

Myths and best practice

Sarah Morgan

Melissa Burke

EMBL-EBI



Webinar series

2 August 2018

EMBL-EBI 

What do we mean by data management?

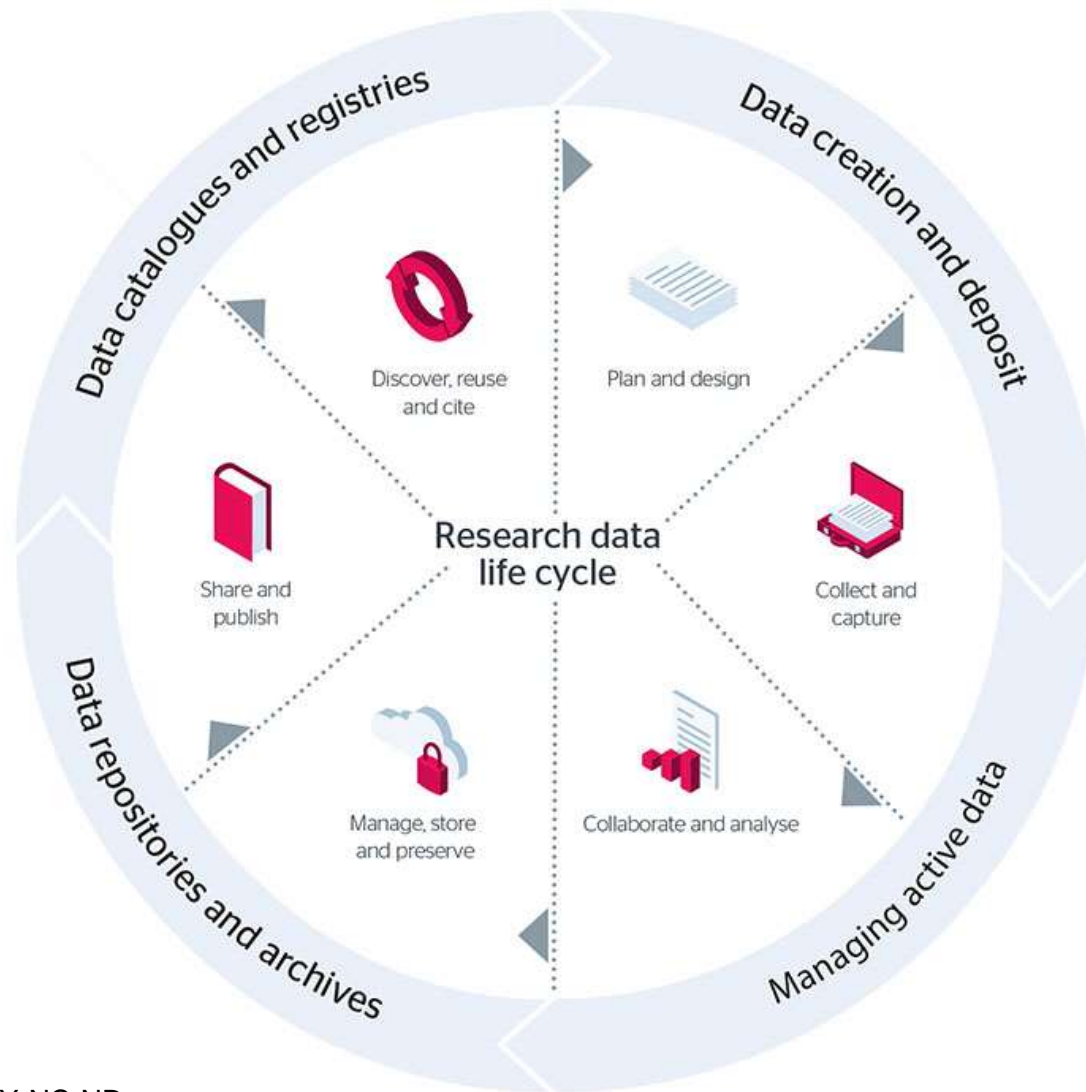


Image from [JISC](#) under CC-BY-NC-ND

Why data management?

