## Molecular Plant Editorial

# FAIR: A Call to Make Published Data More Findable, Accessible, Interoperable, and Reusable

Biology has become an information-dense, data-intensive enterprise that requires the ability to access, integrate, and compute on large amounts of data. To enable better use of data, stakeholders (scientists, curators, librarians, and journal publishers) have established standards to ensure that data are findable, accessible, interoperable, and reusable (FAIR; Wilkinson et al., 2016; https://www.force11.org/group/fairgroup/fairprinciples). In a nutshell, FAIR means:

- (1) Findable data are human and machine readable and attached to persistent identifiers
- (2) Accessible data can be found and retrieved by humans and machines using standard formats
- (3) Interoperable data can be exchanged and used between systems
- (4) Reusable data can be used by others

Good data stewardship is essential to maximize data access and reuse, and to ensure reproducibility in plant sciences (Leonelli et al., 2017), but researchers typically do not have explicit training in data management. Funding agencies increasingly require that proposals include data management and sustainability plans as well as evidence of compliance. Thus, researchers will need to learn how to prepare a data management plan (Schiermeier, 2018). As the research and scholarly communications communities become familiar with and embrace FAIR principles, researchers will also need guidance on how to implement these recommendations for their published data. Scholarly publications are the primary sources by which research results are disseminated, and thousands of plant science articles are published every year. However, many of these articles may never be cited and the associated data may never be reused because authors failed to make them FAIR. Publishers have a role in requiring and enforcing data sharing; but implementation is inconsistent and rarely are explicit guidelines provided (Vasilevsky et al., 2017). To this end, we have prepared a checklist of some key elements for authors to consider when preparing their papers for publication, which would help ensure that the published data are FAIR. Each item is detailed below, and summarized in Figure 1.

### CHECKLIST FOR PREPARING RESEARCH DATA FOR PUBLICATION

#### Submit Your Data to a Stable, Long-Term Repository

Authors are encouraged to publish their research data associated with their research paper as complete as possible. If the data are too large for the tables, figures, and supplemental information, they should reside in a long-term, stable repository and be attached to a persistent identifier such as a Digital Object Identifier (DOI). DOIs should be assigned by the repository and included in the publication. Check the journal's author instructions for preferred/recommended repositories (e.g., http://www. cell.com/molecular-plant/authors). If there are no recommendations or you are unsure where your data should go, please ask for help by contacting your community database curators or a curator from an "allied" species resource. If there are no recommended repositories that can accept your specific data type, please consider generalist repositories such as Data Dryad (https://www.datadryad.org/) or Zenodo (https://zenodo.org/). Some universities maintain their own institutional repositories to host data generated by members of their academic community. Finally, Re3data.org (re3data.org) is a searchable registry of repositories from all over the world that can be used to identify the right data repository for your data. Having your data available in a public repository, where it can be accessed using a stable identifier, is crucial for making your data FAIR.

#### **Release Your Data with Clearly Defined Terms of Use**

Because copyright law applies to scientific data, permissions for each dataset should be explicit by providing an appropriate copyright license, which specifies who can reuse the dataset and under what conditions. In some cases, journals, funders, or institutions may impose specific restrictions on licenses. Repositories and journals often apply one or more types of Creative Commons licenses (https://creativecommons.org/licenses/) to their submissions. Different license types may be applied to metadata, data, and the actual published article. Authors should review the different types of licenses and what they do or do not allow. For example, the CC-0 license means that all copyright is waived, whereas the CC-BY license allows data to be reused and remixed in any way, as long as the changes are indicated and appropriate credit is given. Ambiguity about permissible reuse of data can slow the pace of research. If it is not made clear who can use the data and how it can be used, authors may unintentionally restrict reuse, as the default copyright is the most restrictive.

#### Format the Data for Easy Reuse

To make it easier for machines to read and process the data, tabular data should be kept in text format as tab- or commaseparated text, or in Excel. In general, it is not recommended to use PDF or other image-based formats for tabular data. A good rule of thumb is: if the table contains data that can be directly reused, keep it as text. Anything that a reader might want to copy (e.g., primer sequences, lists of gene identifiers from a gene family) should stay as text. While it is possible to convert PDF tables into text, it is time-consuming and requires additional curation to clean up the data. If using Excel, the table values

Published by the Molecular Plant Shanghai Editorial Office in association with Cell Press, an imprint of Elsevier Inc., on behalf of CSPB and IPPE, SIBS, CAS.

#### **Molecular Plant**

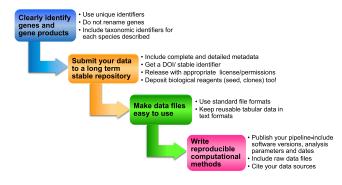


Figure 1. Checklist for Preparing Publications to Enhance Data Discoverability and Reuse.

A summary of the recommendations detailed in this editorial are presented for quick reference.

should be carefully checked to ensure data are not corrupted (Ziemann et al., 2016). Data files should conform to existing standards. For example, SNP data for plants can be submitted to the European Variation Archive (EVA; https://www.ebi.ac. uk/eva/) and formatted using the Variant Call Format (VCF; http://samtools.github.io/hts-specs/). Consult resources such as FAIRsharing (https://fairsharing.org), an emerging resource to aggregate and make searchable information about data standards. Publishing your data in standard formats makes it easier for computers to process and integrate.

