OPINION ARTICLE

Best practice data life cycle approaches for the life sciences
[version 1; referees: 2 approved with reservations]

Philippa C. Griffin¹,², Jyoti Khadake³, Kate S. LeMay⁴, Suzanna E. Lewis⁵, Sandra Orchard⁶, Andrew Pask⁷, Bernard Pope⁵, Ute Roessner⁸, Keith Russell⁴, Torsten Seemann⁵, Andrew Treloar⁴, Sonika Tyagi⁹,¹⁰, Jeffrey H. Christiansen¹¹, Saravanan Dayalan⁸, Simon Gladman¹, Sandra B. Hangartner¹², Helen L. Hayden¹³, William W.H. Ho⁷, Gabriel Keeble-Gagnère⁷,¹³, Pasi K. Korhonen¹⁴, Peter Neish¹⁵, Priscilla R. Prestes¹⁶, Mark F. Richardson¹⁷, Nathan S. Watson-Haigh¹⁸, Kelly L. Wyres¹⁹, Neil D. Young¹⁴, Maria Victoria Schneider²,¹⁵

¹EMBL Australia Bioinformatics Resource, The University of Melbourne, Parkville, VIC, 3010, Australia
²Melbourne Bioinformatics, The University of Melbourne, Parkville, VIC, 3010, Australia
³NIHR BioResource, University of Cambridge and Cambridge University Hospitals NHS Foundation Trust Hills Road, Cambridge, CB2 0QQ, UK
⁴Australian National Data Service, Monash University, Malvern East, VIC, 3145, Australia
⁵Lawrence Berkeley National Laboratory, Environmental Genomics and Systems Biology Division, Berkeley, CA, 94720, USA
⁶European Bioinformatics Institute (EMBL-EBI), European Molecular Biology Laboratory, Cambridge, CB10 1SD, UK
⁷School of BioSciences, The University of Melbourne, Parkville, VIC, 3010, Australia
⁸Metabolomics Australia, School of BioSciences, The University of Melbourne, Parkville, VIC, 3010, Australia
⁹Australian Genome Research Facility Ltd, Parkville, VIC, 3052, Australia
¹⁰Monash Bioinformatics Platform, Monash University, Clayton, VIC, 3800, Australia
¹¹Queensland Cyber Infrastructure Foundation and the University of Queensland Research Computing Centre, St Lucia, QLD, 4072, Australia
¹²School of Biological Sciences, Monash University, Clayton, VIC, 3800, Australia
¹³Agriculture Victoria, AgriBio, Centre for AgriBioscience, Department of Economic Development, Jobs, Transport and Resources (DEDJTR), Bundoora, VIC, 3083, Australia
¹⁴Faculty of Veterinary and Agricultural Sciences, The University of Melbourne, Parkville, VIC, 3010, Australia
¹⁵The University of Melbourne, Parkville, VIC, 3010, Australia
¹⁶Faculty of Science and Engineering, Federation University Australia, Mt Helen, VIC, 3350, Australia
¹⁷Bioinformatics Core Research Group & Centre for Integrative Ecology, Deakin University, Geelong, VIC, 3220, Australia
¹⁸School of Agriculture, Food and Wine, University of Adelaide, Glen Osmond, SA, 5064, Australia
¹⁹Department of Biochemistry and Molecular Biology, Bio21 Molecular Science and Biotechnology Institute, The University of Melbourne, Parkville, VIC, 3010, Australia

Abstract

Throughout history, the life sciences have been revolutionised by technological advances; in our era this is manifested by advances in instrumentation for data generation, and consequently researchers now routinely handle large amounts of heterogeneous data in digital formats. The simultaneous transitions towards biology as a data science and towards a 'life cycle' view of research data pose...
new challenges. Researchers face a bewildering landscape of data management requirements, recommendations and regulations, without necessarily being able to access data management training or possessing a clear understanding of practical approaches that can assist in data management in their particular research domain.

Here we provide an overview of best practice data life cycle approaches for researchers in the life sciences/bioinformatics space with a particular focus on ‘omics’ datasets and computer-based data processing and analysis. We discuss the different stages of the data life cycle and provide practical suggestions for useful tools and resources to improve data management practices.

Keywords
data sharing, data management, open science, bioinformatics, reproducibility

This article is included in the Global Open Data for Agriculture and Nutrition gateway.

This article is included in the International Society for Computational Biology Community Journal gateway.

This article is included in the EMBL-EBI gateway.

This article is included in the Science Policy Research gateway.
Corresponding authors: Philippa C. Griffin (pip.griffin@gmail.com), Maria Victoria Schneider (mvschneider@gmail.com)

Author roles: Griffin PC: Conceptualization, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; Khadake J: Writing – Review & Editing; LeMay KS: Writing – Review & Editing; Lewis SE: Writing – Review & Editing; Orchard S: Writing – Review & Editing; Pask A: Writing – Review & Editing; Pope B: Writing – Review & Editing; Roessner U: Writing – Review & Editing; Russell K: Writing – Review & Editing; Seemann T: Writing – Review & Editing; Treloar A: Writing – Review & Editing; Tyagi S: Writing – Review & Editing; Christiansen JH: Writing – Review & Editing; Dayalan S: Writing – Review & Editing; Gladman S: Writing – Review & Editing; Hangartner SB: Writing – Review & Editing; Hayden HL: Writing – Review & Editing; Ho WWH: Writing – Review & Editing; Keeble-Gagnère G: Writing – Review & Editing; Korhonen PK: Writing – Review & Editing; Neish P: Writing – Review & Editing; Prestes PR: Writing – Review & Editing; Richardson MF: Writing – Review & Editing; Watson-Haigh NS: Writing – Review & Editing; Wyres KL: Writing – Review & Editing; Young ND: Writing – Review & Editing; Schneider MV: Conceptualization, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This publication was possible thanks to funding support from the University of Melbourne and Bioplatforms Australia (BPA) via an Australian Government NCRIS investment (to EMBL-ABR).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2017 Griffin PC et al. This is an open access article distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Griffin PC, Khadake J, LeMay KS et al. Best practice data life cycle approaches for the life sciences [version 1; referees: 2 approved with reservations] F1000Research 2017, 6:1618 (https://doi.org/10.12688/f1000research.12344.1)

Introduction

Technological data production capacity is revolutionising biology\(^1\), but is not necessarily correlated with the ability to efficiently analyse and integrate data, or with enabling long-term data sharing and reuse. There are selfish as well as altruistic benefits to making research data reusable\(^2\): it allows one to find and reuse one’s own previously-generated data easily; it is associated with higher citation rates\(^3\); and it ensures eligibility for funding from and publication in venues that mandate data sharing, an increasingly common requirement (e.g. Final NIH statement on sharing research data, Wellcome Trust policy on data management and sharing, Bill & Melinda Gates Foundation open access policy). Currently we are losing data at a rapid rate, with up to 80% unavailable after 20 years\(^4\). This affects reproducibility - assessing the robustness of scientific conclusions by ensuring experiments and findings can be reproduced - which underpins the scientific method. Once access to the underlying data is lost, replicability, reproducibility and extensibility\(^5\) are reduced.

