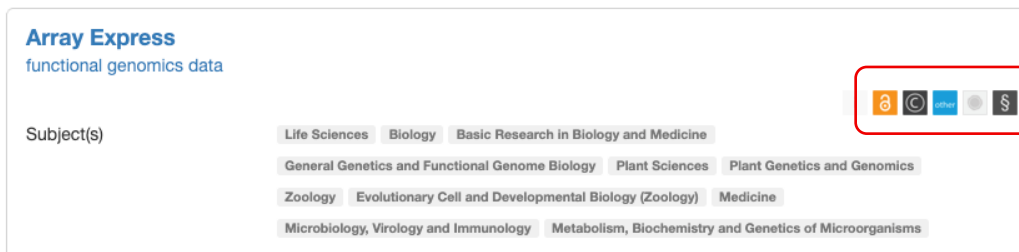
Country United States

Database and knowledgebase of authenticated microbial genomics data wi Collection's (ATCC) biorepository and culture collections. Data includes whc genome assemblies, metadata, drug susceptibility data, and more. All data web-based interface suite a REST API. The goal is to provide the research







- By looking at the selection options more carefully and choosing specific medical topic subsections, the number of results narrows down.

Medicine (211)
 Microbiology, Virology and Immunology (106)
 Metabolism, Biochemistry and Genetics of Microorganisms (29)
 Microbial Ecology and Applied Microbiology (3)
 Medical Microbiology and Mycology, Hygiene, Molecular Infection Biology (7)
 Virology (9)
 Immunology (8)
 Medicine (126)
 Epidemiology and Medical Biometry/Statistics (11)
 Public Health, Healthcare Research, Social and Occupational Medicine (8)
 Human Genetics (89)
 Anatomy and Physiology (6)
 Pathology (3)
 Medical Informatics and Medical Bioinformatics (1)
 Pharmacy (2)
 Pharmacology (3)
 Toxicology, Laboratory Medicine (3)
 Hematology, Oncology (8)
 Gastroenterology (1)
 Medical Physics, Biomedical Technology (3)
 Nuclear Medicine, Radiotherapy, Radiobiology (1)
 Neurosciences (16)
 Molecular Biology and Physiology of Neurons and Glial Cells (3)
 Clinical Neurology; Neurosurgery and Neuroradiology (1)
 Human Cognitive and Systems Neuroscience (2)
 Clinical Psychiatry, Psychotherapy, Child and Adolescent Psychiatry (1)
 Ophthalmology (1)

- The remaining selection options can be used as needed – depending on the publication or funder's requirements.
- Next, you need to evaluate the selected repository, for which we will look at Array Express as an example.






It is necessary to pay attention to the indicated emblems in the upper right corner of the repository description – each icon represents a specific functionality or element of the repository.

-  Offers a description of the services it provides.
-  Offers open access to data.
-  Provides licensing.
-  Assigns a permanent unique identifier (e.g., DOI).
-  The repository is certified or uses a generally accepted repository standard.
-  The repository has its own policy document created and available.

If the particular emblem is coloured, the repository provides the given element. So we can conclude that Array Express provides open access, licensing, permanent unique identifier assignment, and a policy document is available.

Unfortunately, in very rare cases, repositories provide all 6 elements, but this does not necessarily mean that this is a reason not to use the repository. In order for a repository to meet the basic requirements set by RSU, it must necessarily provide the following elements:

-  Offers open access to data.
-  Provides licensing.
-  Assigns a permanent unique identifier (e.g., DOI).

The repository you plan to use should have at least these 3 badges highlighted.

It should also be noted that sometimes the information in the register may not be updated, and some of the elements provided in the repository may not yet be appropriately indicated in the register.

In the case of sensitive data, it is also important to ensure that the data is stored and processed within the EU or the European Economic Area (EEA). If, however, the chosen repository stores the data outside the EU or EEA, it is necessary to contact the RSU data protection officer or data curators.

Specific cases – large amounts of genetic data

In cases where, due to technical limitations (large amount of data or specific file format), it is not possible to deposit a dataset in RSU Dataverse, it is recommended to use Re3Data to search for another appropriate repository.