Why do I need to worry about this?

Data management is the
bioinformatician's job

Data management is just for big
data/omics data

Decision makers realised that
rewards system needs to change



Evaluation of Research Careers fully acknowledging Open Science Practices

Rewards, incentives and/or recognition for researchers
practicing Open Science

Be the front-runner!

Be the one eligible for grants in the future!

https://cdn1.euraxess.org/sites/default/files/policy_library/os-rewards-wgreport-final_integrated_0.pdf

Selfish/altruistic benefits

We are the first (re)users of our own data

Makes accessing and using your own data easy

Give back to the community

Reduce duplication of effort

Good for your reputation

Makes your work more visible, reproducible (and citable)

Shows that you know what you're doing

You will get nice emails thanking you!



How do I manage
my data?

Make a plan

Make a plan



Plan and design

**In preparing for battle I have
always found that plans are
useless, but planning is
indispensable.**

Dwight D. Eisenhower

Make a data management plan



Things to think about:

- Collection/Documentation
- Analysis
- Storage/back-up
- Share
- Make use of checklists and online tools



[DMPonline](#)



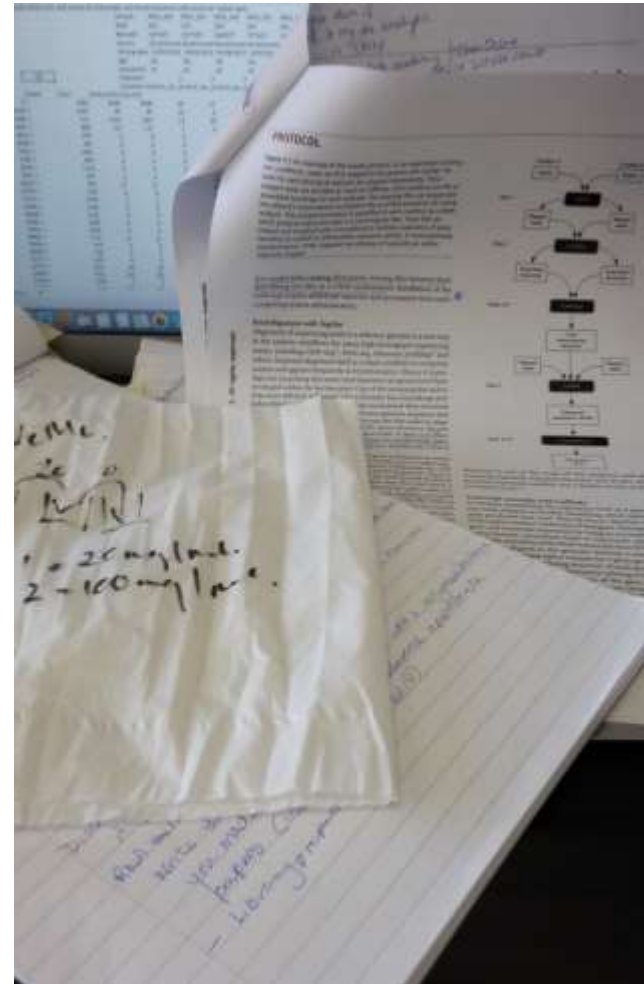
[Data stewardship wizard](#)

Talk to people

Data management made simple *Nature* **555**, 403–405 (2018) doi: 10.1038/d41586-018-03071-1

Keep good
records

Poll



Is this good data management?

Image credit: Elisabeth Busch, Melissa Burke

Collecting and storing data

What kind of data?

Where, how, what?