At a broader societal level, the full value of research data may go beyond the initial use case in unforeseen ways\(^6\), so ensuring data quality and reusability is crucial to realising its potential value\(^7\). The recent publication of the FAIR principles\(^8\) identifies four key criteria for high-quality research data: the data should be Findable, Accessible, Interoperable and Reusable. Whereas a traditional view of data focuses on collecting, processing, analysing and publishing results only, a life cycle view reveals the additional importance of finding, storing and sharing data\(^9\). Throughout this article, we present a researcher-focused data life cycle framework that has commonalities with other published frameworks [e.g. the DataONE Data Life Cycle, the US geological survey science data lifecycle model and \(^{10,11,12}\)], but is aimed at life science researchers specifically (Figure 1).

Learning how to find, store and share research data is not typically an explicit part of undergraduate or postgraduate training in the biological sciences\(^13\). The scope, size and complexity of datasets in many fields has increased dramatically over the last 10–20 years, but the knowledge of how to manage this data is currently limited to specific cohorts of ‘information managers’ (e.g. research data managers, research librarians, database curators and IT professionals with expertise in databases and data schemas\(^14\)). In response to institutional and funding requirements around data availability, a number of tools and educational programs have been developed to help researchers create Data Management Plans to address elements of the data lifecycle\(^15\); however, even when a plan is mandated, there is often a gap between the plan and the actions of the researcher\(^16\).

During the week of 24–28 October 2016, EMBL Australia Bioinformatics Resource (EMBL-ABR)\(^17\) led workshops on the data life cycle for life science researchers working in the plant, animal, microbial and medical domains. The workshops provided opportunities to (i) map the current approaches to the data life cycle in biology and bioinformatics, and (ii) present and discuss best practice approaches and standards for key international projects with Australian life scientists and bioinformaticians. Discussions during these workshops have informed this publication, which targets life science researchers wanting to improve their data management practice; throughout we highlight some specific data management challenges mentioned by participants.

An earlier version of this article can be found on bioRxiv (https://doi.org/10.1101/167619).

Finding data

In biology, research data is frequently published as supplementary material to articles, on personal or institutional websites, or in non-discipline-specific repositories like Figshare and Dryad\(^18\). In such cases, data may exist behind a paywall, there is no guarantee it will remain extant, and, unless one already knows it exists and its exact location, it may remain undiscovered\(^19\). It is only when a dataset is added to public data repositories, along with accompanying standardized descriptive metadata (see Collecting data), that it can be indexed and made publicly available\(^20\). Data repositories also provide unique identifiers that increase findability by enabling persistent linking from other locations and permanent association between data and its metadata.

In the field of molecular biology, a number of bioinformatics-relevant organisations host public data repositories. National and international-level organisations of this kind include the European Bioinformatics Institute (EMBL-EBI)\(^21\), the National Centre for Biotechnology Information (NCBI)\(^22\), the DNA Data Bank of Japan

![Figure 1. The Data Life Cycle framework for bioscience, biomedical and bioinformatics data that is discussed throughout this article. Black arrows indicate the ‘traditional’, linear view of research data; the green arrows show the steps necessary for data reusability. This framework is likely to be a simplified representation of any given research project, and in practice there would be numerous ‘feedback loops’ and revisiting of previous stages. In addition, the publishing stage can occur at several points in the data life cycle.](https://doi.org/10.1101/167619)
(DDBJ)\(^3\), the Swiss Institute of Bioinformatics (SIB)\(^2\), and the four
data center members of the worldwide Protein Data Bank\(^2\), which
mirror their shared data with regular, frequent updates. This shared
central infrastructure is hugely valuable to research and develop-
ment. For example, EMBL-EBI resources have been valued at over
£270 million per year and contribute to ~£1 billion in research effi-
ciencies; a 20-fold return on investment\(^2\).

Numerous repositories are available for biological data (see
Table 1 for an overview), though repositories are still lacking
for some data types and sub-domains\(^3\). Many specialised data
repositories exist outside of the shared central infrastructure men-
tioned, often run voluntarily or with minimal funding. Support for
biocuration, hosting and maintenance of these smaller-scale but
key resources is a pressing problem\(^2\). The quality of the user-
submitted data in public repositories\(^4,5\) can mean that public
datasets require extra curation before reuse. Unfortunately, due to low
uptake of established methods (see the EMBL-EBI and NCBI third-
party annotation policies and;\(^6\)) to correct the data\(^7\), the results of
extra curation may not find their way back into the repositories.
Repositories are often not easily searched by generic web
search engines\(^8\). Registries, which form a secondary layer link-
ing multiple, primary repositories, may offer a more convenient
way to search across multiple repositories for data relevant to a
researcher’s topics of interest\(^9\).

Collecting data

The most useful data has associated information about its
creation, its content and its context - called metadata. If
metadata is well structured, uses consistent element names and con-
tains element values with specific descriptions from agreed-upon
vocabulary\(^s\). It enables machine readability, aggregation, integra-
tion and tracking across datasets: allowing for Findability, Inter-
operability and Reusability\(^9,10\). One key approach in best-practice
metadata collection is to use controlled vocabularies built from
ontology terms. Biological ontologies are tools that provide
machine-interpretable representations of some aspect of biologi-
cal reality\(^10,11\). They are a way of organising and defining objects
(i.e. physical entities or processes), and the relationships between
them. Sourcing metadata element values from ontologies
ensures that the terms used in metadata are consistent and clearly
defined. There are several user-friendly tools available to assist
researchers in accessing, using and contributing to ontologies
(Table 2).

Adopting standard data and metadata formats and syntax is
critical for compliance with FAIR principles\(^12,13,14,15,16\). Biological
and biomedical research has been considered an especially chal-
enging research field in this regard, as datatypes are extremely
heterogeneous and not all have defined data standards\(^10,17\); many
existing data standards are complex and therefore difficult to
use\(^18\), or only informally defined, and therefore subject to vari-
tion, misrepresentation, and divergence over time\(^18\). Nevertheless,
well-established standards exist for a variety of biological data
types (Table 3). FAIRsharing is a useful registry of data standards
and policies that also indicates the current status of standards for
different data types and those recommended by databases and
research organisations\(^19\).

Most public repositories for biological data (see Table 1 and Storing
data section) require that minimum metadata be submitted accom-
panying each dataset (Table 4). This minimum metadata specific-
ification typically has broad community input\(^2\). Minimum metadata
standards may not include the crucial metadata fields that give the
full context of the particular research project\(^3\), so it is important
to gather metadata early, understand how to extend a minimum
metadata template to include additional fields in a structured
way, and think carefully about all the relevant pieces of metadata
information that might be required for reuse.

Processing and analysing data

Recording and reporting how research data is processed and
analysed computationally is crucial for reproducibility and
assessment of research quality\(^2\). Full reproducibility requires
access to the software, software versions, dependencies and oper-
ating system used as well as the data and software code itself\(^2\).
Therefore, although computational work is often seen as enabling
reproducibility in the short term, in the long term it is fragile and
reproducibility is limited (e.g. discussion by D. Katz, K. Hinsen
and C.T. Brown). Best-practice approaches for preserving data
processing and analysis code involve hosting source code in a
repository where it receives a unique identifier and is under ver-
sion control; where it is open, accessible, interoperable and
reusable - broadly mapping to the FAIR principles for data. Github
and Bitbucket, for example, fulfil these criteria, and Zenodo
additionally generates Digital Object Identifiers (DOIs) for
submissions and guarantees long-term archiving. Several recent
publications have suggested ways to improve current practice in
research software development\(^15,16,17,18\).