In such situations, it is advisable to be guided by the following parameters:




- the repository is included in Re3Data;
- it provides open access to data;
- it complies with GDPR requirements (if the dataset contains sensitive data).

GDPR-compliant repositories

Below are listed the most popular open non-commercial repositories where this type of data can be deposited.

It should be noted that currently the registry does not mark all 3 basic functionality conditions for the repositories shown below, although in fact these repositories provide it for the majority of deposited datasets. For example, ENA and the assignment of independent identifiers to a dataset, DOIs are not assigned in this repository, but rather accession numbers. These numbers are not as functional as DOIs in finding datasets, therefore this functionality is not indicated as existing in the registry. Taking into account the fact that there are not many repositories in which the functional capabilities allow uploading large amounts of data and at the same time ensure GDPR requirements, not assigning DOIs does not limit the use of this repository.

European Nucleotide Archive
ENA

Subject(s)   

Life Sciences Biology Basic Research in Biology and Medicine General Genetics and Functional Genome Biology

Bioinformatics and Theoretical Biology Medicine Microbiology, Virology and Immunology

Repository type(s) disciplinatory

Provider type(s) dataProvider serviceProvider

Country European Union International United Kingdom

The European Nucleotide Archive (ENA) captures and presents information relating to experimental workflows that are based around nucleotide sequencing. A typical workflow includes the isolation and preparation of material for sequencing, a run of a sequencing machine in which sequencing data are produced and a subsequent bioinformatic analysis pipeline. ENA records this information in a data model that covers input information (sample, experimental setup, machine configuration), output machine data (sequence traces, reads and quality scores) and interpreted information (assembly, mapping, functional annotation). Data arrive at ENA from a variety of sources. These include submissions of raw data, assembled sequences and annotation from small-scale sequencing efforts, data provision from the major European sequencing centres and routine and comprehensive exchange with our partners in the International Nucleotide Sequence Database Collaboration (INSDC). Provision of nucleotide sequence data to ENA or its INSDC partners has become a central and mandatory step in the dissemination of research findings to the scientific community. ENA works with publishers of scientific literature and funding bodies to ensure compliance with these principles and to provide optimal submission systems and data access tools that work seamlessly with the published literature.

The European Genome-phenome Archive
EGA

Subject(s) Life Sciences Biology Basic Research in Biology and Medicine Structural Biology Medicine Medicine Human Genetics

Repository type(s) disciplinary

Provider type(s) dataProvider serviceProvider

Country European Union Spain

The European Genome-phenome Archive (EGA) is designed to be a repository for all types of sequence and genotype experiments, including case-control, population, and family studies. We will include SNP and CNV genotypes from array based methods and genotyping done with re-sequencing methods. The EGA will serve as a permanent archive that will archive several levels of data including the raw data (which could, for example, be re-analysed in the future by other algorithms) as well as the genotype calls provided by the submitters. We are developing data mining and access tools for the database. For controlled access data, the EGA will provide the necessary security required to control access, and maintain patient confidentiality, while providing access to those researchers and clinicians authorised to view the data. In all cases, data access decisions will be made by the appropriate data access-granting organisation (DAO) and not by the EGA. The DAO will normally be the same organisation that approved and monitored the initial study protocol or a designate of this approving organisation. The European Genome-phenome Archive (EGA) allows you to explore datasets from genomic studies, provided by a range of data providers. Access to datasets must be approved by the specified Data Access Committee (DAC).

Array Express
functional genomics data

Subject(s) Life Sciences Biology Basic Research in Biology and Medicine General Genetics and Functional Genome Biology Plant Sciences
Plant Genetics and Genomics Zoology Evolutionary Cell and Developmental Biology (Zoology) Medicine
Microbiology, Virology and Immunology Metabolism, Biochemistry and Genetics of Microorganisms

Repository type(s) disciplinary

Provider type(s) dataProvider serviceProvider

Country United Kingdom United States European Union

ArrayExpress is one of the major international repositories for high-throughput functional genomics data from both microarray and high-throughput sequencing studies, many of which are supported by peer-reviewed publications. Data sets are submitted directly to ArrayExpress and curated by a team of specialist biological curators. In the past (until 2018) datasets from the NCBI Gene Expression Omnibus database were imported on a weekly basis. Data is collected to MIAME and MINSEQE standards.

