The STORE database; towards FAIR

Paul Schofield, University of Cambridge

PIANOFORTE
EUROPEAN PARTNERSHIP FOR RADIATION PROTECTION RESEARCH

RadoNorm
Managing risks from radon and NORM

BfS Bundesamt für Strahlenschutz

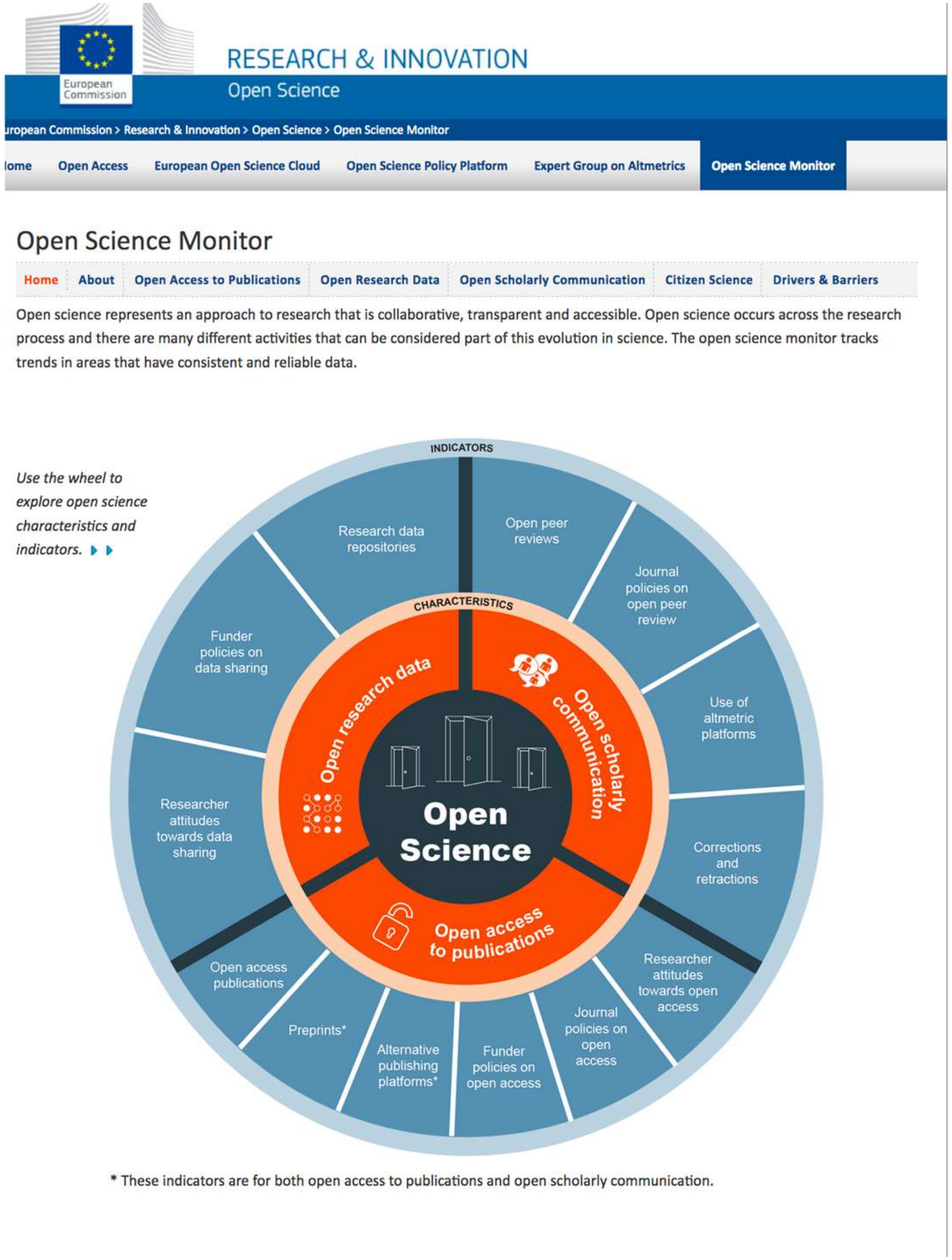
UNIVERSITY OF CAMBRIDGE

UNIVERSITY OF BIRMINGHAM

NASA GeneLab
OPEN SCIENCE FOR LIFE IN SPACE

Open Science

- Open publication access
- Open scholarly communication
- Open research data



European Commission

RESEARCH & INNOVATION

Open Science

uropean Commission > Research & Innovation > Open Science > Open Science Monitor

lome | Open Access | European Open Science Cloud | Open Science Policy Platform | Expert Group on Altmetrics | Open Science Monitor

Open Science Monitor

Home | About | Open Access to Publications | Open Research Data | Open Scholarly Communication | Citizen Science | Drivers & Barriers

Open science represents an approach to research that is collaborative, transparent and accessible. Open science occurs across the research process and there are many different activities that can be considered part of this evolution in science. The open science monitor tracks trends in areas that have consistent and reliable data.

Use the wheel to explore open science characteristics and indicators. ▸ ▸

INDICATORS
CHARACTERISTICS
Open Science
Open research data
Open scholarly communication
Open access to publications
Research data repositories
Open peer reviews
Journal policies on open peer review
Use of altmetric platforms
Corrections and retractions
Researcher attitudes towards open access
Journal policies on open access
Funder policies on open access
Alternative publishing platforms*
Preprints*
Open access publications
Researcher attitudes towards data sharing
Funder policies on data sharing

* These indicators are for both open access to publications and open scholarly communication.

Open Science

“Open science is an approach based on open cooperative work and systematic sharing of knowledge and tools as early and widely as possible in the process”

European Commission

- Open Access to publications
- Responsible management of data (FAIR principles)
- Open access to data ‘as open as possible, as closed as necessary’
- Information about outputs / tools / instruments to validate/re-use results and data
- Digital /physical access to results to validate the conclusions

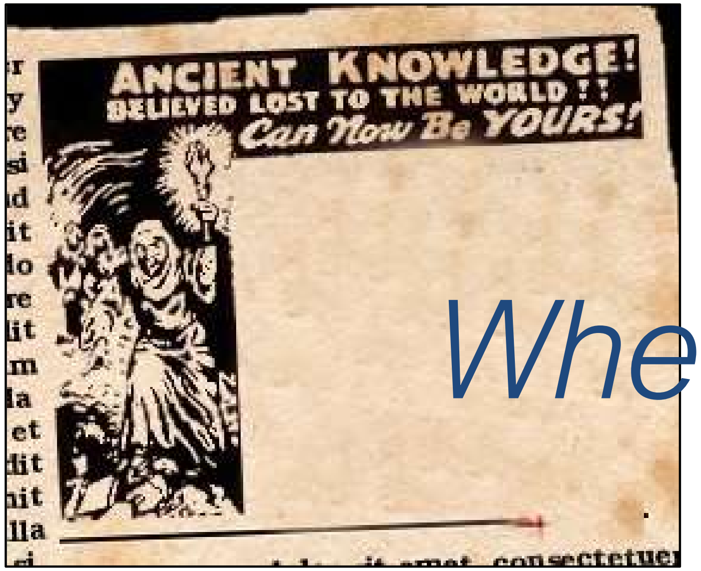
EC Requirements

- Must manage the digital research data in line with the **FAIR principles** (Findable, Accessible, Interoperable, Reusable)
- **Data Management Plan** (DMP) is required by M6; updated mid-project and at end of project
- **Deposit (meta)data as soon as possible** after production/generation or after processing and quality controls
- Deposit data in a **trusted repository** and make them **open as soon as possible** (deadlines set in DMP), following the “as open as possible, as closed as necessary” (open by default) principles
- Data closed if necessary, but **metadata must be FAIR and unde r CCO** (trusted repositories will automatically share metadata in CCO)
- Open licence, preferentially CC-BY or CC0 licence
- Detailed information about research outputs or tools/instruments needed to re-use or validate the data (e.g. data, software, algorithms, protocols, models, workflows, electronic notebooks)

Examples of metadata
author(s) name, author(s) ORCID, DOI, licence, language, journal, title, etc.