The same points hold for wet-lab data production: for full repro-
ducibility, it is important to capture and enable access to specimen
cell lines, tissue samples and/or DNA as well as reagents. Wet-lab
methods can be captured in electronic laboratory notebooks and
reported in the Biosamples database\(^9\), protocols.io or OpenWat-
ter; specimens can be lodged in biobanks, culture or museum col-
collections\(^6,54\); but the effort involved in enabling full reproducibility
remains extensive. Electronic laboratory notebooks are frequently
suggested as a sensible way to make this information openly avail-
able and archived\(^19\). Some partial solutions exist (e.g. LabTrove,
BlogMyData, Benchling and others\(^8\)), including tools for specific
domains such as the Scratchpad Virtual Research Environment for
natural history research\(^2\). Other tools can act as or be combined
to produce notebooks for small standalone code-based projects
[Boettiger, 2017\(^15\) and update], including Jupiter Notebook,
Rmarkdown, and Docker. However, it remains a challenge to imple-
ment online laboratory notebooks to cover both field/lab work and
computer-based work, especially when computer work is extensive,
involved and non-modular\(^2\). Currently, no best-practice guidelines
or minimum information standards exist for use of electronic labo-
atory notebooks\(^8\). We suggest that appropriate minimum informa-
tion to be recorded for most computer-based tasks should include
<table>
<thead>
<tr>
<th>Database/registry</th>
<th>Name</th>
<th>Description</th>
<th>Datatypes</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Database</td>
<td>OrthoDB</td>
<td>Repository for gene ortholog information</td>
<td>Protein sequences and orthologous group annotations for evolutionarily related species groups</td>
<td><a href="http://www.orthodb.org">http://www.orthodb.org</a></td>
</tr>
<tr>
<td>Database with analysis layer</td>
<td>eggNOG</td>
<td>Repository for gene ortholog information with functional annotation prediction tool</td>
<td>Protein sequences, orthologous group annotations and phylogenetic trees for evolutionarily related species groups</td>
<td><a href="http://eggnogdb.embl.de">http://eggnogdb.embl.de</a></td>
</tr>
<tr>
<td>Database</td>
<td>European Nucleotide Archive (ENA)</td>
<td>Repository for nucleotide sequence information</td>
<td>Raw next-generation DNA sequencing and genome assembly and annotation data</td>
<td><a href="http://www.ebi.ac.uk/ena">http://www.ebi.ac.uk/ena</a></td>
</tr>
<tr>
<td>Database</td>
<td>ArrayExpress</td>
<td>Repository for genomic expression data</td>
<td>RNA-seq, microarray, ChIP-seq, BSA-seq and more (see <a href="https://www.ebi.ac.uk/arrayexpress/help/experiment_types.htm">https://www.ebi.ac.uk/arrayexpress/help/experiment_types.htm</a> for full list)</td>
<td><a href="https://www.ebi.ac.uk/arrayexpress">https://www.ebi.ac.uk/arrayexpress</a></td>
</tr>
<tr>
<td>Database</td>
<td>PRIDE</td>
<td>Repository for proteomics data</td>
<td>Protein and peptide identifications, post-translational modifications and supporting evidence</td>
<td><a href="https://www.ebi.ac.uk/pride/archive">https://www.ebi.ac.uk/pride/archive</a></td>
</tr>
<tr>
<td>Database</td>
<td>Protein Data Bank (PDB)</td>
<td>Repository for protein structure information</td>
<td>3D structures of proteins, nucleic acids and complexes</td>
<td><a href="https://www.wwpdb.org">https://www.wwpdb.org</a></td>
</tr>
<tr>
<td>Database</td>
<td>Metabolights</td>
<td>Repository for metabolomics experiments and derived information</td>
<td>Metabolite structures, reference spectra and 3D model files for metabolite profiles</td>
<td><a href="https://www.ebi.ac.uk/metabolights">https://www.ebi.ac.uk/metabolights</a></td>
</tr>
<tr>
<td>Database</td>
<td>ChEBI</td>
<td>Repository for chemical entities</td>
<td>Small molecule structures and chemical data for organisms in public NCBI databases</td>
<td><a href="https://www.ebi.ac.uk/chembl">https://www.ebi.ac.uk/chembl</a></td>
</tr>
<tr>
<td>Database</td>
<td>Taxonomy</td>
<td>Repository for taxonomic classification information</td>
<td>Taxonomic classification and nomenclature data for organisms</td>
<td><a href="https://www.ebi.ac.uk/taxonomy">https://www.ebi.ac.uk/taxonomy</a></td>
</tr>
<tr>
<td>Database</td>
<td>BioStudies</td>
<td>Repository for descriptions of biological studies, with links to data in other databases and publications</td>
<td>Study descriptions and supplementary files</td>
<td><a href="https://www.ebi.ac.uk/biostudies">https://www.ebi.ac.uk/biostudies</a></td>
</tr>
<tr>
<td>Name</td>
<td>Description</td>
<td>Datatypes</td>
<td>URL</td>
<td></td>
</tr>
<tr>
<td>-------------------------------</td>
<td>-----------------------------------------------------------------------------</td>
<td>---------------------------------------------------------------------------</td>
<td>--------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td><strong>Database/registry</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Database</strong></td>
<td>Repository for information about biological samples, with links to data generated from these samples located in other databases</td>
<td>Molecular interactions and evidence type</td>
<td><a href="https://www.ebi.ac.uk/biosamples/">https://www.ebi.ac.uk/biosamples/</a></td>
<td></td>
</tr>
<tr>
<td><strong>IntAct</strong></td>
<td>Repository for protein sequence and function and evidence type, raw, processed and/or analysed, uncurated, curated (Uniprot/TrEMBL) databases</td>
<td>Raw, processed and/or analysed sequence and genotype data along with phenotype information</td>
<td><a href="http://www.ebi.ac.uk/intact">http://www.ebi.ac.uk/intact</a></td>
<td></td>
</tr>
<tr>
<td><strong>UniProtKB, SwisssProt and TrEMBL</strong></td>
<td>Repository for protein sequence and function data. Combines curated (UniProtKB/SwissProt) and automatically annotated (TrEMBL) databases</td>
<td>Protein sequences, protein function and evidence type</td>
<td><a href="http://www.uniprot.org">http://www.uniprot.org</a></td>
<td></td>
</tr>
<tr>
<td><strong>European Genome-Phenome Archive</strong></td>
<td>Repository for molecular interaction information</td>
<td>Molecular interactions and evidence type</td>
<td><a href="https://www.ebi.ac.uk/ega">https://www.ebi.ac.uk/ega</a></td>
<td></td>
</tr>
<tr>
<td><strong>EBI Metagenomics</strong></td>
<td>Repository and analysis service for metagenomics data.</td>
<td>Next-generation sequencing metagenomic and metatranscriptomics data</td>
<td><a href="https://www.ebi.ac.uk/metagenomics/">https://www.ebi.ac.uk/metagenomics/</a></td>
<td></td>
</tr>
<tr>
<td><strong>MG-RAST</strong></td>
<td>Repository and analysis service for metatranscriptomics data.</td>
<td>Next-generation sequencing metagenomic and metatranscriptomics data</td>
<td><a href="http://metagenomics.anl.gov">http://metagenomics.anl.gov</a></td>
<td></td>
</tr>
<tr>
<td><strong>Omics DI</strong></td>
<td>Data set deposited in ENA.</td>
<td>Metagenomics, metatranscriptomics data</td>
<td><a href="http://www.omicsdi.org">http://www.omicsdi.org</a></td>
<td></td>
</tr>
<tr>
<td><strong>DataMed</strong></td>
<td>Registry for biomedical dataset discovery that currently spans 66 data repositories</td>
<td>Genomic, transcriptomic, proteomic and metabolomic data</td>
<td><a href="https://datamed.org">https://datamed.org</a></td>
<td></td>
</tr>
<tr>
<td><strong>Biosharing</strong></td>
<td>Curated registry for biological database repositories across multiple research disciplines</td>
<td>Information on database repositories, data standards, and policies</td>
<td><a href="https://biosharing.org">https://biosharing.org</a></td>
<td></td>
</tr>
<tr>
<td><strong>re3data</strong></td>
<td>Registry for research data repositories across multiple research disciplines</td>
<td>Information on research data repositories, data standards, and policies</td>
<td><a href="http://re3data.org">http://re3data.org</a></td>
<td></td>
</tr>
</tbody>
</table>