“Your primary collaborator is yourself six months from now, and your past self doesn’t answer e-mails”

Rachael Ainsworth, astrophysicist, University of Manchester, UK. in [Nature 555, 403-405 \(2018\) doi: 10.1038/d41586-018-03071-1](#)

F.A.I.R

F indable

A ccessible

I nteroperable

R eproducible

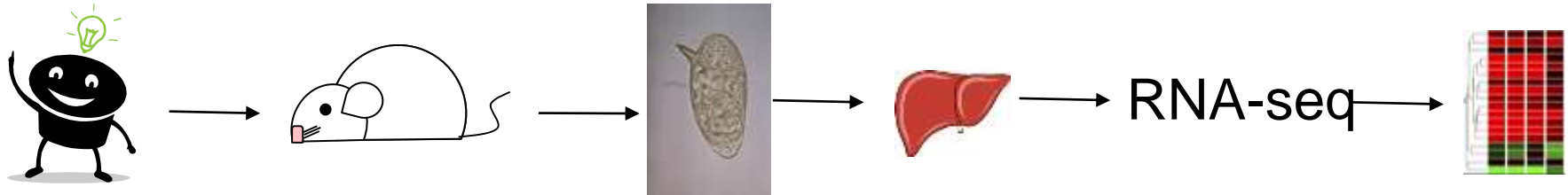


The screenshot shows the top portion of a scientific article page. At the top is a dark blue navigation bar with a 'MENU' dropdown and the 'SCIENTIFIC DATA' logo. Below this is a light blue bar containing a set of colored squares, the text 'Altmetric: 1145 Citations: 311', and a 'More detail >>' link. The main content area has a 'Comment | OPEN | Published: 15 March 2016' line. The article title is 'The FAIR Guiding Principles for scientific data management and stewardship'. The authors are listed as 'Mark D. Wilkinson, Michel Dumontier [...] Barend Mons'. At the bottom of the snippet, it says 'Scientific Data 3, Article number: 160018 (2016) | Download Citation ↓'.

<https://www.nature.com/articles/sdata201618>

What to record (metadata)

Minimal information required to unambiguously describe the experiment



- Bio background
- Experiment aim
- Who's submitting
- Main

experimental
factor/variable

Sample
annotation

Wet/dry lab protocols

Raw data
files

Processed
files

Look up metadata standards in <https://fairsharing.org/collection/MIBBI>

MINSEQE guidelines

Examples of checklists



<https://fairsharing.org/collection/MIBBI>



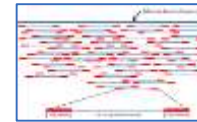
MICROARRAY: **MIAME** (2001)

Minimal **I**nformation **A**bout a **M**icroarray
Experiment
([http://www.mged.org/Workgroups/MIA
ME/miame_2.0.html](http://www.mged.org/Workgroups/MIA ME/miame_2.0.html))



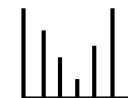
METABOLOMICS: **MSI** (2005)

MSI = **M**etabolomics **S**tandards **I**nitiative
(<http://www.metabolomics-msi.org/>)



SEQUENCING (e.g. RNA-seq):
MINSEQE (2008/2012)

Minimal **I**nformation about a high-
throughput **N**ucleotide **SEQ**uencing
Experiment
(<http://www.mged.org/minseqe>)



PROTEOMICS: **MIAPE** (2007)
Minimum **I**nformation **A**bout a **P**roteomics
Experiment
(<http://www.psidev.info/node/91>)

Keep it consistent

Sample Name	Treatment	File
Sample1	Not treated	13.fq.gz
Sample 2	INH	22.fastq.gz
Sample-3	normal	36.fq.gz
Sample_4	none	4.fq.gz
Sample#5	Isoniazid	48.fastq.gz
Sample6	Treated	Sample6.fq.gz
Sample-7	N/A	39a.fastq.gz
Sample8	C27H32N8O15P 2	11_2.fq.gz

