

# Thematic Data Repositories

## Contents

Natural Sciences.....	1
Engineering and Technology.....	3
Medical and Health Sciences .....	3
Agricultural and Veterinary Sciences .....	4
Social Sciences & Humanities .....	4

## Natural Sciences

### [PubChem](#) (Chemical Sciences)

Pubchem contains 3 databases:

1. PubChem BioAssay: The PubChem BioAssay Database contains bioactivity screens of chemical substances described in PubChem Substance. It provides searchable descriptions of each bioassay, including descriptions of the conditions and readouts specific to that screening procedure.
2. PubChem Compound: The PubChem Compound Database contains validated chemical depiction information provided to describe substances in PubChem Substance. Structures stored within PubChem Compounds are pre-clustered and cross-referenced by identity and similarity groups.
3. PubChem Substance. The PubChem Substance Database contains descriptions of samples from a variety of sources and links to biological screening results that are available in PubChem BioAssay. If the chemical contents of a sample are known, the description includes links to PubChem Compound.

### [STRENDA](#) (Biochemistry)

STRENDA (**S**tandards for **R**eporting **E**nzymology **D**ata) DB is a storage and search platform supported by the Beilstein-Institut that incorporates the [STRENDA Guidelines](#) in a user-friendly, web-based system. Data entered in the STRENDA DB submission form are automatically checked for compliance with the STRENDA Guidelines; users receive notifications informing them when necessary information is missing.

### [National Geoscience Data Centre](#) (Earth and related Environmental sciences)

The National Geoscience Data Centre (NGDC) collects and preserves geoscientific data and information, making them available for the long-term to a wide range of users and communities. The NGDC is recognised as the [NERC Environmental Data Centre](#) for geoscience data. The NGDC is responsible for managing geo-scientifically valuable information and datasets. The data it holds covers

many geological disciplines dealing with the physical structure of the Earth and the processes that act on it, as measured from geological time to near-real-time sensor and data streams.

### [NASA Open Data Portal](#) (Earth and related Environmental Sciences)

The NASA Open Data portal centralises NASA's open geospatial data generated from its rich history of planetary, lunar and terrestrial missions. It has nearly 20,000 unique monthly users and more than 40,000 datasets. A small percentage of these datasets are hosted directly on [data.nasa.gov](#), but in most cases metadata and links to other space agency projects are provided. [Data.nasa.gov](#) includes a [wide range of subject matter](#), from data related to rocket testing to geologic maps of Mars. The data is offered in multiple formats, depending on publisher. The site is part of the Open Innovation Sites project, along with [api.nasa.gov](#), a space for sharing information about NASA APIs, and [code.nasa.gov](#), where NASA's open source projects are collected.

### [Copernicus](#) (Earth and related Environmental sciences)

COPERNICUS is the [Earth observation program of the European Union](#). Led by the European Commission, with the collaboration of the member states and various European agencies and organisations, it collects, stores, combines and analyses data obtained through satellite observation and in situ sensor systems on land, air and sea. It provides data through six services: emergency, security, marine monitoring, land monitoring, climate change and atmospheric monitoring.

### [HEPData](#) (High Energy Physics)

The Durham High-Energy Physics Database (HEPData) is an open access repository for scattering data from experimental particle physics. It currently comprises the data points from plots and tables related to several thousand publications including those from the Large Hadron Collider (LHC).

### [MathRepo](#) (Mathematics)

Repository of the [Max Planck Institute for Mathematics in the Sciences](#) in Leipzig, dedicated to mathematical research data. The purpose of this repository is to collect scripts and code, to explain applications of mathematical software, to showcase additional examples to paper publications, and more generally to host supplementary material developed for research projects or discussed in workshops.

### [NCBI Databases](#) (Biology)

Various databases with biomedical and genomic information, including: [DNA & RNA, Domains & Structures, Genes & Expression, Genetics & Medicine, Genomes & Maps, Homology, Proteins, Sequence Analysis, etc.](#)

Other natural sciences repositories can be found in [re3data](#).