**Registry**

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Datatypes</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Database/registry</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Database</strong></td>
<td>Repository for information about biological samples, with links to data generated from these samples located in other databases</td>
<td>Molecular interactions and evidence type</td>
<td><a href="https://www.ebi.ac.uk/biosamples/">https://www.ebi.ac.uk/biosamples/</a></td>
</tr>
<tr>
<td><strong>IntAct</strong></td>
<td>Repository for protein sequence and function and evidence type, raw, processed and/or analysed, uncurated, curated (Uniprot/TrEMBL) databases</td>
<td>Raw, processed and/or analysed sequence and genotype data along with phenotype information</td>
<td><a href="http://www.ebi.ac.uk/intact">http://www.ebi.ac.uk/intact</a></td>
</tr>
<tr>
<td><strong>UniProtKB, SwisssProt and TrEMBL</strong></td>
<td>Repository for protein sequence and function data. Combines curated (UniProtKB/SwissProt) and automatically annotated (TrEMBL) databases</td>
<td>Protein sequences, protein function and evidence type</td>
<td><a href="http://www.uniprot.org">http://www.uniprot.org</a></td>
</tr>
<tr>
<td><strong>European Genome-Phenome Archive</strong></td>
<td>Repository for molecular interaction information</td>
<td>Molecular interactions and evidence type</td>
<td><a href="https://www.ebi.ac.uk/ega">https://www.ebi.ac.uk/ega</a></td>
</tr>
<tr>
<td><strong>EBI Metagenomics</strong></td>
<td>Repository and analysis service for metagenomics data.</td>
<td>Next-generation sequencing metagenomic and metatranscriptomics data</td>
<td><a href="https://www.ebi.ac.uk/metagenomics/">https://www.ebi.ac.uk/metagenomics/</a></td>
</tr>
<tr>
<td><strong>MG-RAST</strong></td>
<td>Repository and analysis service for metatranscriptomics data.</td>
<td>Next-generation sequencing metagenomic and metatranscriptomics data</td>
<td><a href="http://metagenomics.anl.gov">http://metagenomics.anl.gov</a></td>
</tr>
<tr>
<td><strong>Omics DI</strong></td>
<td>Data set deposited in ENA.</td>
<td>Metagenomics, metatranscriptomics data</td>
<td><a href="http://www.omicsdi.org">http://www.omicsdi.org</a></td>
</tr>
<tr>
<td><strong>DataMed</strong></td>
<td>Registry for biomedical dataset discovery that currently spans 66 data repositories</td>
<td>Genomic, transcriptomic, proteomic and metabolomic data</td>
<td><a href="https://datamed.org">https://datamed.org</a></td>
</tr>
<tr>
<td><strong>Biosharing</strong></td>
<td>Curated registry for biological database repositories across multiple research disciplines</td>
<td>Information on database repositories, data standards, and policies</td>
<td><a href="https://biosharing.org">https://biosharing.org</a></td>
</tr>
<tr>
<td><strong>re3data</strong></td>
<td>Registry for research data repositories across multiple research disciplines</td>
<td>Information on research data repositories, data standards, and policies</td>
<td><a href="http://re3data.org">http://re3data.org</a></td>
</tr>
</tbody>
</table>
date, task name and brief description, aim, actual command(s) used, software names and versions used, input/output file names and locations, script names and locations.

During the EMBL-ABR workshop series, participants identified the data processing and analysis stage as one of the most challenging for openness. A few participants had put intensive individual effort into developing custom online lab (and code) notebook approaches, but the majority had little awareness of this as a useful goal. This suggests a gap between modern biological research as a field of data science, and biology as it is still mostly taught in undergraduate courses, with little or no focus on computational analysis, or project or data management. As reported elsewhere\textsuperscript{16-18}, this gap has left researchers lacking key knowledge and skills required to implement best practices in dealing with the life cycle of their data.

Publishing data
Traditionally, scientific publications included raw research data, but in recent times datasets have grown beyond the scope of practical inclusion in a manuscript\textsuperscript{11,44}. Selected data outputs are often included without sharing or publishing the underlying raw data\textsuperscript{11}. Journals increasingly recommend or require deposition of raw data in a public repository [e.g. \textsuperscript{59}], although exceptions have been made for publications containing commercially-relevant data\textsuperscript{60}. The current data-sharing mandate is somewhat field-dependent\textsuperscript{1,61} and also varies within fields\textsuperscript{62}. For example, in the field of bioinformatics, the UPSIDE principle\textsuperscript{63} is referred to by some journals (e.g. Bioinformatics), while others have journal- or publisher-specific policies (e.g. BMC Bioinformatics).

The vast majority of scientific journals require inclusion of processing and analysis methods in ‘sufficient detail for reproduction’ (e.g. Public Library of Science submission and data availability guidelines; International Committee of Medical Journal Editors manuscript preparation guidelines; Science instructions for authors; Elsevier Cell Press STAR Methods; and\textsuperscript{64}), though journal requirements are diverse and complex\textsuperscript{65}, and the level of detail authors provide can vary greatly in practice\textsuperscript{66,67}. More recently, many authors have highlighted that full reproducibility requires sharing data and resources at all stages of the scientific process, from raw data (including biological samples) to full methods and analysis workflows\textsuperscript{1,6,57,67}. However, this remains a challenge\textsuperscript{68,69}, as discussed in the Processing and analysing data section. To our knowledge, strategies for enabling computational reproducibility are currently not mandated by any scientific journal.

A recent development in the field of scientific publishing is the establishment of ‘data journals’: scientific journals that publish papers describing datasets. This gives authors a vehicle to accrue citations (still a dominant metric of academic impact) for data production alone, which can often be labour-intensive and expensive yet is typically not well recognised under the traditional publishing model. Examples of this article type include the Data Descriptor in Scientific Data and the Data Note in GigaScience, which do not include detailed new analysis but rather focus on describing and enabling reuse of datasets.