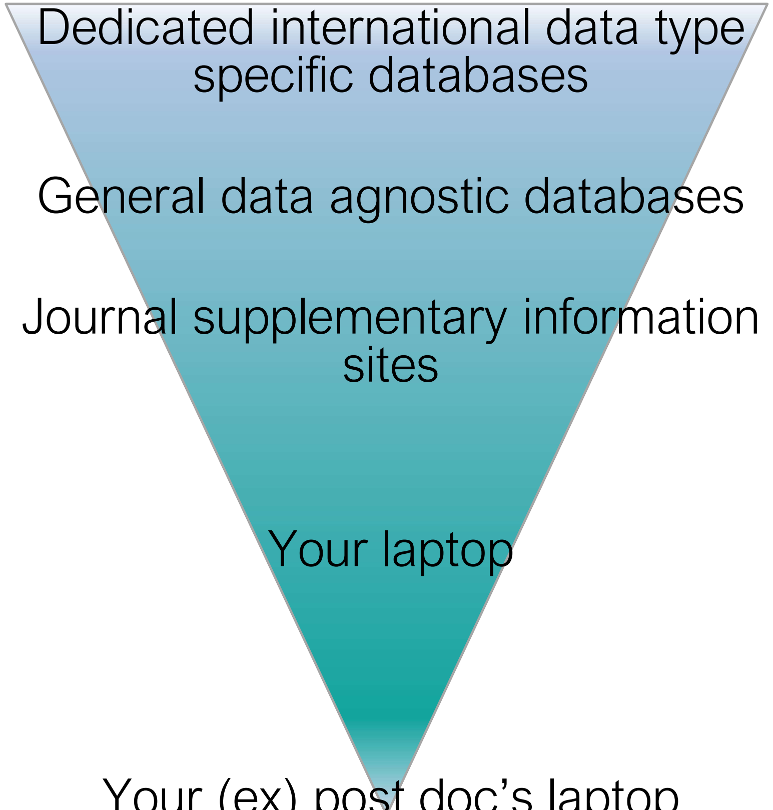
Why do we need to share data?

- Evidence support for publications
    - Reproducibility
    - Accountability
- Data and resource reuse and reanalysis
    - Resilience
    - New discovery
    - Prevention of duplication
    - Supports RRRs
    - Value for funding
- Coordination and integration of projects
- Training and education resources



Where's the data?

Data resources



Dedicated international data type specific databases

General data agnostic databases

Journal supplementary information sites

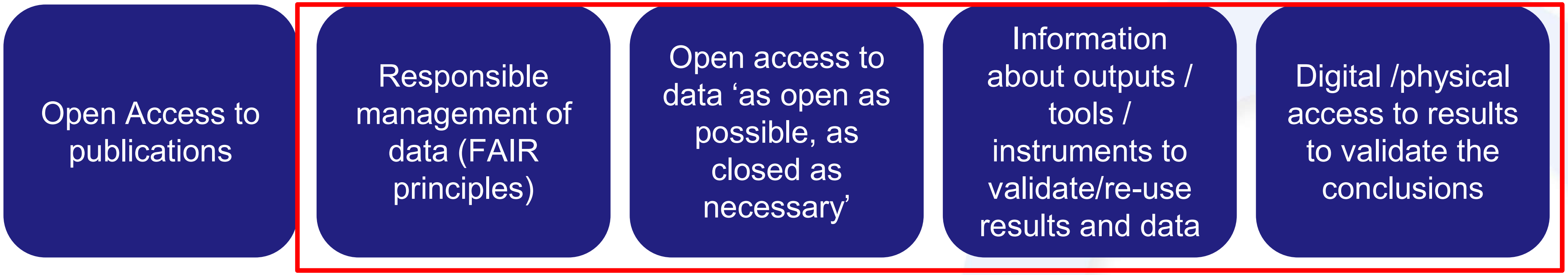
Your laptop

Your (ex) post doc's laptop

ENA
European Nucleotide Archive
MGI
PRIDE
NASA GeneLab
OPEN SCIENCE FOR LIFE IN SPACE
AE
figshare
ERA
european radiobiological archives
STORE DB

Open Science

> “Open science is an approach based on open cooperative work and systematic sharing of knowledge and tools as early and widely as possible in the process”
>
> European Commission



- Open Access to publications
- Responsible management of data (FAIR principles)
- Open access to data ‘as open as possible, as closed as necessary’
- Information about outputs / tools / instruments to validate/re-use results and data
- Digital /physical access to results to validate the conclusions

The STORE database

Funded by the European Commission EURATOM Programme and the Bundesamt fuer Strahlenschutz since 2009

European Commission | Horizon 2020 European Union funding for Research & Innovation

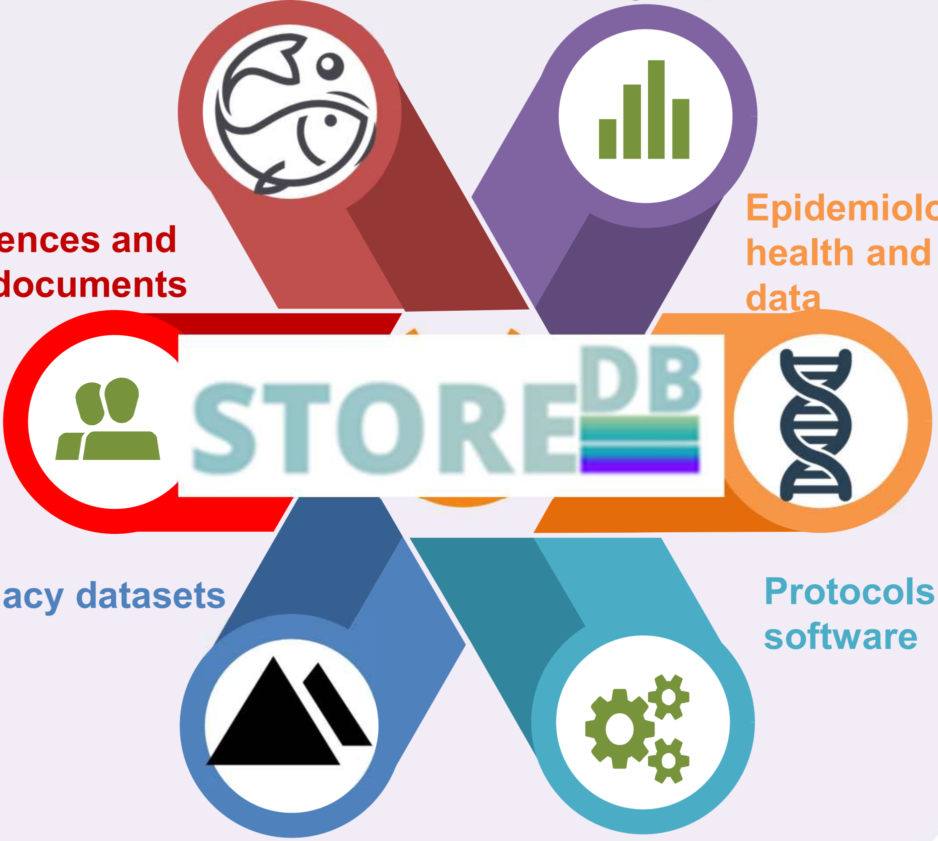
RadoNorm Managing risks from radon and NORM

elixir

re3data.org REGISTRY OF RESEARCH DATA REPOSITORIES

http://doi.org/10.17616/R3732R

STOREDB



Environmental data

Primary experimental data

Social sciences and policy documents

Epidemiological, health and genetic data

Legacy datasets

Protocols and software

http://www.storedb.org

Free to the entire community and sustained by the BfS Self curation, upload and controlled access.