## Engineering and Technology

### [4TU.ResearchData](#)

4TU.ResearchData is an international data repository for science, engineering and design. Its services include curation, sharing, long-term access and preservation of research datasets. These services are available to anyone around the world. In addition, 4TU.ResearchData also offers training and resources to researchers to support them in making research data findable, accessible, interoperable and reproducible (FAIR). 4TU.ResearchData guarantees that datasets remain findable, usable and accessible in the future. 4TU.Research Data offers space for thousands of data sets and contains complete, valuable data sets from technical-scientific research. This may be raw data, code, processed data or specific data that is underlying a publication.

### [DEEDS](#)

The Digital Environment for Enabling Data-Driven Science (DEEDS) is a full-service platform that provides end-to-end support for your research investigations to scientific and engineering communities worldwide. DEEDS is a systematic, secure, reliable way to conduct your research, with powerful yet user-friendly interactive support for shared data, computing, and scientific workflows. The design and development of DEEDS is a collaborative effort, bringing together scientists and engineers from chemistry, agriculture, electrical engineering, civil engineering, health and human services and computer science. The platform supports research activities throughout the research investigation and is effective across science domains.

## Medical and Health Sciences

### [Vivli](#)

Vivli is a global clinical research data sharing platform. It facilitates data sharing and data access from clinical trials to advanced science. It provides a secure environment and provides robust analytical tools.

### [OpenTrials](#)

Open Trials is a collaboration between Open Knowledge International and Dr Ben Goldacre from the University of Oxford DataLab. It aims to locate, match and share all publicly accessible data and documents, on all trials conducted, on all medicine and other treatments, globally.

### [Project Data Sphere](#)

The Project Data Sphere aims to break down the barriers to cancer clinical trial data sharing - barriers that historically have kept valuable trial data from ultimately benefiting the patients who so selflessly participate in them. By aggregating trial data from biopharmaceutical companies, academic medical centres, and government organisations and making it freely available on this open access platform, makes it a premier resource for the global oncology research community. The relationships with renowned oncology experts allow to *convene* research collaborations that leverage the power of

pooled clinical trial data and which ultimately position the project data sphere to be a catalyst for the discovery of urgently needed new treatments while helping to make cancer trials faster, more effective, and less expensive.

### [ClinicalCodes.org](https://clinicalcodes.org)

The ClinicalCodes repository aims to hold code lists for all published electronic medical record studies, irrespective of code type (e.g. Read, ICD9-10, SNOMED) and database (CPRD, QResearch, THIN, etc.). Once deposited, code lists will be freely available, with no login needed to download codes.

## Agricultural and Veterinary Sciences

### [AG Data Commons](https://datacommons.usda.gov/)

Data repository of the US Department of Agriculture's National Agricultural Library to provide open access to data, multimedia, and other documents funded in whole or in part by the department. Projects span subjects such as agronomy, genomics, hydrology, soils, agro-ecosystems, sustainability science, and economic statistics.

### [Movebank](https://movebank.org/)

The Movebank Data Repository is hosted by the Communication, Information, Media Centre (KIM) of the University of Konstanz that publishes data packages based on data stored on the Movebank platform - a global, online research platform for managing, sharing, analysing and archiving animal tracking and animal-borne sensor data hosted by the Max Planck Institute of Animal Behavior. Data owners manage and retain ownership of their data within user-controlled studies in Movebank. This repository allows data owners to archive their data as data packages that include a citation, persistent identifier (DOI), and Creative Commons Zero licence.

## Social Sciences & Humanities

### [DARIAH-EU](https://dariah.eu/)

DARIAH-EU is a digital research infrastructure for the Arts and Humanities which aims to enhance and support digitally-enabled research and teaching across the arts and humanities. Working with communities of practice, DARIAH brings together individual state-of-the-art digital activities across Europe. This helps to preserve and provide access to and disseminate research that stems from collaborations and ensure that best practices, methodological and technical standards are adhered to.

### [CROSSDA](https://crossda.eu/) (Croatian data archive for social sciences)

Croatian national public service for research data and service provider for the Consortium of European Social Science Data Archives CESSDA ERIC.