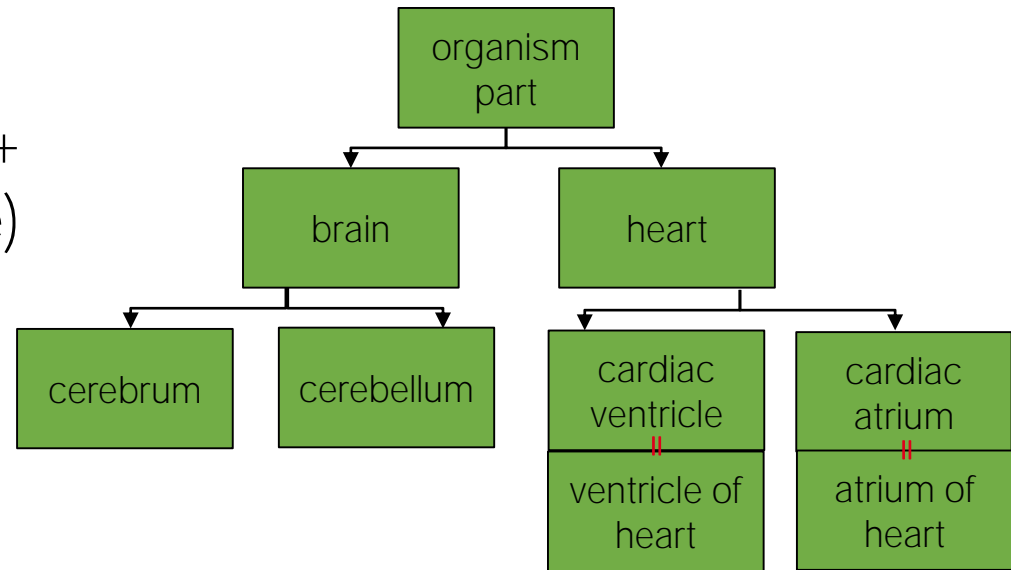
Keep it consistent

Sample Name	Treatment	File
Control_1	None	Control_1.fq.gz
Control_2	None	Control_2.fq.gz
Control_3	None	Control_3.fq.gz
Control_4	None	Control_4.fq.gz
Treated_1	Isoniazid	Treated_1.fq.gz
Treated_2	Isoniazid	Treated_2.fq.gz
Treated_3	Isoniazid	Treated_3.fq.gz
Treated_4	Isoniazid	Treated_4.fq.gz

Ontology – more than just controlled vocab

➤ Organizes information with controlled vocab + hierarchy (i.e. structure)

➤ Human + machine read-able



➤ Some databases allow you to broaden data discovery when searching by ontology terms:

- By synonyms (e.g. “human” = “man” = “Homo sapiens”)
- By child terms (e.g. “brain” → “cerebrum” and “cerebellum”)

How to find ontology terms

The screenshot shows the OLS homepage with a search bar, navigation menu, and data content statistics.

Ontology Lookup Service

Home | Ontologies | Documentation | About | Contact Us

Welcome to the EMBL-EBI Ontology Lookup Service.

Search OLS: [input field] [search icon]

Examples: diabetes, GO:0098743

[Looking for a particular ontology?](#)

Data Content

Updated 22 Apr 2016 02:41

- 148 ontologies
- 4,568,669 terms
- 12,129 properties
- 33,115 individuals

<http://www.ebi.ac.uk/ols/>

Examples of ontologies in biology

The screenshot shows the EFO homepage with a search bar and navigation menu.

Experimental Factor Ontology

Search EFO: [input field] [search icon]

Home | Browse EFO | Submit Terms | EFO RDF Platform

Representing experimental variables with EFO

The Experimental Factor Ontology (EFO) provides a systematic description of many experimental variables available in EBI databases, and for external projects such as the NHGRI GWAS catalog. It combines parts of several biological ontologies, such as UMLS/NCIT anatomy, ChEBI chemical compounds, and Cell Ontology. The scope of EFO is to support the innovation, analysis and visualization of data handled by many groups at the EBI and as the core ontology for Open Targets. We also add terms for external users when requested. If you are new to ontologies, there is a short introduction on the subject available and a blog post by James Malone on what ontologies are for.