FAIRsharing.org standards, databases, policies

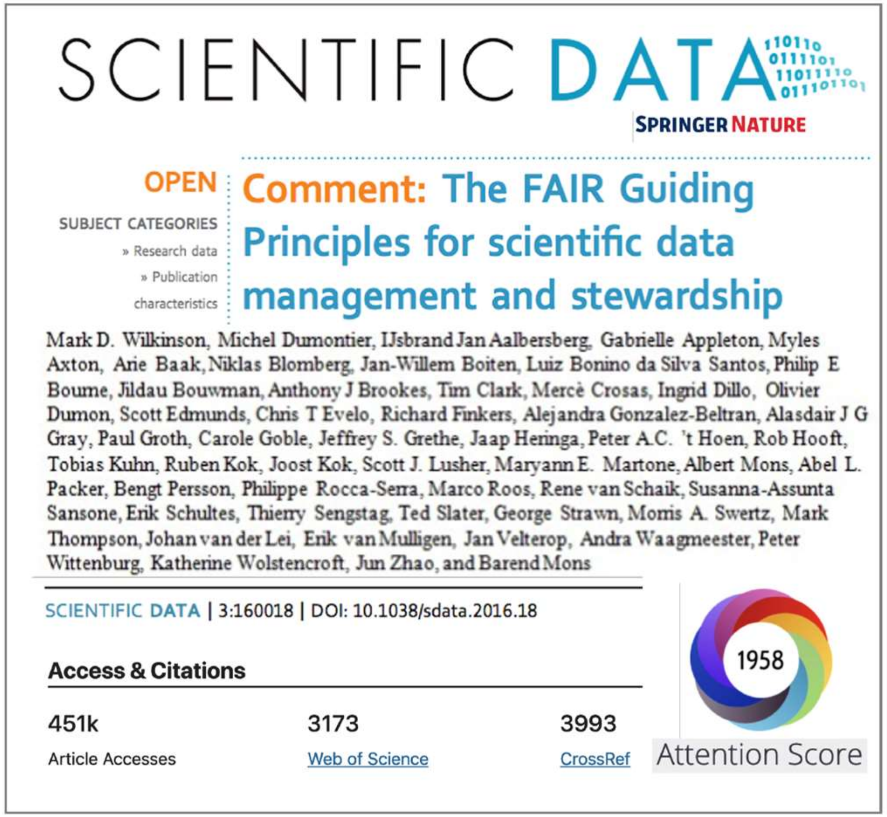
DataCite FIND, ACCESS, AND REUSE DATA

Findable Accessible Interoperable Reusable

Funded by Euratom research and training programme 2014-2018 under grant agreement No 900009.

FAIR sharing

Use of data at scale by humans and machines



SCIENTIFIC DATA

SPRINGER NATURE

OPEN

SUBJECT CATEGORIES
» Research data
» Publication characteristics

Comment: The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, Jildau Bouwman, Anthony J Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J G Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C. 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons

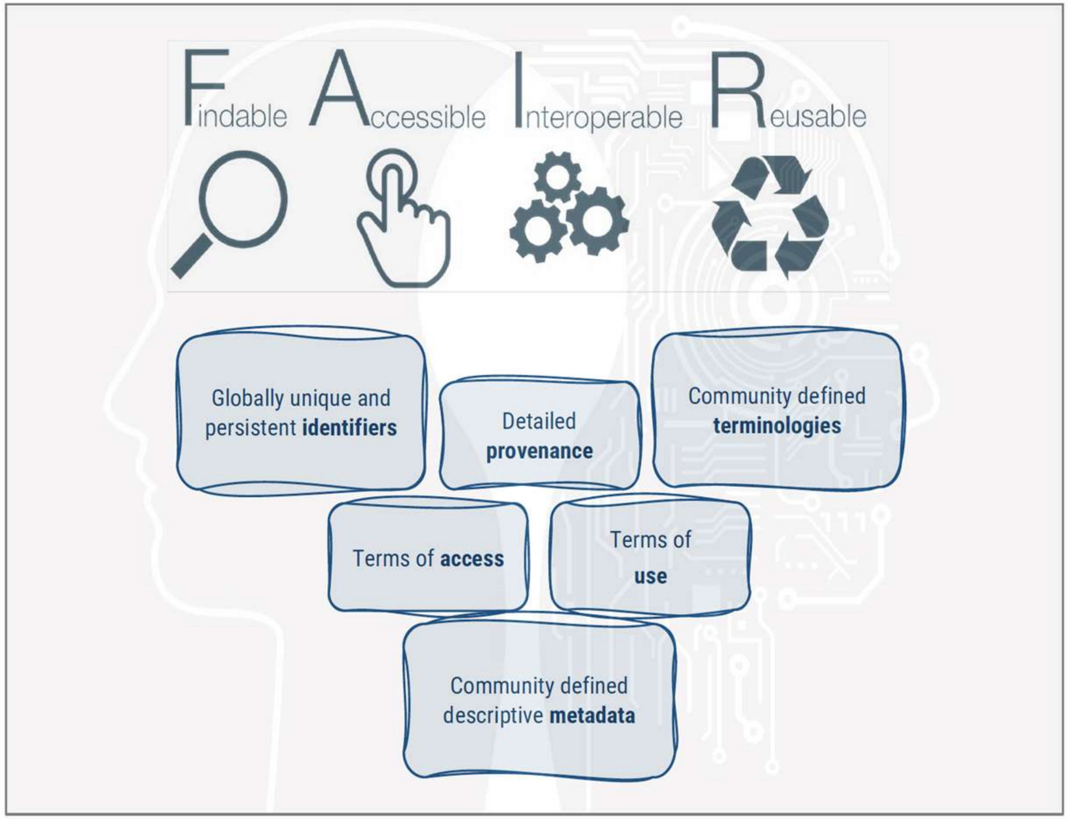
SCIENTIFIC DATA | 3:160018 | DOI: 10.1038/sdata.2016.18

Access & Citations

451k	3173	3993
Article Accesses	Web of Science	CrossRef

1958

Attention Score



Findable
Accessible
Interoperable
Reusable
Globally unique and persistent **identifiers**
Detailed **provenance**
Community defined **terminologies**
Terms of **access**
Terms of **use**
Community defined descriptive **metadata**

OPEN AND *FAIR* ARE NOT THE SAME THING

“Open” is about data rights and licensing.
“FAIR” is about mechanics.

Anne Raugh

FAIR IS PRIMARILY CONCERNED WITH PROGRAMMATIC PROCESSING

FAIR recognizes that data is diverse and scattered across cyberspace. Software processing levels the playing field.