The movement towards sharing research publications themselves (‘Open Access Publishing’) has been discussed extensively elsewhere [e.g. \textsuperscript{22,70,71}]. Publications have associated metadata (creator, date, title etc.; see Dublin Core Metadata Initiative metadata terms) and unique identifiers (PubMed ID for biomedical and some life science journals, DOIs for the vast majority of journals; see Table 5). The ORCID system enables researchers to claim their own unique identifier, which can be linked to their publications. The use of unique identifiers within publications referring to repository records (e.g. genes, proteins, chemical entities) is not generally mandated by journals, although it would ensure a common vocabulary is used and so make scientific results more interoperable and reusable\textsuperscript{72}. Some efforts are underway to make this easier for researchers: for example, Genetics and other Genetics Society of America journals assist authors in linking gene names to model organism database entries.

Table 2. Useful ontology tools to assist in metadata collection.

<table>
<thead>
<tr>
<th>Tool</th>
<th>Task</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ontology Lookup Service</td>
<td>Discover different ontologies and their contents</td>
<td><a href="http://www.ebi.ac.uk/ols/">http://www.ebi.ac.uk/ols/</a></td>
</tr>
<tr>
<td>OBO Foundry</td>
<td>Table of open biomedical ontologies with information on development status, license and content</td>
<td><a href="http://obofoundry.org/">http://obofoundry.org/</a></td>
</tr>
<tr>
<td>Zooma</td>
<td>Assign ontology terms using curated mapping</td>
<td><a href="http://www.ebi.ac.uk/spot/zooma/">http://www.ebi.ac.uk/spot/zooma/</a></td>
</tr>
<tr>
<td>Webulous</td>
<td>Create new ontology terms easily</td>
<td><a href="https://www.ebi.ac.uk/efo/webulous/">https://www.ebi.ac.uk/efo/webulous/</a></td>
</tr>
<tr>
<td>Ontobee</td>
<td>A linked data server that facilitates ontology data sharing, visualization, and use.</td>
<td><a href="http://www.ontobee.org">http://www.ontobee.org</a></td>
</tr>
</tbody>
</table>