Browse | **Submit** | **Download**

Browse EFO with EBI's OLS OR NCBO BioPortal (external). You can also search EFO using the search box, above.

Submit: new terms or report bugs using our JIRA ticket system or EFO discussion template, or join EFO mailing list

Download the latest releases of EFO in OWL format. There is an OBO format version and an aligned OBO, also. [Read the](#)

[Experimental Factor Ontology](http://www.ebi.ac.uk/ols/ontologies/efo/)

The screenshot shows the GO homepage with a search bar and navigation menu.

Gene Ontology Consortium

Home | Documentation | Downloads | Community | Tools | About | Contact Us

Enrichment analysis

Search GO data: [input field] [search icon]

Ontology

Annotations

Search documentation

User stories

What is the Gene

[Gene Ontology](http://www.geneontology.org/)

The screenshot shows the EPO homepage with a search bar and navigation menu.

GRAMENE Search

Ontologies Search | Browse | Ontology Submission | Tutorial | FAQ | Help

Ontology Search: Find: ED:0007359

Annotation Search: Find: [input field]

Ontology Type: All Ontology Types

Annotation Type: All Class Types

Options: Exact Match Include Obsolete Terms

E.g. flower or TO:0006331

E.g. dt or wavy

plant_environment_ontology Term "plant environment ontology" (EPO:0007359)

Term Name	plant environment ontology
Term Accession	EPO:0007359
Aspect	plant_environment_ontology
Synonyms (2)	plant treatment ontology; treatment ontology

A set of standardized controlled vocabularies to describe various types of treatments given to an individual plant / a population or a cultured tissue.

[Plant Environment Ontology](http://www.ebi.ac.uk/ols/ontologies/epo/)

Watch out for human error

Sample	Treatment
Culture A_control	Drug X
Culture B_treated	Drug X
Culture C_control	control
Culture D_treated	Drug X

?



Review!