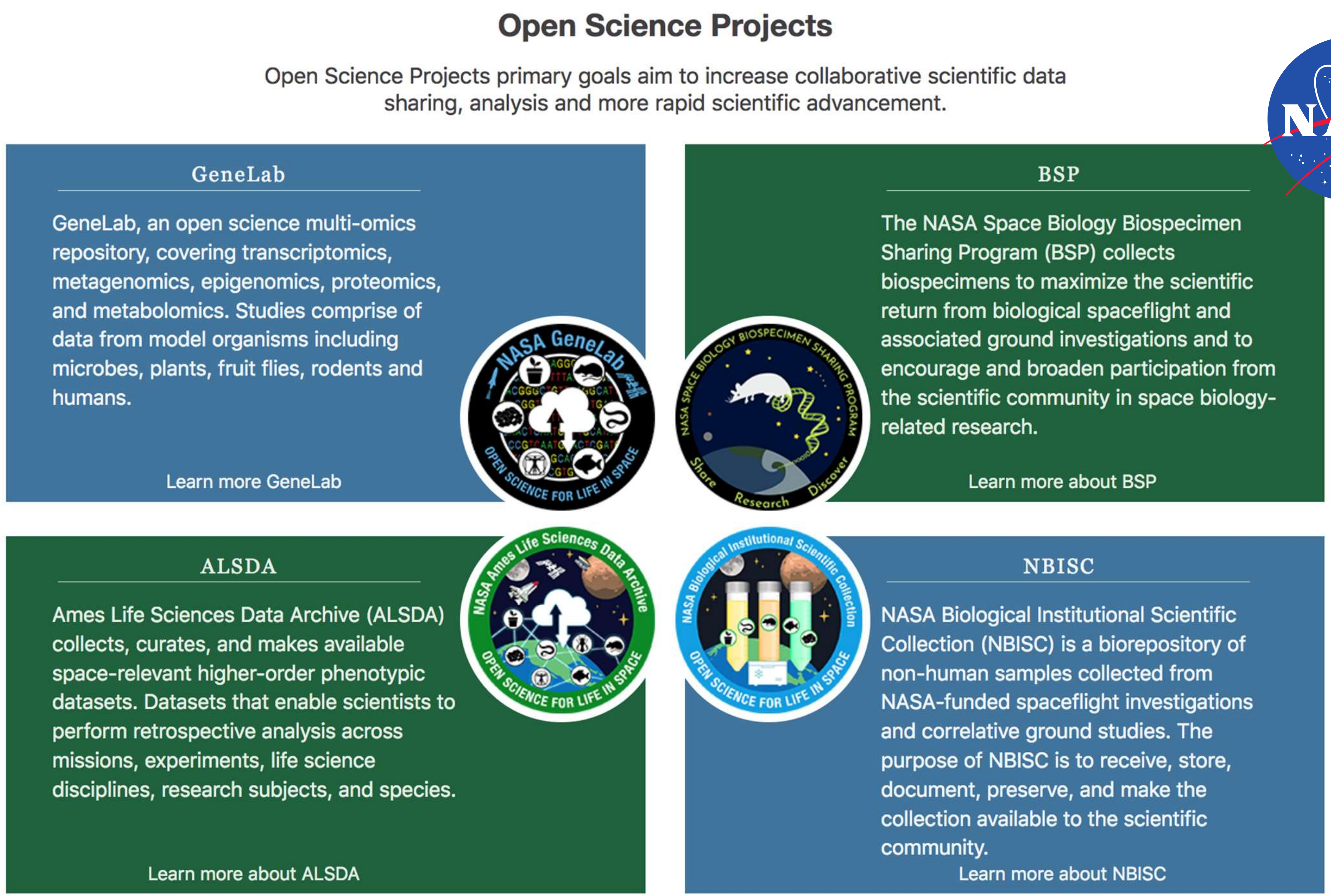
Anne Raugh

FAIR is not a binary state

FAIRness is a spectrum. There are various ways to increase FAIRness and they can be applied incrementally.

BUT

- Use of generic data repositories is not very FAIR
- Lack of granularity in metadata
- Lack of metadata standards relevant to theme or community
- Non-grounded



Open Science Projects

Open Science Projects primary goals aim to increase collaborative scientific data sharing, analysis and more rapid scientific advancement.

GeneLab

GeneLab, an open science multi-omics repository, covering transcriptomics, metagenomics, epigenomics, proteomics, and metabolomics. Studies comprise of data from model organisms including microbes, plants, fruit flies, rodents and humans.

Learn more GeneLab

BSP

The NASA Space Biology Biospecimen Sharing Program (BSP) collects biospecimens to maximize the scientific return from biological spaceflight and associated ground investigations and to encourage and broaden participation from the scientific community in space biology-related research.

Learn more about BSP

ALSDA

Ames Life Sciences Data Archive (ALSDA) collects, curates, and makes available space-relevant higher-order phenotypic datasets. Datasets that enable scientists to perform retrospective analysis across missions, experiments, life science disciplines, research subjects, and species.

Learn more about ALSDA

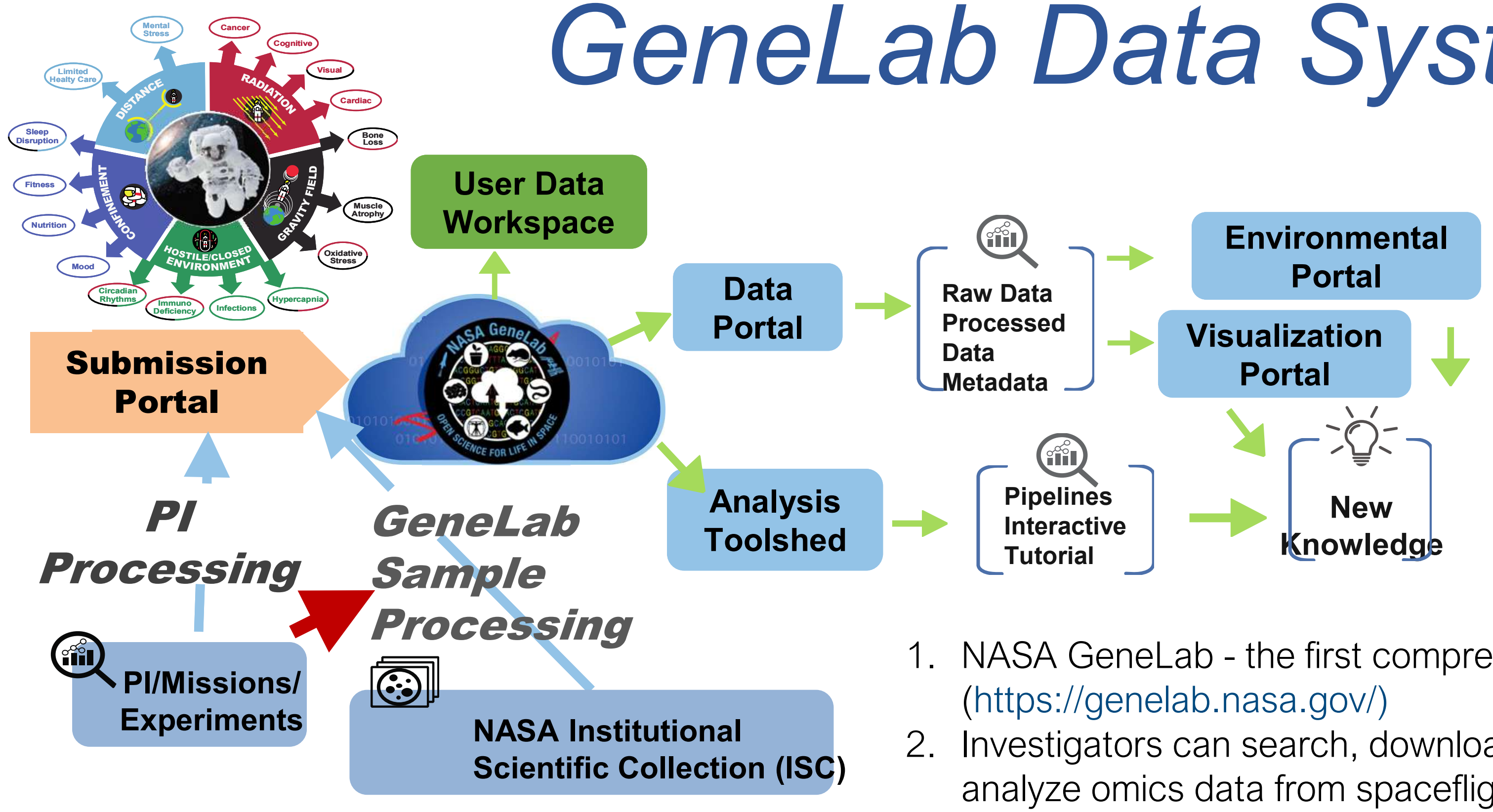
NBISC

NASA Biological Institutional Scientific Collection (NBISC) is a biorepository of non-human samples collected from NASA-funded spaceflight investigations and correlative ground studies. The purpose of NBISC is to receive, store, document, preserve, and make the collection available to the scientific community.

Learn more about NBISC

NASA GeneLab
OPEN SCIENCE FOR LIFE IN SPACE
NASA SPACE BIOLOGY BIOSPECIMEN SHARING PROGRAM
Share
Research
Discover
NASA Ames Life Sciences Data Archive
OPEN SCIENCE FOR LIFE IN SPACE
NASA Biological Institutional Scientific Collection
OPEN SCIENCE FOR LIFE IN SPACE