Storing data
While primary data archives are the best location for raw data and some downstream data outputs (Table 1), researchers also need local data storage solutions during the processing and analysis stages. Data storage requirements vary among research domains, with major challenges often evident for groups working on taxa with large genomes (e.g. crop plants), which require large storage resources, or on human data, where privacy regulations may require local data storage, access controls and conversion to
<table>
<thead>
<tr>
<th>Data type</th>
<th>Format name</th>
<th>Description</th>
<th>Reference or URL for format specification</th>
<th>URLs for repositories accepting data in this format</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw DNA/RNA sequence</td>
<td>FASTA</td>
<td>FASTA is a common text format to store DNA/RNA/Protein sequence and FASTQ combines base quality information with the nucleotide sequence. HDF5 is a newer sequence read formats used by long read sequencers e.g. PacBio and Oxford Nanopore. Raw sequence can also be stored in unaligned SAM/BAM/CRAM format.</td>
<td>41<a href="https://support.hdfgroup.org/HDF5/">https://support.hdfgroup.org/HDF5/</a> 42<a href="https://samtools.github.io/hts-specs/">https://samtools.github.io/hts-specs/</a></td>
<td><a href="https://www.ncbi.nlm.nih.gov/sra/docs/submitformats/">https://www.ncbi.nlm.nih.gov/sra/docs/submitformats/</a> <a href="http://www.ebi.ac.uk/ena/submit/data-formats">http://www.ebi.ac.uk/ena/submit/data-formats</a></td>
</tr>
<tr>
<td></td>
<td>FASTQ</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>HDF5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SAM/BAM/CRAM</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Assembled DNA sequence</td>
<td>FASTA</td>
<td>Assemblies without annotation are generally stored in FASTA format. Annotation can be integrated with assemblies in contig, scaffold or chromosome flat file format. AGP files are used to describe how smaller fragments are placed in an assembly but do not contain the sequence information themselves.</td>
<td>41<a href="http://www.ebi.ac.uk/ena/submit/contig-flat-file">http://www.ebi.ac.uk/ena/submit/contig-flat-file</a> <a href="http://www.ebi.ac.uk/ena/submit/scaffold-flat-file">http://www.ebi.ac.uk/ena/submit/scaffold-flat-file</a> <a href="https://www.ncbi.nlm.nih.gov/assembly/agp/AGP_Specification/">https://www.ncbi.nlm.nih.gov/assembly/agp/AGP_Specification/</a></td>
<td><a href="http://www.ebi.ac.uk/ena/submit/genes-sequences-submission">http://www.ebi.ac.uk/ena/submit/genes-sequences-submission</a></td>
</tr>
<tr>
<td></td>
<td>Flat file</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>AGP</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aligned DNA sequence</td>
<td>SAM/BAM</td>
<td>Sequences aligned to a reference are represented in sequence alignment and mapping format (SAM). Its binary version is called BAM and further compression can be done using the CRAM format.</td>
<td><a href="https://samtools.github.io/hts-specs/">https://samtools.github.io/hts-specs/</a></td>
<td><a href="https://www.ncbi.nlm.nih.gov/sra/docs/submitformats/#bam">https://www.ncbi.nlm.nih.gov/sra/docs/submitformats/#bam</a></td>
</tr>
<tr>
<td></td>
<td>CRAM</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gene model or genomic feature</td>
<td>GTF/GFF</td>
<td>General feature format or general transfer format are commonly used to store genomic features in tab-delimited flat text format. GFF3 is a more advanced version of the basic GFF that allows description of more complex features. BED format is a tab-delimited text format that also allows definition of how a feature should be displayed (e.g. on a genome browser). GenBank flat file format (GB/GBK) is also commonly used but not well standardised.</td>
<td><a href="https://github.com/The-Sequence-Ontology/Specifications/blob/master/gff3.md">https://github.com/The-Sequence-Ontology/Specifications/blob/master/gff3.md</a> <a href="https://genome.ucsc.edu/FAQ/FAQformat.html">https://genome.ucsc.edu/FAQ/FAQformat.html</a></td>
<td><a href="http://www.ensembl.org/info/website/upload/gff.html">http://www.ensembl.org/info/website/upload/gff.html</a> <a href="http://www.ensembl.org/info/website/upload/gff3.html">http://www.ensembl.org/info/website/upload/gff3.html</a></td>
</tr>
<tr>
<td>annotation</td>
<td>GFF3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>BED</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>GB/GBK</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data type</td>
<td>Format name</td>
<td>Description</td>
<td>Reference or URL for format specification</td>
<td>URLs for repositories accepting data in this format</td>
</tr>
<tr>
<td>------------------------------</td>
<td>-------------</td>
<td>-----------------------------------------------------------------------------</td>
<td>---------------------------------------------</td>
<td>-----------------------------------------------------</td>
</tr>
<tr>
<td>Gene functional annotation</td>
<td>GAF</td>
<td>A GAF file is a GO Annotation File containing annotations made to the GO by a contributing resource such as FlyBase or Pombase. However, the GAF standard is applicable outside of GO, e.g. using other ontologies such as PO. GAF (v2) is a simple tab-delimited file format with 17 columns to describe an entity (e.g. a protein), its annotation and some annotation metadata.</td>
<td><a href="http://geneontology.org/page/go-annotation-file-format-20">http://geneontology.org/page/go-annotation-file-format-20</a></td>
<td><a href="http://geneontology.org/page/submitting-go-annotations">http://geneontology.org/page/submitting-go-annotations</a></td>
</tr>
<tr>
<td>Genetic/genomic variants</td>
<td>VCF</td>
<td>A tab-delimited text format to store meta-information as header lines followed by information about variants position in the genome. The current version is VCF4.2</td>
<td><a href="https://samtools.github.io/hts-specs/VCFv4.2.pdf">https://samtools.github.io/hts-specs/VCFv4.2.pdf</a></td>
<td><a href="http://www.ensembl.org/info/website/upload/var.htm">http://www.ensembl.org/info/website/upload/var.htm</a></td>
</tr>
<tr>
<td>Interaction data</td>
<td>PSI-MI XML</td>
<td>Data formats developed to exchange molecular interaction data, related metadata and fully describe molecule constructs.</td>
<td><a href="http://psidev.info/groups/molecular-interactions">http://psidev.info/groups/molecular-interactions</a></td>
<td><a href="http://www.ebi.ac.uk/intact">http://www.ebi.ac.uk/intact</a></td>
</tr>
<tr>
<td>Raw metabolite profile</td>
<td>mzML</td>
<td>XML based data formats that define mass spectrometry and nuclear magnetic resonance raw data in Metabolomics</td>
<td><a href="http://www.psidev.info/mzml">http://www.psidev.info/mzml</a></td>
<td><a href="http://nmrml.org/">http://nmrml.org/</a></td>
</tr>
<tr>
<td></td>
<td>nmrML</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Protein sequence</td>
<td>FASTA</td>
<td>A text-based format for representing nucleotide sequences or protein sequences, in which nucleotides or amino acids are represented using single-letter codes.</td>
<td>41</td>
<td><a href="http://www.uniprot.org">www.uniprot.org</a></td>
</tr>
<tr>
<td>Raw proteome profile</td>
<td>mzML</td>
<td>A formally defined XML format for representing mass spectrometry data. Files typically contain sequences of mass spectra, plus metadata about the experiment.</td>
<td><a href="http://www.psidev.info/mzml">http://www.psidev.info/mzml</a></td>
<td><a href="http://www.ebi.ac.uk/pride">www.ebi.ac.uk/pride</a></td>
</tr>
<tr>
<td>Organisms and specimens</td>
<td>Darwin Core</td>
<td>The Darwin Core (DwC) standard facilitates the exchange of information about the geographic location of organisms and associated collection specimens.</td>
<td><a href="http://rs.tdwg.org/dwc/">http://rs.tdwg.org/dwc/</a></td>
<td></td>
</tr>
<tr>
<td>Name</td>
<td>Description</td>
<td>Examples of projects/databases that use this specification</td>
<td>URL</td>
<td></td>
</tr>
<tr>
<td>---------------------------</td>
<td>------------------------------------------------------------------------------</td>
<td>---------------------------------------------------------------------------------------------------------------------------</td>
<td>----------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>MINSEQE</td>
<td>Minimum Information about a high-throughput SEQuencing Experiment</td>
<td>Developed by the Functional Genomics Data Society. Used in the NCBI Sequence Read Archive, ArrayExpress</td>
<td><a href="http://fged.org/site_media/pdf/MINSEQE_1.0.pdf">http://fged.org/site_media/pdf/MINSEQE_1.0.pdf</a></td>
<td></td>
</tr>
<tr>
<td>MIMARKS</td>
<td>Minimum Information about a MARKer gene Sequence. This is an extension of MIGS/MIMS for environmental sequences</td>
<td>Developed by the Genomic Standards Consortium. Numerous adopters including NCBI/EBI/DDBJ databases</td>
<td><a href="http://wiki.gensc.org/index.php?title=MIMARKS">http://wiki.gensc.org/index.php?title=MIMARKS</a></td>
<td></td>
</tr>
<tr>
<td>MiMIX</td>
<td>Minimum Information about a Molecular Interaction eXperiment</td>
<td>Developed by the Proteomics Standards Initiative. Adopted by the IMEx Consortium databases</td>
<td><a href="http://www.psidev.info/mimix">http://www.psidev.info/mimix</a></td>
<td></td>
</tr>
<tr>
<td>MIAPE</td>
<td>Minimum Information About a Proteomics Experiment</td>
<td>Developed by the Proteomics Standards Initiative. Adopted by PRIDE, World-2DPAGE and ProteomeXchange databases</td>
<td><a href="http://www.psidev.info/miape">http://www.psidev.info/miape</a></td>
<td></td>
</tr>
<tr>
<td>Metabolomics Standards Initiative (MSI) standards</td>
<td>Minimal reporting structures that represent different parts of the metabolomics workflow</td>
<td>Developed by the Metabolomics Standards Initiative (MSI) and the Coordination of Standards in Metabolomics (COSMOS) consortium</td>
<td><a href="http://www.metabolomics-msi.org/">http://www.metabolomics-msi.org/</a></td>
<td></td>
</tr>
<tr>
<td>MIRIAM</td>
<td>Minimal Information Required In the Annotation of Models. For annotation and curation of computational models in biology</td>
<td>Initiated by the BioModels.net effort. Adopted by the EBI BioModels database and others</td>
<td><a href="http://co.mbine.org/standards/miriam">http://co.mbine.org/standards/miriam</a></td>
<td></td>
</tr>
<tr>
<td>Name</td>
<td>Description</td>
<td>Examples of projects/databases that use this specification</td>
<td>URL</td>
<td></td>
</tr>
<tr>
<td>-------------------------------</td>
<td>-----------------------------------------------------------------------------</td>
<td>------------------------------------------------------------</td>
<td>-------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>MDM</td>
<td>Minimal Data for Mapping for sample and experimental metadata for pathogen genome-scale sequence data</td>
<td>Developed by the Global Microbial Identifier Initiative and EBI. Complies with EBI ENA database submission requirements</td>
<td><a href="http://www.ebi.ac.uk/ena/submit/pathogen-data">http://www.ebi.ac.uk/ena/submit/pathogen-data</a></td>
<td></td>
</tr>
<tr>
<td>FAANG sample metadata specification</td>
<td>Metadata specification for biological samples derived from animals (animals, tissue samples, cells or other biological materials). Complies with EBI database requirements and BioSamples database formats</td>
<td>Developed and used by the Functional Annotation of Animal Genomes Consortium</td>
<td><a href="https://github.com/FAANG/faang-metadata/blob/master/docs/faang_sample_metadata.md">https://github.com/FAANG/faang-metadata/blob/master/docs/faang_sample_metadata.md</a></td>
<td></td>
</tr>
<tr>
<td>FAANG experimental metadata specification</td>
<td>Metadata specification for sequencing and array experiments on animal samples</td>
<td>Developed and used by the Functional Annotation of Animal Genomes Consortium</td>
<td><a href="https://github.com/FAANG/faang-metadata/blob/master/docs/faang_experiment_metadata.md">https://github.com/FAANG/faang-metadata/blob/master/docs/faang_experiment_metadata.md</a></td>
<td></td>
</tr>
<tr>
<td>FAANG analysis metadata specification</td>
<td>Metadata specification for analysis results</td>
<td>Developed and used by the Functional Annotation of Animal Genomes Consortium. NB no public repository exists for this specific datatype</td>
<td><a href="https://github.com/FAANG/faang-metadata/blob/master/docs/faang_analysis_metadata.md">https://github.com/FAANG/faang-metadata/blob/master/docs/faang_analysis_metadata.md</a></td>
<td></td>
</tr>
<tr>
<td>SNOMED-CT</td>
<td>Medical terminology and pharmaceutical product standard</td>
<td>Commercial but collaboratively-designed product</td>
<td><a href="http://www.snomed.org/snomed-ct">http://www.snomed.org/snomed-ct</a></td>
<td></td>
</tr>
<tr>
<td>Name</td>
<td>Relevant stage of data life cycle</td>
<td>Description</td>
<td>URL</td>
<td></td>
</tr>
<tr>
<td>------------------------------------------------</td>
<td>-----------------------------------</td>
<td>-----------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>Digital Object Identifier (DOI)</td>
<td>Publishing, Sharing, Finding</td>
<td>A unique identifier for a digital (or physical or abstract) object</td>
<td><a href="https://www.doi.org/">https://www.doi.org/</a></td>
<td></td>
</tr>
<tr>
<td>Open Researcher and Contributor ID (ORCID)</td>
<td>Publishing</td>
<td>An identifier for a specific researcher that persists across publications and other research outputs</td>
<td><a href="https://orcid.org/">https://orcid.org/</a></td>
<td></td>
</tr>
<tr>
<td>Repository accession number</td>
<td>Finding, Processing/ Analyzing, Publishing, Sharing, Storing</td>
<td>A unique identifier for a record within a repository. Format will be repository-specific. Examples include NIH UIDs (unique identifiers) and accession numbers; ENA accession numbers; PDB IDs</td>
<td>For example, <a href="https://support.ncbi.nlm.nih.gov/link/portal/28045/28049/Article/499/">https://support.ncbi.nlm.nih.gov/link/portal/28045/28049/Article/499/</a> <a href="http://www.ebi.ac.uk/ena/submit/accession-number-formats">http://www.ebi.ac.uk/ena/submit/accession-number-formats</a></td>
<td></td>
</tr>
<tr>
<td>International Standard Serial Number (ISSN)</td>
<td>Publishing</td>
<td>A unique identifier for a journal, magazine or periodical</td>
<td><a href="http://www.issn.org/">http://www.issn.org/</a></td>
<td></td>
</tr>
</tbody>
</table>
non-identifiable data if data is to be shared (see the Australian National Data Service de-identification guide, the National Health and Medical Research Council statement on ethical conduct in human research, and the Australian National Medical Research Storage Facility discussion paper on legal, best practice and security frameworks). In addition, long-term preservation of research data should consider threats such as storage failure, mistaken erasure, bit rot, outdated media, outdated formats, loss of context and organisational failure\(^3\).