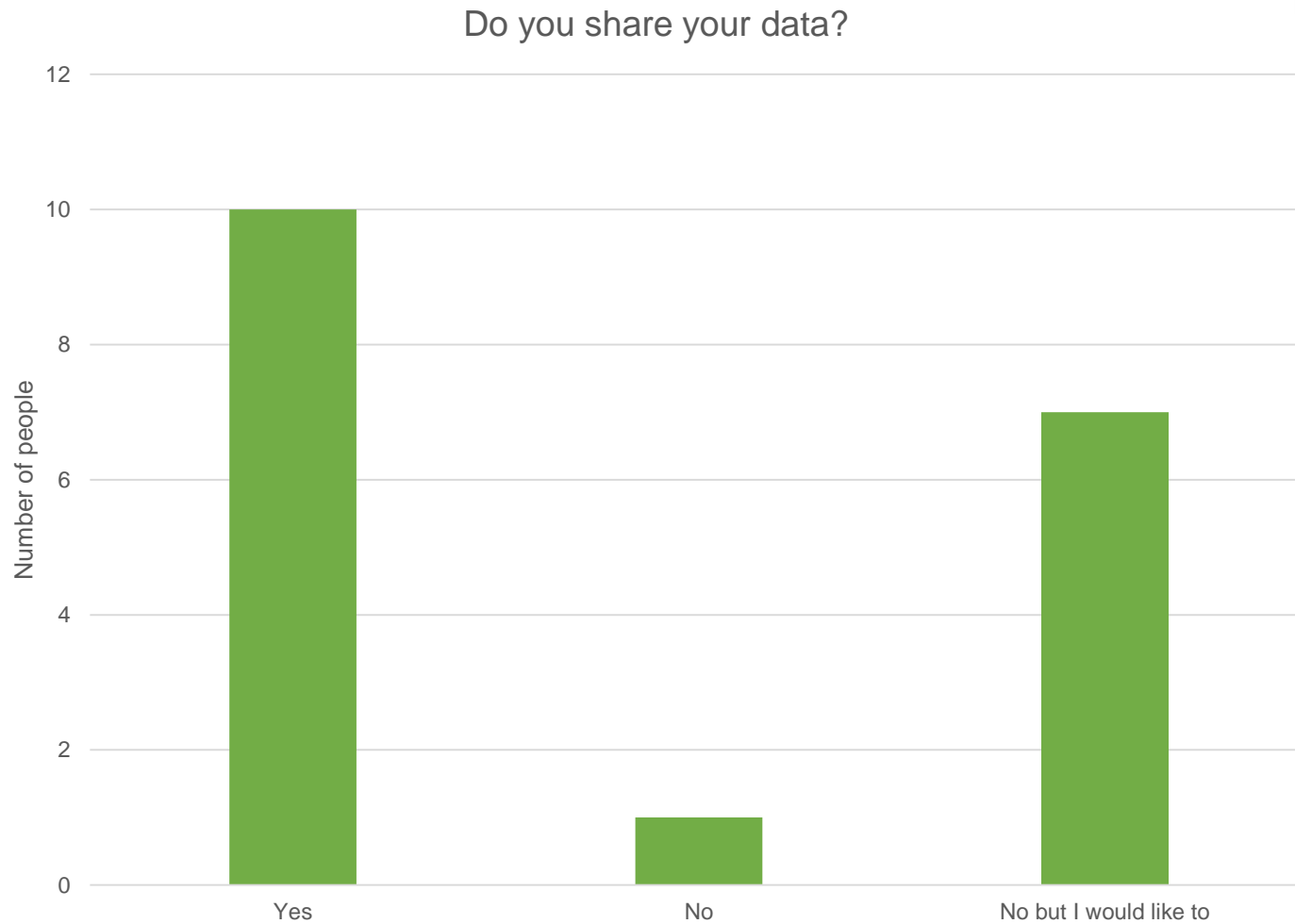


Sharing your data

Sharing your data



Share and
publish



“I can’t share data that isn’t published”

“My data will be scooped if I share it before publication”

“I won’t benefit from sharing my data”

“My files are too big to share”

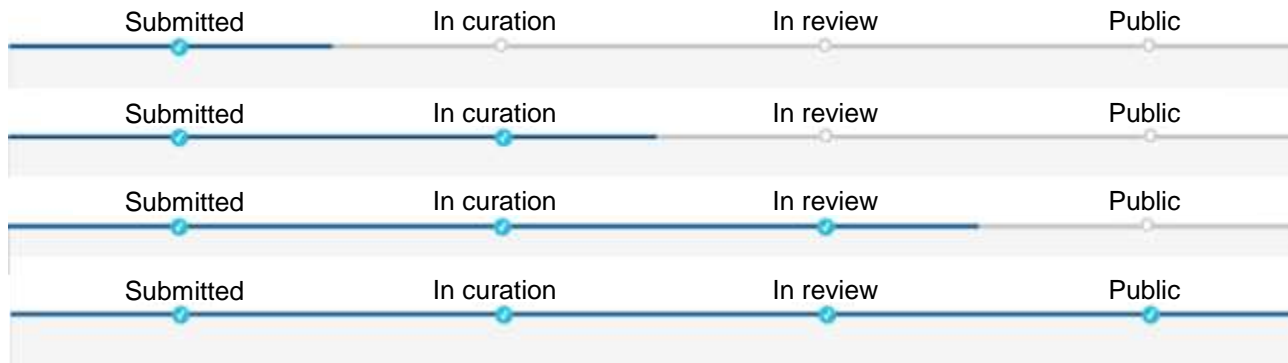
“How can we trust other scientist to cite us when we for example share our genome data before publishing our related paper?”