GeneLab Data System



User Data Workspace

Submission Portal

Data Portal

Raw Data
Processed Data
Metadata

Environmental Portal

Visualization Portal

Analysis Toolshed

Pipelines
Interactive
Tutorial

New Knowledge

PI Processing

GeneLab Sample Processing

PI/Missions/ Experiments

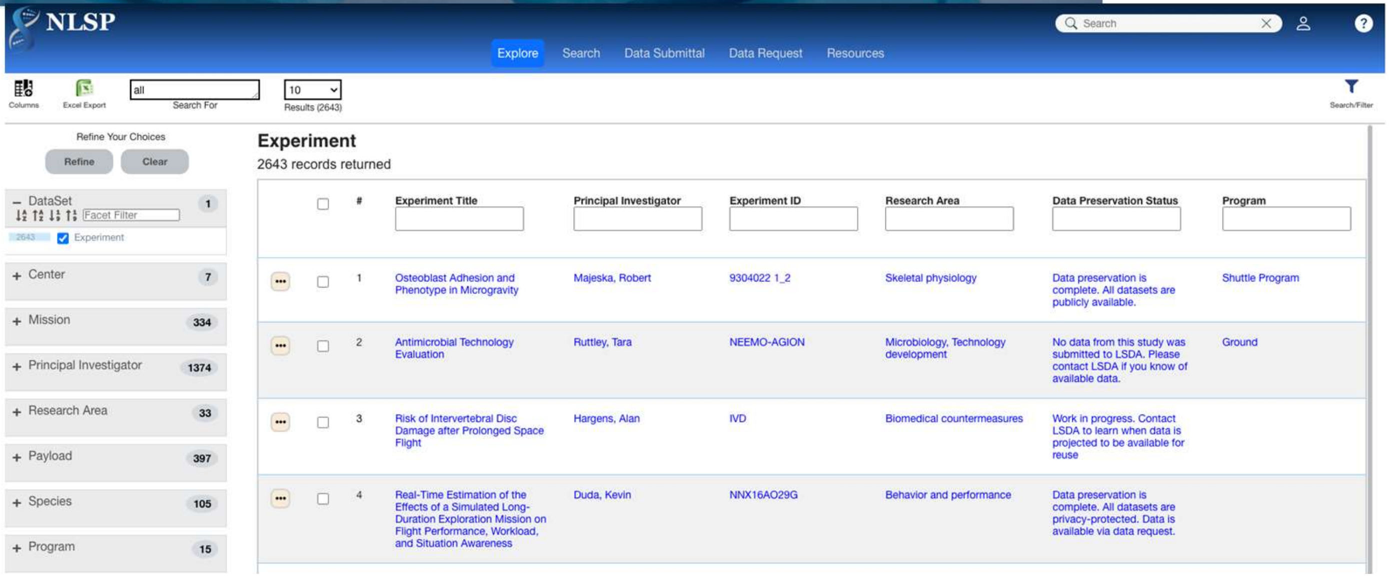
NASA Institutional Scientific Collection (ISC)

1. NASA GeneLab - the first comprehensive space-related omics database (https://genelab.nasa.gov/)
2. Investigators can search, download, submit, privately share, and/or analyze omics data from spaceflight and corresponding ground-analog experiments.
3. GeneLab Data Systems users can explore GeneLab datasets in the Data Repository, submit omics data through the Submission Portal, analyze data using GeneLab Analysis Platform tools, and Visualize study results through Visualization Portal.
4. GeneLab also offers biospecimen processing services and NASA's Institutional Scientific Collection is a space-research biobank that offers potential investigators hundreds of biospecimens for further analysis.

NASA Life Sciences Data Archive (LSDA)

The Next Generation of the Life Sciences Data Archive for Human, Animal and Plant Research

NASA's Human Research Program (HRP) conducts research and develops technologies that allow humans to travel safely and productively in space. The Program uses evidence from data collected on astronauts, as well as other supporting studies. These data are stored in the research data repository, Life Sciences Data Archive (LSDA).



Experiment

2643 records returned

	□	#	Experiment Title	Principal Investigator	Experiment ID	Research Area	Data Preservation Status	Program
•••	□	1	Osteoblast Adhesion and Phenotype in Microgravity	Majeska, Robert	9304022 1_2	Skeletal physiology	Data preservation is complete. All datasets are publicly available.	Shuttle Program
•••	□	2	Antimicrobial Technology Evaluation	Ruttley, Tara	NEEMO-AGION	Microbiology, Technology development	No data from this study was submitted to LSDA. Please contact LSDA if you know of available data.	Ground
•••	□	3	Risk of Intervertebral Disc Damage after Prolonged Space Flight	Hargens, Alan	IVD	Biomedical countermeasures	Work in progress. Contact LSDA to learn when data is projected to be available for reuse	
•••	□	4	Real-Time Estimation of the Effects of a Simulated Long-Duration Exploration Mission on Flight Performance, Workload, and Situation Awareness	Duda, Kevin	NNX16AO29G	Behavior and performance	Data preservation is complete. All datasets are privacy-protected. Data is available via data request.	

The LSDA is a collection of NASA's life sciences research data and information, consisting of human, animal, microbe, and plant studies conducted from 1958 to present.

Standardisation of metadata

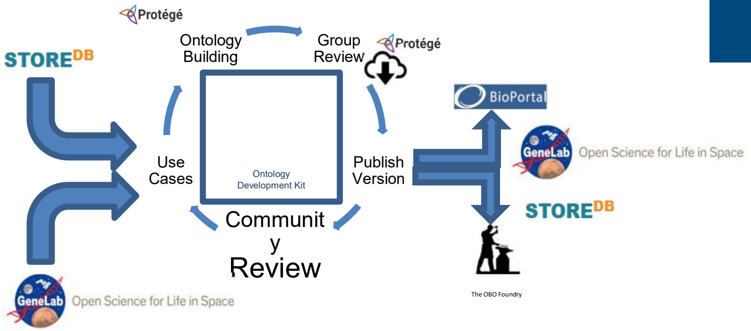
Ontologies

Comprised of standardised hierarchical concepts and relationships

Capture semantic data and metadata

Radiation Biology Ontology (RBO)

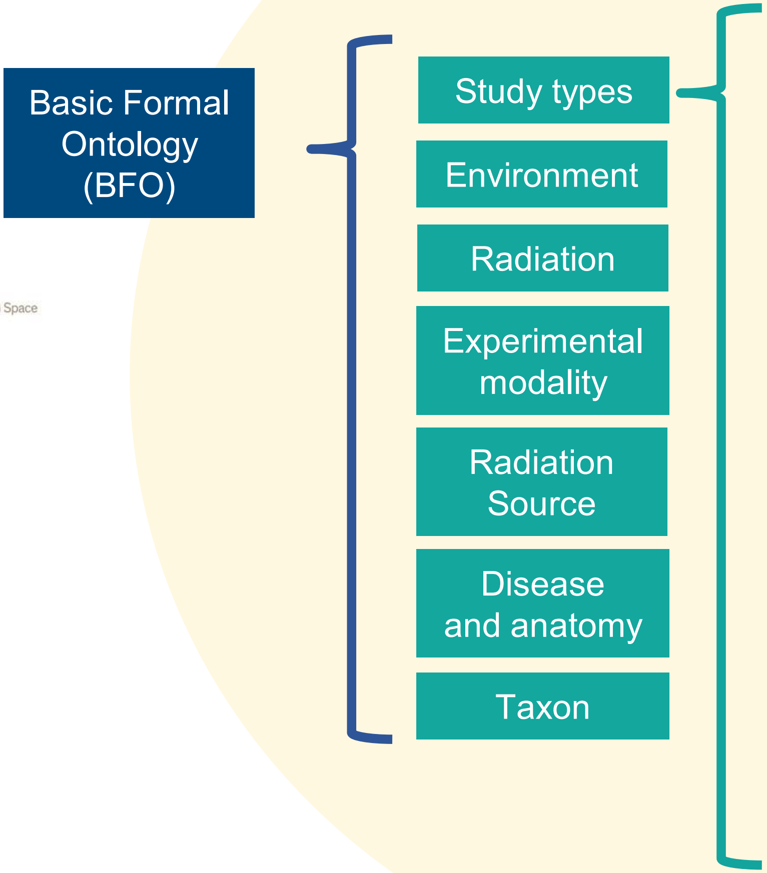
- Facilitates data retrieval and query expansion from NASA's GeneLab 'omics database and STORE