**Sharing data**

The best-practice approach to sharing biological data is to deposit it (with associated metadata) in a primary archive suitable for that datatype\(^7\) that complies with FAIR principles. As highlighted in the Storing data section, these archives assure both data storage and public sharing as their core mission, making them the most reliable location for long-term data storage. Alternative data sharing venues (e.g. FigShare, Dryad) do not require or implement specific metadata or data standards. This means that while these venues have a lower barrier to entry for submitters, the data is not FAIR unless submitters have independently decided to comply with more stringent criteria. If available, an institutional repository may be a good option if there is no suitable archive for that datatype. Importantly, plans for data sharing should be made at the start of a research project and reviewed during the project, to ensure ethical approval is in place and that the resources and metadata needed for effective sharing are available at earlier stages of the data life cycle\(^7\).

During the EMBL-ABR workshop series, the majority of participants were familiar with at least some public primary data repositories, and many had submitted data to them previously. A common complaint was around usability of current data submission tools and a lack of transparency around metadata requirements and the rationale for them. A few workshop participants raised specific issues about the potential limitations of public data repositories where their data departed from the assumptions of the repository (e.g. unusual gene models supported by experimental evidence that were rejected by the automated NCBI curation system). Most workshop participants were unaware they could provide feedback to the repositories to deal with such situations, and this could also be made clearer on the repository websites. Again, this points in part to existing limitations in the undergraduate and postgraduate training received by researchers, where the concepts presented in this article are presented as afterthoughts, if at all. On the repository side, while there is a lot of useful information and training material available to guide researchers through the submission process (e.g. the EMBL-EBI Train Online webinars and online training modules), it is not always linked clearly from the database portals or submission pages themselves. Similarly, while there are specifications and standards available for many kinds of metadata [Table 4; also see FAIRsharing], many do not have example templates available, which would assist researchers in implementing the standards in practice.

**What can the research community do to encourage best practice?**

We believe that the biological/biomedical community and individual researchers have a responsibility to the public to help advance knowledge by making research data FAIR for reuse\(^8\), especially if the data were generated using public funding. There are several steps that can assist in this mission:

1. **Senior scientists should lead by example** and ensure all the data generated by their laboratories is well-managed, fully annotated with the appropriate metadata and made publicly available in an appropriate repository.

2. **The importance of data management and benefits of data reuse should be taught** at the undergraduate and postgraduate levels\(^9\). Computational biology and bioinformatics courses in particular should include material about data repositories, data and metadata standards, data discovery and access strategies. Material should be domain-specific enough for students to attain learning outcomes directly relevant to their research field.

3. Funding bodies are already taking a lead role in this area by requiring the incorporation of a data management plan into grant applications. A next step would be for a formal check, at the end of the grant period, that this plan has been adhered to and data is available in an appropriate format for reuse\(^10\).

4. **Funding bodies and research institutions should judge quality dataset generation as a valued metric when evaluating grant or promotion applications.**

5. Similarly, leadership and participation in community efforts in data and metadata standards, and open software and workflow development should be recognised as academic outputs.

6. **Data repositories should ensure that the data deposition and third-party annotation processes are as FAIR and painless as possible** to the naive researcher, without the need for extensive bioinformatics support\(^3\).

7. **Journals should require editors and reviewers to check manuscripts to ensure that all data, including research software code and samples where appropriate, have been made publicly available in an appropriate repository,** and that methods have been described in enough detail to allow re-use and meaningful reanalysis\(^6\).

8. **Finally, researchers reusing any data should openly acknowledge this fact and fully cite the dataset, including unique identifiers\(^4,8,30\).**
Conclusions
While the concept of a life cycle for research data is appealing from an Open Science perspective, challenges remain for life science researchers to put this into practice. During the EMBL-ABR Data Life Cycle workshop series, we noted limited awareness among attendees of the resources available to researchers that assist in finding, collecting, processing, analysis, publishing, storing and sharing FAIR data. We believe this article provides a useful overview of the relevant concepts and an introduction to key organisations, resources and guidelines to help researchers improve their data management practices.