Things to know about sharing data

Data can be kept private



Study status



Private; submitter access only

Private; private link available for eg. journal reviewers

Public; open access

There are ways to handle BIG data



Specialised databases for human data



<https://ega-archive.org/>

How can a researcher choose the best repository?



Where to submit your data

The screenshot shows the EMBL-EBI website homepage. At the top, there is a navigation bar with links for Services, Research, Training, and About us. The main header features the EMBL-EBI logo and the text "The home for big data in biology". A prominent "Open data" banner states "EMBL-EBI shares data from life science experiments." Below this, there are several call-to-action buttons: "More about EBI Search", "Find a tool for your data analysis", and "Share your scientific data with the world." The "Share your scientific data with the world" section contains a button labeled "Deposit data" which is highlighted with a red rectangular box. A search bar is also visible with the text "Find a gene, protein or chemical" and example searches: "blast keratin bf11".

Explore EMBL-EBI and our mission

The European Bioinformatics Institute (EMBL-EBI) shares [data from life science experiments](#), performs [basic research](#) in computational biology and offers an extensive [user training programme](#), supporting researchers in [academia](#) and [industry](#). We are part of [EMBL](#), Europe's flagship laboratory for the life sciences.

[More about EMBL-EBI and our impact](#) >

Services

We provide freely available data and bioinformatics services to all facets of the scientific community >

Research

We contribute to the advancement of biology through basic investigator-driven research >

Training

We provide advanced bioinformatics training to scientists at all levels >

Industry

We help disseminate cutting-edge technologies to industry >

ELIXIR

We support, as an ELIXIR node, the coordination of biological data provision throughout Europe >

Data submission

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

1 You have **DNA/RNA sequence data** [Change](#)

2 Your data **do not** require controlled access [Change](#)

3 You have **experimental data** [Change](#)

4 You have **environmental community data** [Change](#)

 You can submit your data to the following database:


 [Metagenomics](#)

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
- [DGVs](#) > structural genetic variation data
- [EFO](#) > experimental variables
- [EGA](#) > human data that requires controlled access
- [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
- [ww-PDB 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](#).

How/When

- Start early!
- Get in touch with the database, talk to the curators
- Find out the requirements and learn about the standards, file formats etc
- Data can be kept private until it is published or you ask for it to be released (whichever occurs first)