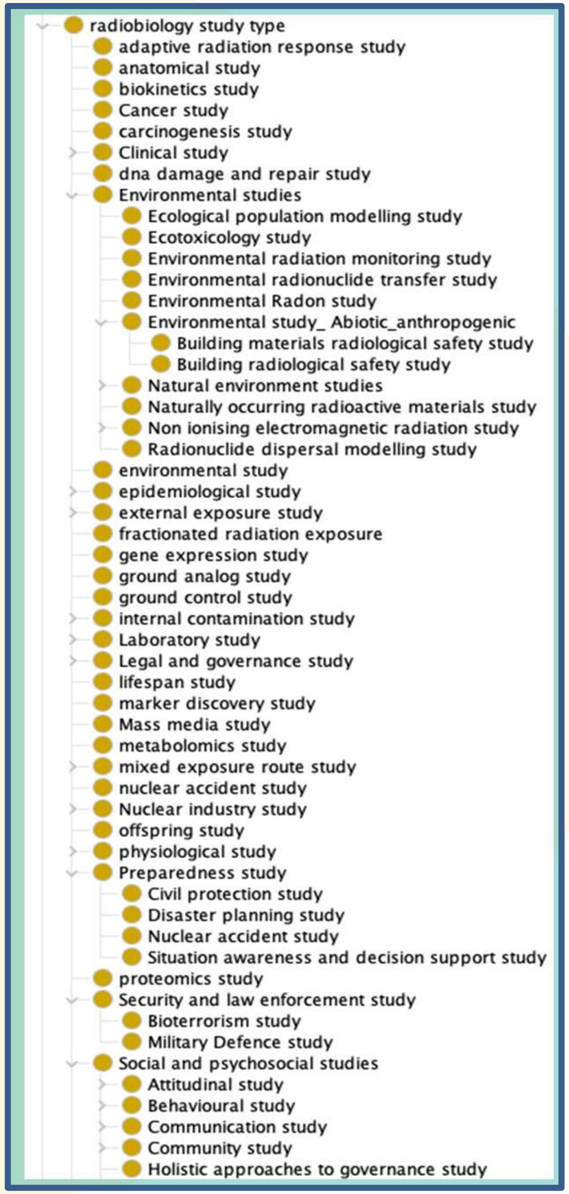


- Contains over 200 annotated classes and instances specific to the study of radiation on biological systems, as well as imports of more than 3500 additional classes from 13 other OBO Foundry ontologies

- Published through the OBO Foundry

- Available through NCBI Bioportal web site and application programming interface at https://bioportal.bioontology.org/ontologies/RBO

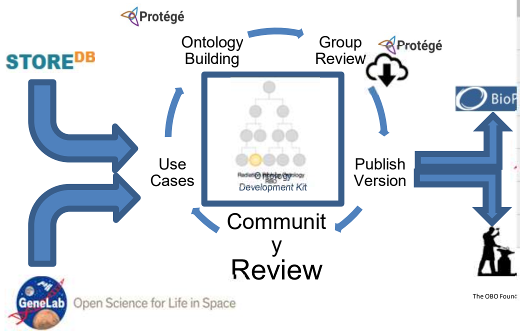




- radiobiology study type
  - adaptive radiation response study
  - anatomical study
  - biokinetics study
  - Cancer study
  - carcinogenesis study
  - Clinical study
  - dna damage and repair study
  - Environmental studies
    - Ecological population modelling study
    - Ecotoxicology study
    - Environmental radiation monitoring study
    - Environmental radionuclide transfer study
    - Environmental Radon study
    - Environmental study_ Abiotic_anthropogenic
      - Building materials radiological safety study
      - Building radiological safety study
    - Natural environment studies
    - Naturally occurring radioactive materials study
    - Non ionising electromagnetic radiation study
    - Radionuclide dispersal modelling study
  - environmental study
  - epidemiological study
  - external exposure study
  - fractionated radiation exposure
  - gene expression study
  - ground analog study
  - ground control study
  - internal contamination study
  - Laboratory study
  - Legal and governance study
  - lifespan study
  - marker discovery study
  - Mass media study
  - metabolomics study
  - mixed exposure route study
  - nuclear accident study
  - Nuclear industry study
  - offspring study
  - physiological study
  - Preparedness study
    - Civil protection study
    - Disaster planning study
    - Nuclear accident study
    - Situation awareness and decision support study
  - proteomics study
  - Security and law enforcement study
    - Bioterrorism study
    - Military Defence study
  - Social and psychosocial studies
    - Attitudinal study
    - Behavioural study
    - Communication study
    - Community study
    - Holistic approaches to governance study

Radiation Biology Ontology (RBO)

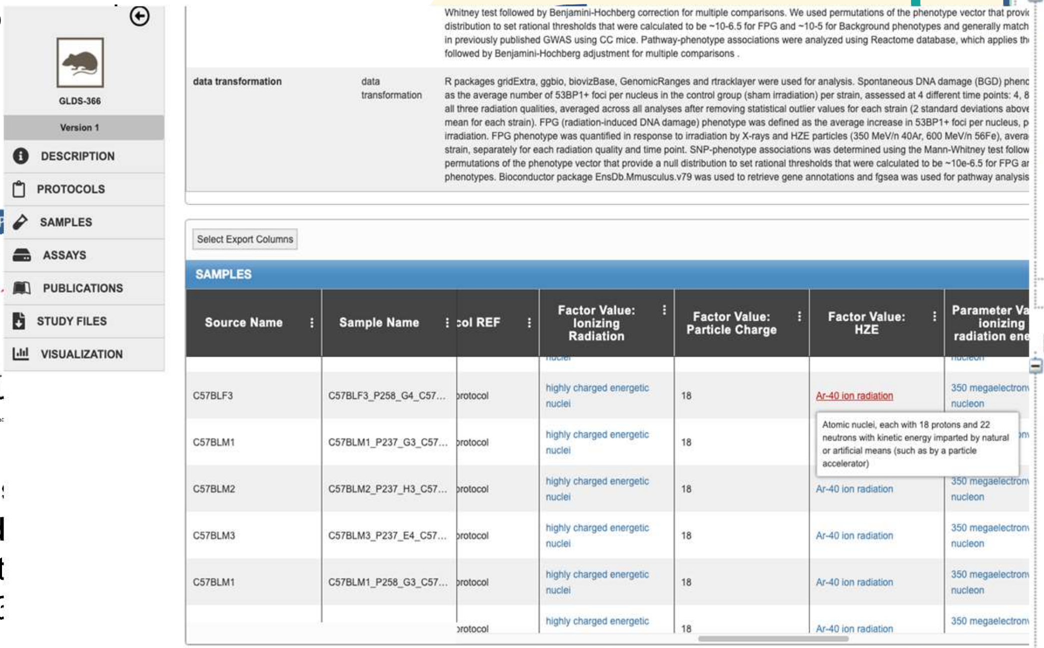
- Facilitates data retrieval and query expansion from NASA's GeneLab 'omics datab STORE



- Contains over 200 annotated classe instances specific to the study of rad biological systems, as well as import than 3500 additional classes from 13 Foundry ontologies

- Published through the OBO Foundry

- Available through NCBI Bioportal web site and application programming interface at https://bioportal.bioontology.org/ontologies/RBO



Whitney test followed by Benjamini-Hochberg correction for multiple comparisons. We used permutations of the phenotype vector that provic distribution to set rational thresholds that were calculated to be ~10-6.5 for FPG and ~10-5 for Background phenotypes and generally match in previously published GWAS using CC mice. Pathway-phenotype associations were analyzed using Reactome database, which applies th followed by Benjamini-Hochberg adjustment for multiple comparisons .

data transformation — data transformation — R packages gridExtra, ggbio, biovizBase, GenomicRanges and rtracklayer were used for analysis. Spontaneous DNA damage (BGD) phenc as the average number of 53BP1+ foci per nucleus in the control group (sham irradiation) per strain, assessed at 4 different time points: 4, 8 all three radiation qualities, averaged across all analyses after removing statistical outlier values for each strain (2 standard deviations above mean for each strain). FPG (radiation-induced DNA damage) phenotype was defined as the average increase in 53BP1+ foci per nucleus, p irradiation. FPG phenotype was quantified in response to irradiation by X-rays and HZE particles (350 MeV/n 40Ar, 600 MeV/n 56Fe), avera strain, separately for each radiation quality and time point. SNP-phenotype associations was determined using the Mann-Whitney test follow permutations of the phenotype vector that provide a null distribution to set rational thresholds that were calculated to be ~10e-6.5 for FPG ar phenotypes. Bioconductor package EnsDb.Mmusculus.v79 was used to retrieve gene annotations and fgsea was used for pathway analysis