If the data is not sensitive, options for depositing data sets in the following repositories may be considered:

GenBank

Subject(s) Life Sciences Biology Basic Research in Biology and Medicine General Genetics and Functional Genome Biology
Bioinformatics and Theoretical Biology Medicine Microbiology, Virology and Immunology

Repository type(s) disciplinary

Provider type(s) dataProvider serviceProvider

Country United States United Kingdom

GenBank® is a comprehensive database that contains publicly available nucleotide sequences for almost 260 000 formally described species. These sequences are obtained primarily through submissions from individual laboratories and batch submissions from large-scale sequencing projects, including whole-genome shotgun (WGS) and environmental sampling projects. Most submissions are made using the web-based BankIt or standalone Sequin programs, and GenBank staff assigns accession numbers upon data receipt. Daily data exchange with the European Nucleotide Archive (ENA) and the DNA Data Bank of Japan (DDBJ) ensures worldwide coverage. GenBank is accessible through the NCBI Entrez retrieval system, which integrates data from the major DNA and protein sequence databases along with taxonomy, genome, mapping, protein structure and domain information, and the biomedical journal literature via PubMed. BLAST provides sequence similarity searches of GenBank and other sequence databases. Complete bimonthly releases and daily updates of the GenBank database are available by FTP.

Gene Expression Omnibus
GEO

Subject(s) Life Sciences Biology Basic Research in Biology and Medicine General Genetics and Functional Genome Biology

Repository type(s) disciplinary


Provider type(s) serviceProvider

Country United States

Gene Expression Omnibus: 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.

Sequence Read Archive

SRA



Subject(s)
 Life Sciences
Biology
Basic Research in Biology and Medicine
General Genetics and Functional Genome Biology
Plant Sciences

Plant Genetics and Genomics
Zoology
Evolutionary Cell and Developmental Biology (Zoology)

Repository type(s)
 disciplinary


Provider type(s)
 dataProvider

Country
 United States

The Sequence Read Archive stores the raw sequencing data from such sequencing platforms as the Roche 454 GS System, the Illumina Genome Analyzer, the Applied Biosystems SOLID System, the Helicos Heliscope, and the Complete Genomics. It archives the sequencing data associated with RNA-Seq, ChIP-Seq, Genomic and Transcriptomic assemblies, and 16S ribosomal RNA data.

DNA Data Bank of Japan

DDBJ



Subject(s)
 Life Sciences
Biology
Medicine
Medicine

Repository type(s)
 disciplinary

Provider type(s)
 dataProvider
serviceProvider

Country
 Japan
International

DDBJ; DNA Data Bank of Japan is the sole nucleotide sequence data bank in Asia, which is officially certified to collect nucleotide sequences from researchers and to issue the internationally recognized accession number to data submitters. Since we exchange the collected data with EMBL-Bank/EBI; European Bioinformatics Institute and GenBank/NCBI; National Center for Biotechnology Information on a daily basis, the three data banks share virtually the same data at any given time. The virtually unified database is called "INSD; International Nucleotide Sequence Database DDBJ collects sequence data mainly from Japanese researchers, but of course accepts data and issue the accession number to researchers in any other countries.

This list is not complete, as, depending on the additional specifications of the particular repository (e.g., certain tumour types or specific diseases), there are also other repositories in Re3Data where it is possible to deposit large amounts of genetic data.

! In the case of large amounts of genetic data, it is recommended to select a deposit location in a timely manner (if possible - before data generation), because:

- discipline-specific repositories have specific data structure, format, and metadata requirements, so it would be easier to follow them from the first day of data acquisition;
- it is often necessary to include detailed information about the hardware used, its settings, and the software used in the metadata, even specifying its version, so it is more convenient to record this information during data acquisition rather than trying to obtain the information retrospectively;
- open access repositories are non-commercial, with limited human resources, where quality control of data sets can take several weeks, so the duration of the deposit process needs to be taken into account when planning the publication deadlines for project results.