Citing data

Data ownership

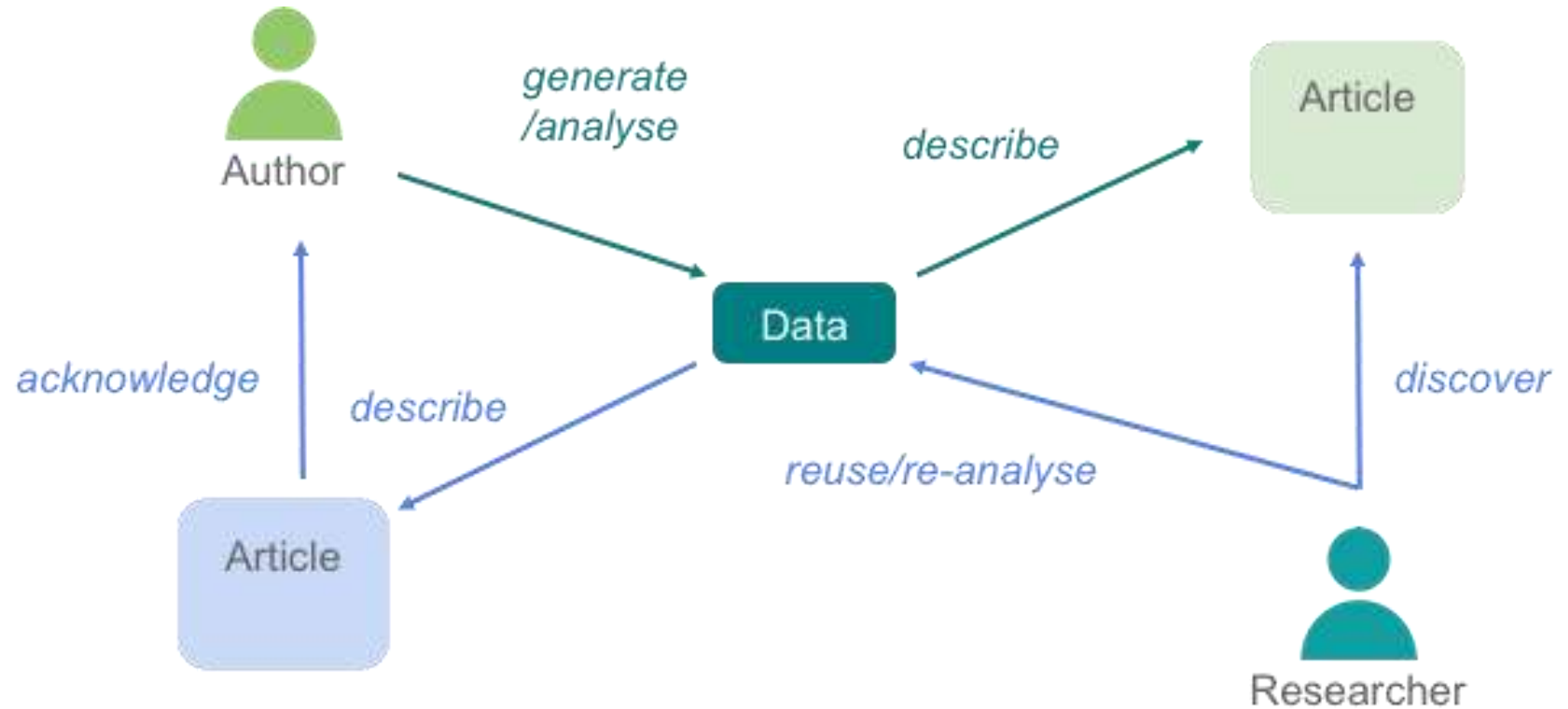


Discover, reuse
and cite

How can we trust other scientist to cite us when we for example share our genome data before publishing our related paper?

There is no need to cite the data if I am citing the publication

Giving credit to research work



Make it easy for others to cite and find you

Include the accession number/DOI and cite the database

Use your ORCID

Pre-publication data sharing

Data Reuse statement

This is a pre-publication release in accordance with [the Fort Lauderdale Agreement](#). Feel free to search and download data on your genes of interest.

Equally, you can use the dataset to show developmental expression profiles for specific genes in your publications.

However, we ask that you refrain from publishing larger scale or genome-wide analyses of this dataset for 12 months from the time of deposition in Expression Atlas or until we have published our transcriptional time-course paper, whichever comes first.

For citations in publications before the paper is out please use this link to the Expression Atlas site (<http://www.ebi.ac.uk/gxa/experiments/E-ERAD-475>) and acknowledge us: “We would like to thank the Busch-Nentwich lab for providing RNA-seq data.”

FAQs – Data ownership

- Data ownership – who owns my data when it is submitted?
- Can I link my datasets to my ORCID ID?
- Can I link to my publication/BioRxiv?
- Can I link different datasets across resources?

Ask for help

Where to get help

- [Fairsharing.org](https://fairsharing.org)
 - Information on standards and FAIR principles
- The databases and curators
- Check out support pages at your institution
- [Digital curation centre](#)
 - Further training on data management
 - Data management plan tools and advice
- Re-watch the webinars in Train online

Quick wins for data management and sharing

- Start early
- Make a plan, keep good records
- Collect metadata
- Learn about standards
- Be consistent, look out for human error
- Cite the data and the database
- Be brave, ask for help!