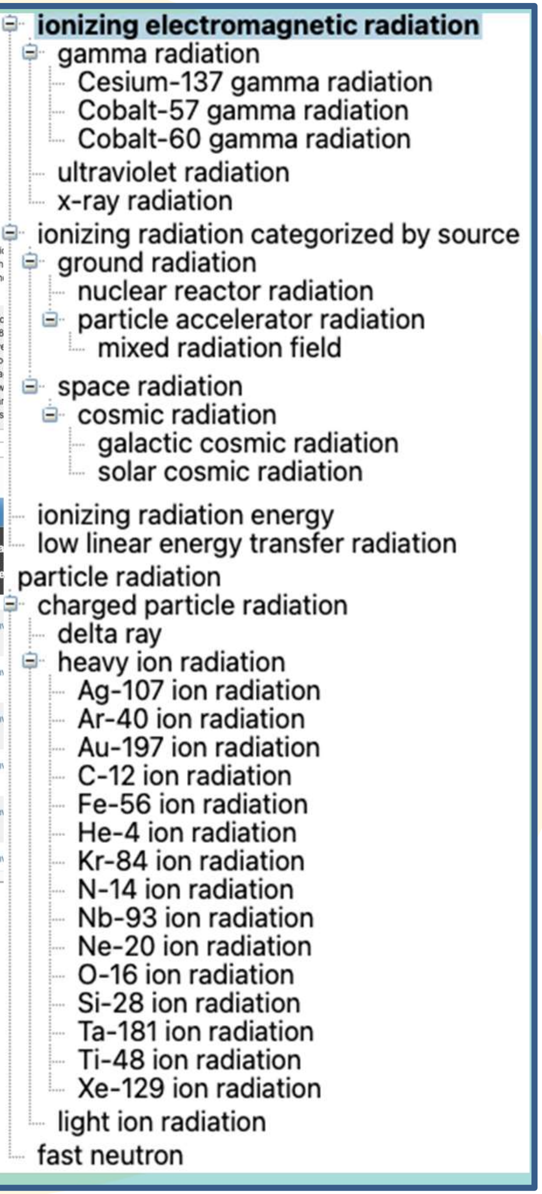
Select Export Columns

SAMPLES

Source Name	Sample Name	col REF	Factor Value: Ionizing Radiation	Factor Value: Particle Charge	Factor Value: HZE	Parameter Va ionizing radiation ene
			nuclei			nucleon
C57BLF3	C57BLF3_P258_G4_C57...	protocol	highly charged energetic nuclei	18	Ar-40 ion radiation	350 megaelectronv nucleon
C57BLM1	C57BLM1_P237_G3_C57...	protocol	highly charged energetic nuclei	18		
C57BLM2	C57BLM2_P237_H3_C57...	protocol	highly charged energetic nuclei	18	Ar-40 ion radiation	350 megaelectronv nucleon
C57BLM3	C57BLM3_P237_E4_C57...	protocol	highly charged energetic nuclei	18	Ar-40 ion radiation	350 megaelectronv nucleon
C57BLM1	C57BLM1_P258_G3_C57...	protocol	highly charged energetic nuclei	18	Ar-40 ion radiation	350 megaelectronv nucleon
		protocol	highly charged energetic	18	Ar-40 ion radiation	350 megaelectronv

Atomic nuclei, each with 18 protons and 22 neutrons with kinetic energy imparted by natural or artificial means (such as by a particle accelerator)

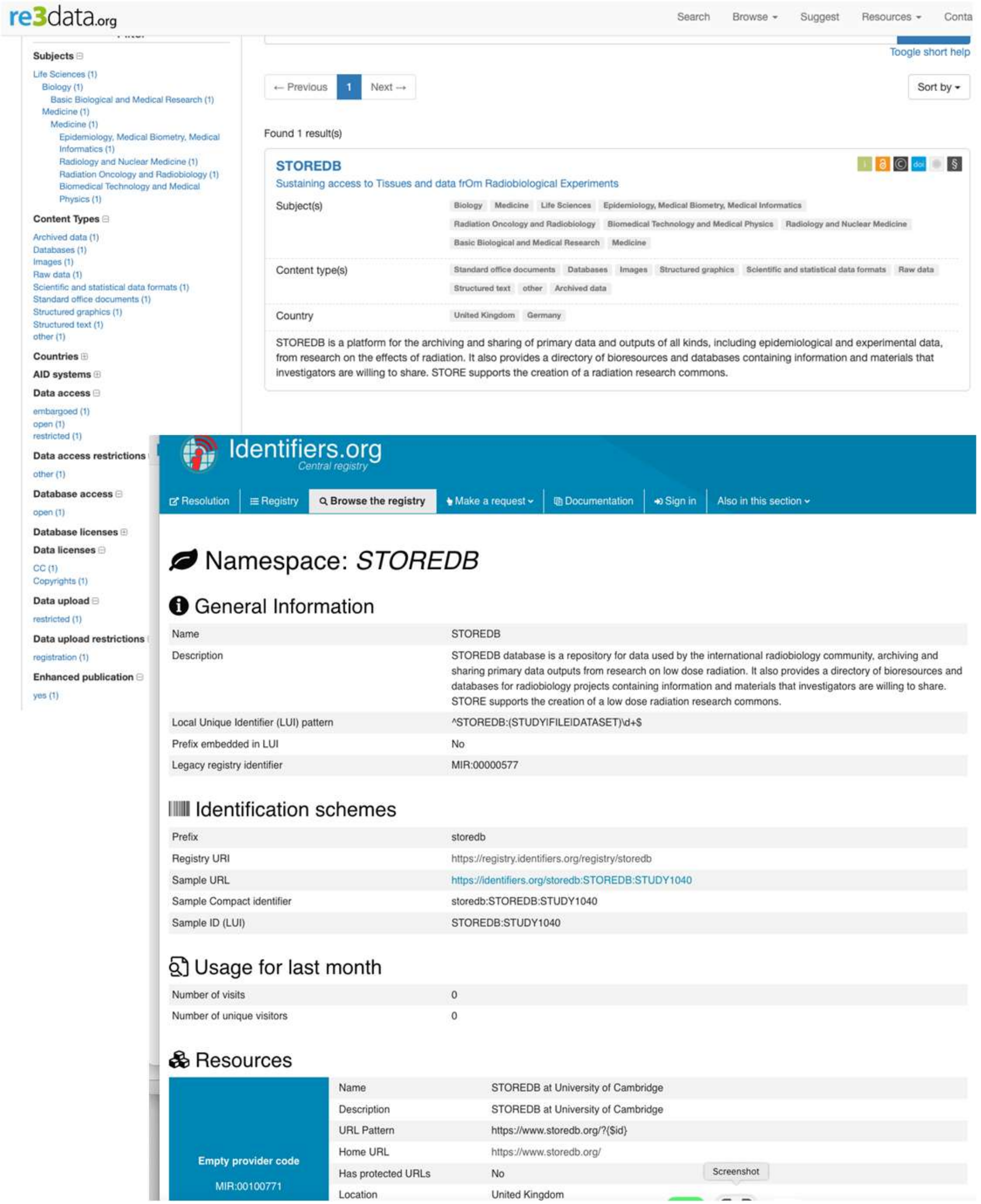
Taxon



- **ionizing electromagnetic radiation**
  - gamma radiation
    - Cesium-137 gamma radiation
    - Cobalt-57 gamma radiation
    - Cobalt-60 gamma radiation
  - ultraviolet radiation
  - x-ray radiation
- ionizing radiation categorized by source
  - ground radiation
    - nuclear reactor radiation
    - particle accelerator radiation
      - mixed radiation field
  - space radiation
    - cosmic radiation
      - galactic cosmic radiation
      - solar cosmic radiation
- ionizing radiation energy
- low linear energy transfer radiation
- particle radiation
- charged particle radiation
  - delta ray
  - heavy ion radiation
    - Ag-107 ion radiation
    - Ar-40 ion radiation
    - Au-197 ion radiation
    - C-12 ion radiation
    - Fe-56 ion radiation
    - He-4 ion radiation
    - Kr-84 ion radiation
    - N-14 ion radiation
    - Nb-93 ion radiation
    - Ne-20 ion radiation
    - O-16 ion radiation
    - Si-28 ion radiation
    - Ta-181 ion radiation
    - Ti-48 ion radiation
    - Xe-129 ion radiation
  - light ion radiation
- fast neutron

How FAIR is STORE?