Furthermore, we note that data management in the era of biology as a data science is a complex and evolving topic and both best practices and challenges are highly domain-specific, even within the life sciences. This factor may not always be appreciated at the organisational level, but has major practical implications for the quality and interoperability of shared life science data. Finally, domain-specific education and training in data management would be of great value to the life science research workforce, and we note an existing gap at the undergraduate, postgraduate and short course level in this area.

Competing interests
No competing interests were disclosed.

Grant information
This publication was possible thanks to funding support from the University of Melbourne and Bioplatforms Australia (BPA) via an Australian Government NCRIS investment (to EMBL-ABR).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements
The authors thank Dan Bolser for his involvement in the EMBL-ABR Data Life Cycle workshops, and all workshop participants for sharing their experiences and useful discussions.

References
The article "Best practice data life cycle approaches for the life sciences", submitted by Griffin et al. reports opinions on how to best manage the growing complexity of scientific data in the life sciences.

The article touches on an extremely important topic that is currently very purely covered in the literature. In fact, data-driven approaches in the biosciences will strongly rely on professional concepts of data management. In brief, I recommend the indexing of the article, as we urgently need stronger awareness of this topic, upon the implementation of some (probably rather minor) changes to the article. The article nicely illustrates the needs in data life cycle management and also suggests best concepts to be followed by researchers. The main content of the article has been compiled based on a workshop that was attended by the authors. At some statements the article reads like the minutes of this meeting; I suggest editing the corresponding paragraphs to avoid the impression of reading meeting minutes.

I suggest the following issues to be fixed before indexing:

- Figure 1: This illustration is very important and can be used by many readers. I suggest to use figures wherever possible to replace the words such as “finding”, “integrating”, …

- The reference to Figure 1 in the second paragraph states that it illustrates a specific aim to the life sciences. I don’t see which of these points should be specific to the life science, but would rather argue that these principles are rather generic and provides a cycle for business intelligence processes in general. It might also be a good location to reference the DAMA (Data management association internation, dama.org) and specifically to the DAMA Body of Knowledge, which is one of the few references for data management and also data life cycle considerations. Further needed references should hint to the Global Alliance for Genomics and Health (ga4gh.org).

- Page 13: The paragraph on data sharing missing some discussion on authentication issues. I would like see some introduction and discussion to the OpenID concept. Especially for medical data there need to be appropriate mechanisms to trace users, concepts for data privacy and so on. As a best practice use case for these topics, the mechanism from ICGC could be introduced.

- The following paragraph states: “A few workshop participants…”. Rephrase, no meeting minutes..

- I would have loved to see more use cases/examples for the individual best practices. E.g. for the data sharing the ICGC efforts could be described more thoroughly.
The article would benefit for 2-3 additional figures. I guess it could be a nice figure to illustrate the concept of controlled vocabularies and/or ontologies. While this seems to be trivial for bioinformaticians/computer scientists, it is not that obvious what it means to non-computer scientists; inspiration for figures can also be obtained by the data sharing mechanisms for the Global alliance for Genomics and Health

Minor things:
- The forth paragraph in the introduction starts with “During the week of 24-28 October 2016…”. I suggest either avoiding that paragraph or formulating it differently. The reader should not be reading the meeting minutes.

Is the topic of the opinion article discussed accurately in the context of the current literature? 
Yes

Are all factual statements correct and adequately supported by citations?
Yes

Are arguments sufficiently supported by evidence from the published literature?
Yes

Are the conclusions drawn balanced and justified on the basis of the presented arguments?
Partly

Competing Interests: No competing interests were disclosed.

Referee Expertise: Data management, multi-omics bioinformatics

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Johannes Starlinger

1 Department of Anesthesiology and Operative Intensive Care Medicine, Charité – Universitätsmedizin Berlin, Berlin, Germany
2 Department of Computer Science, Humboldt-Universität zu Berlin, Berlin, Germany

The article gives a brief overview of the data life cycle in the life sciences and offers an entry point for accessing relevant information about current approaches to increasing compliance with the FAIR data sharing principles at each step of this life cycle. It expressly targets "life science researchers wanting to improve their data management practice" and is labeled as an Opinion article.

The article is well written and comfortable to read, and the concise presentation follows a clear structure. While to me as a biomedical data researcher, who may not strictly belong to the target audience, the article provided only little additional insight, I can well see how - as an entry point - the article provides
valuable information to its target audience.

That said, I believe the article needs clarification and some extension in a few places:

- The list of authors is quite extensive. Please clarify the roles of the authors in conception/conduction/preparation of the manuscript.

- How exactly does the proposed data life cycle differ from related (cited) suggestions, and why? How is it 'aimed at life science researchers specifically'? (Introduction)

- The tabular overviews of existing resources are a nice asset but they are, of course, not exhaustive. Please clarify how the selections of databases/registries, tools, ontologies etc were made for inclusion in the article - and possibly state where to find more complete lists of resources for the life sciences.

- The integrating step of the life cycle has no description in the article - even though this is a very intricate step that often has great influence when collecting data (e.g., the choice of ontologies to use for describing collected data and metadata will often depend on the ontologies used in re-used (found) data), and, even more, is at the core of making datasets interoperable, i.e., making them integratable with newly collected data.

- In the processing step, you make no mention of Scientific Workflows as a means of integrating, processing, and analyzing data. Your first reference (currently cited in a rather different context) would provide a very good hook for this thriving topic that is all about sharing, reproducibility, and reusability of data processing and analysis methods. On the same lines, containerized computing (e.g., Docker) is only very briefly metioned. Even more than with data, using technologies such as these is crucial for ensuring reproducibility over longer periods of time (when software versions of dependencies have changed, web-services have become unavailable, and so forth).

- The section "What can the research community do to encourage best practice?" gives a rather remote, high level view that addresses several different institutional entities - except for the individual researcher within the target audience who actually has to follow the discussed best practices to enable the data life cycle.

Additionally, here are some suggestions for increasing the usefulness and potential impact of the article within the current target audience, and possibly beyond:

- Important interdependencies between the different steps of the life cycle could be mentioned. For instance, the choice of which ontologies to use for metadata and data in the collection step will necessarily be influenced by a) the ontologies used in the data found in public repositories and reused in the current experiment, b) the ontologies mandated by the repositories the data product is to be published in, and c) the ontologies required and used by the (third party, reused) software applied in the processing of the data. These interdependencies often not only put a limit to the choices available regarding the ontologies to be used but also raise a barrier when conversion and mapping between different ontologies is necessary between steps in the life cycle.

- The topic of data privacy is only very briefly touched but fundamental when it comes to sharing and publishing data. It may be out of scope of this article, but a slightly more thorough discussion of the issue would to its importance more justice, I feel.
An additional figure that maps the best practices enumerated in the text to the rather coarse life cycle shown in Figure 1 could prove highly instructive. Something like a 'data life cycle best practices cheat sheet' ;)

If you (the authors) have any questions regarding this review, please do not hesitate to contact me.

**Is the topic of the opinion article discussed accurately in the context of the current literature?**
Yes

**Are all factual statements correct and adequately supported by citations?**
Yes

**Are arguments sufficiently supported by evidence from the published literature?**
Yes

**Are the conclusions drawn balanced and justified on the basis of the presented arguments?**
Yes

**Competing Interests:** No competing interests were disclosed.

**Referee Expertise:** Biomedical knowledge management, systems architectures, clinical informatics

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

---

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com