- ✓ Unique identifier
- ✓ Identifier persistence
- ✓ Resolvable identifier (identifiers.org)
- ✓ Structured metadata
- ✓ Grounded metadata (resolvable metadata IDs)
- • Data identifier explicitly in metadata
- • Metadata identifier explicitly in metadata
- ✓ Programmatic access
- ✓ Licenses



Subjects ⊟

Life Sciences (1)
Biology (1)
Basic Biological and Medical Research (1)
Medicine (1)
Medicine (1)
Epidemiology, Medical Biometry, Medical Informatics (1)
Radiology and Nuclear Medicine (1)
Radiation Oncology and Radiobiology (1)
Biomedical Technology and Medical Physics (1)

Content Types ⊟

Archived data (1)
Databases (1)
Images (1)
Raw data (1)
Scientific and statistical data formats (1)
Standard office documents (1)
Structured graphics (1)
Structured text (1)
other (1)

Countries ⊞

AID systems ⊞

Data access ⊟

embargoed (1)
open (1)
restricted (1)

Data access restrictions

other (1)

Database access ⊟

open (1)

Database licenses ⊞

Data licenses ⊟

CC (1)
Copyrights (1)

Data upload ⊟

restricted (1)

Data upload restrictions

registration (1)

Enhanced publication ⊟

yes (1)

Found 1 result(s)

STOREDB

Sustaining access to Tissues and data frOm Radiobiological Experiments

Subject(s)	Biology Medicine Life Sciences Epidemiology, Medical Biometry, Medical Informatics Radiation Oncology and Radiobiology Biomedical Technology and Medical Physics Radiology and Nuclear Medicine Basic Biological and Medical Research Medicine
Content type(s)	Standard office documents Databases Images Structured graphics Scientific and statistical data formats Raw data Structured text other Archived data
Country	United Kingdom Germany

STOREDB is a platform for the archiving and sharing of primary data and outputs of all kinds, including epidemiological and experimental data, from research on the effects of radiation. It also provides a directory of bioresources and databases containing information and materials that investigators are willing to share. STORE supports the creation of a radiation research commons.

Namespace: *STOREDB*

General Information

Name	STOREDB
Description	STOREDB database is a repository for data used by the international radiobiology community, archiving and sharing primary data outputs from research on low dose radiation. It also provides a directory of bioresources and databases for radiobiology projects containing information and materials that investigators are willing to share. STORE supports the creation of a low dose radiation research commons.
Local Unique Identifier (LUI) pattern	^STOREDB:(STUDYIFILEIDATASET)\d+$
Prefix embedded in LUI	No
Legacy registry identifier	MIR:00000577

Identification schemes

Prefix	storedb
Registry URI	https://registry.identifiers.org/registry/storedb
Sample URL	https://identifiers.org/storedb:STOREDB:STUDY1040
Sample Compact identifier	storedb:STOREDB:STUDY1040
Sample ID (LUI)	STOREDB:STUDY1040

Usage for last month

Number of visits	0
Number of unique visitors	0

Resources

Empty provider code MIR:00100771	Name	STOREDB at University of Cambridge
	Description	STOREDB at University of Cambridge
	URL Pattern	https://www.storedb.org/?{$id}
	Home URL	https://www.storedb.org/
	Has protected URLs	No
	Location	United Kingdom

Work to be done

- Full implementation of the RBO ontology
- Implementation of the REST services
- Implementation of automated versioning
- Refinement of user interface
- Implement ISA-TAB to add licences, metadata and data IDs etc to data header.
- Investigate automatic extraction of candidate metadata classes from data record using NLP.

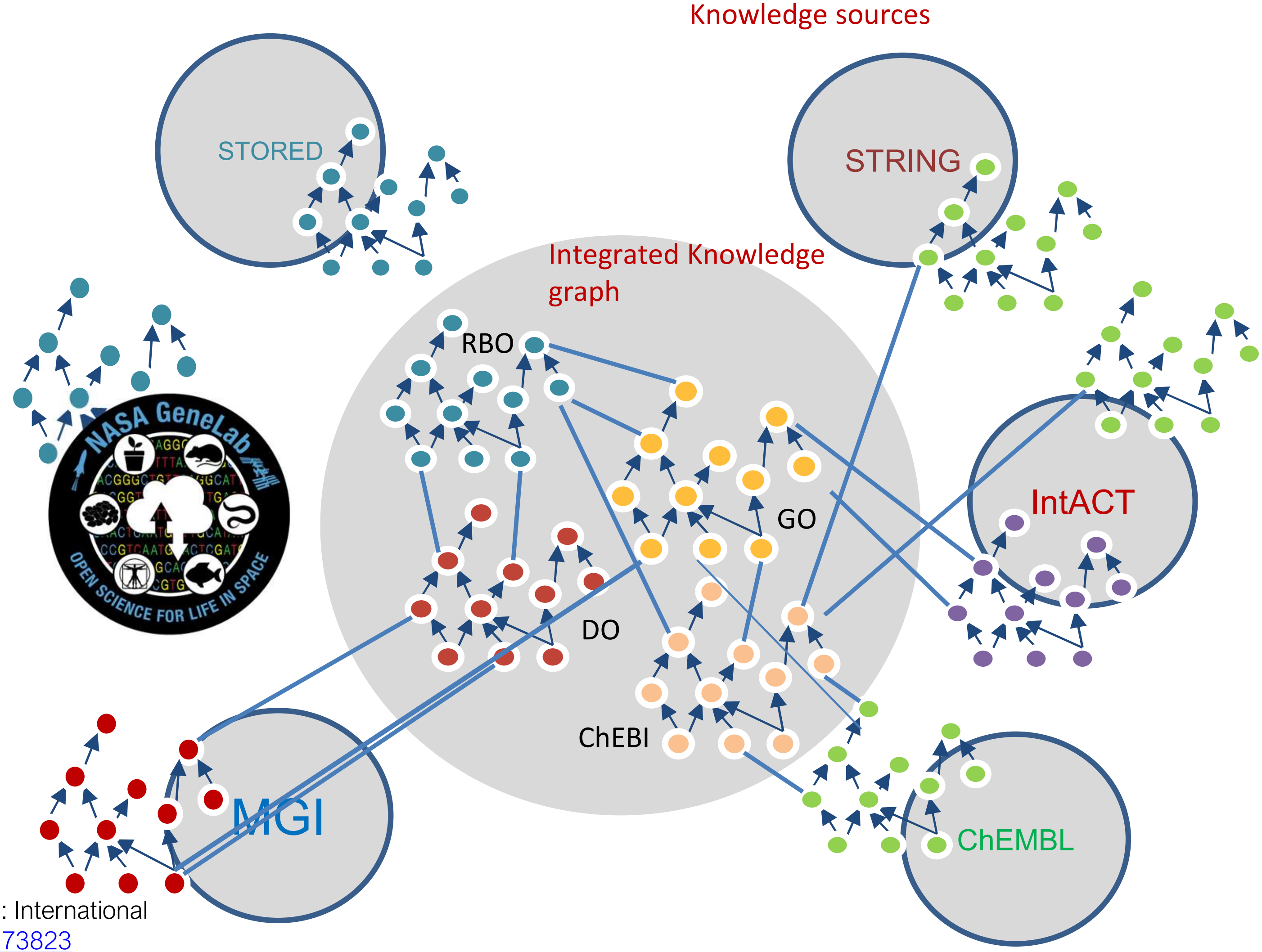
Using open data to create new knowledge

Semantic data standards and metadata

- Discovering and integrating data between databases
- Federated queries and query expansion
- Semantic integration
- Ontologies permit the assertion of defined relationships between concepts

RBO cross references 13 OBO ontologies directly providing semantic linkage to most relevant databases

- Construction of Knowledge Graphs
- Graph embeddings for data representation
- Classification and similarity
- Inductive inference
- Learning over graph convolutional neural networks



Wilson et. Al. (2023) Machine intelligence for radiation science: International Journal of Radiation Biology, DOI: 10.1080/09553002.2023.2173823

Acknowledgements

Daniel Berrios

Karin T. Slater

Jack Miller

Kristen Peach

Sylvain V. Costes

Ulrike Kulka

Michael Gruenberger

Funded by the United States National Aeronautics and Space Administration

This project has received funding from the Euratom research and training programme 2019-2020 under grant agreement Nos. 900009 (Radonorm) and 10106103 (Pianoforte)

UNIVERSITY OF CAMBRIDGE

RadoNorm
Managing risks from radon and NORM

UNIVERSITY OF BIRMINGHAM

Bundesamt für Strahlenschutz

NASA GeneLab
OPEN SCIENCE FOR LIFE IN SPACE

PIANOFORTE
EUROPEAN PARTNERSHIP FOR RADIATION PROTECTION RESEARCH