#### Attach Complete and Detailed Metadata

Metadata are the data about your data, which enable anyone to understand and use your data. This can be as simple as a README file that describes the contents of a data file (e.g., column headers and descriptions) or metadata forms that use standard file formats and naming conventions. Examples of metadata standards include MIxS, minimum information about any sequence (http://gensc.org) used to provide data about any sequence such as germplasm sources, tissue types, environmental conditions, and experimental treatments, and MIAPPE, minimum information about a plant phenotyping experiment (http://www.miappe.org/) for describing plant phenotyping experiments. These metadata are often described using ontologies, which are hierarchically structured, controlled vocabularies that computers can understand, rather than free text. Most repositories will require at least a minimal amount of metadata in order to accept a submission. For example, at NCBI you will be prompted to provide information about a BioSample such as species, developmental stage, and environmental conditions. The more metadata you provide, the more accessible and reusable your data will be.

#### **Use Established Gene and Gene Product Names**

In general, authors should refrain from inventing new names for named genes and gene products. When a gene is published under different names, it becomes difficult to aggregate published data about that gene. For example the *Arabidopsis* gene AT1G01040 has been published with eight distinct gene names (https://www.arabidopsis.org/servlets/TairObject? name=AT1G01040&type=locus). If, when performing searches, researchers and text mining software are not aware of all possible names, information will be missed. Authors should also avoid creating duplicate gene names or symbols for genes in the same genome to avoid accidental mis-association of the data to the wrong gene. If you have identified a gene or gene product that you think has not yet been named, please always check with your community database (Table 1) to find if it has a name. For genes/gene products that have not yet been named, it is recommended to refer to the established nomenclature standards for the organism you are studying (e.g., https://www.arabidopsis.org/portals/nomenclature/symbol\_main.jsp and https://www.maizegdb.org/nomenclature).

## Use Unambiguous Identifiers for Genes and Gene Products

The best way to avoid any confusion over the identity of genes and gene products is to use a unique identifier that unambiguously identifies the sequence. For a sequenced gene or gene product, this could be a UniProt ID (protein), GenBank ID (DNA or RNA sequence), miRBASE (microRNA), RNACentral ID (non-coding RNA) or in the case of *Arabidopsis*, a locus (e.g., AT5G41410) or gene model ID (e.g., AT5G41410.1). The use of unique identifiers assures greater accuracy when using computational methods to associate a gene with its data. Because gene sequences and annotations can change with different genome assemblies, authors should ensure that gene identifiers correspond exactly to the gene/gene products in the published study. For example, in maize each gene model from each assembly version has its own unique identifier. Precise identification of the sequence is also essential to ensure experimental reproducibility.

#### Make Sure to Clearly Define the Study Organisms

To facilitate searching and discoverability of research results, it is important to include taxonomic information about the genes being described in an article. For example, if your paper describes data about genes from tomato or grape, please state that explicitly in the paper. When discussing genes from several different species in a single publication, it is recommended to use unique sequence identifiers or indicate in the text which organism each gene is from (e.g., sorghum FLC). As the number of genomes increases, it will become important to establish a standard plant gene nomenclature, possibly based on orthology. Until then, it has become a common practice to give orthologs the same symbolic name with a species prefix (e.g., AtFLC, ZmTFL).

#### Write Reproducible Computational Methods

While many researchers are experienced in writing experimental methods, critical details are often left out when describing computational analysis. Computational methods should name the specific software, versions, and all parameters. If a series of computational methods (pipelines) are used, this should be precisely described and cited, including all inputs and their sources. Some software platforms have integrations for publishing reproducible methods. For example, analysis routines performed using the Cyverse environment can be published with a unique URL that others can use to access and reuse your pipeline. All data sources should be cited, including the name and URL of the source database, the name of the dataset, the date it was accessed, and names, version numbers, and exact sources of the raw download files or a link to a DOI where the data can be found.

Database name	URL	Species covered
Alfalfa Breeders Toolbox	alfalfatoolbox.org	Alfalfa
CassavaBase	cassavabase.org	Cassava
Citrus Genome Database	citrusgenomedb.org	Lemons, oranges, and more
Cool Season Food Legume Database	coolseasonfoodlegume.org	Lentil, pea, fava bean, chickpea
CottonGen	cottongen.org	Cotton, many species
Genome Database for Rosaceae	rosaceae.org	Apple, strawberry, rose, plums, pears, and more
Genome Database for Vaccinium	vaccinium.org	Blueberry, cranberry, bilberry, and more
GrainGenes	wheat.pw.usda.gov/GG3	Wheat, barley, oats
Hardwood Genomics	hardwoodgenomics.org	Oaks, poplars, maples, chestnuts, and more
KnowPulse	knowpulse.usask.ca	Chickpea, common bean, lentil
Legume Information System	legumeinfo.org	Soybean, Medicago, cowpea, chickpea, and more
MaizeGDB	maizegdb.org	Maize
Medicago truncatula Genome Database	medicagogenome.org	Clover
MusaBase	musabase.org	Banana
Oryzabase, Rap-Db	shigen.nig.ac.jp/rice/oryzabase/	Rice
PeanutBase	peanutbase.org	Peanuts
Solanaceae Genomics Network	solgenomics.net	Tomato, potato, eggplant, petunia, and more
Soybase	soybase.org	Soybean
Sweet Potato Database	sweetpotatobase.org	Sweet potato
ТЗ	triticeaetoolbox.org/wheat	Wheat, barley, oats
TAIR	arabidopsis.org	Arabidopsis thaliana
TreeGenes	treegenesdb.org	1792 species of trees
YamBase	yambase.org	Yams

#### Table 1. Plant Gene Nomenclature Resources.

Plant community databases with resources and guidelines for naming genes and other biological entities. It is recommended to consult with appropriate database curator before naming any sequenced or genetically defined locus.

It is recommended to cite the database as well and include that in your references. Not only does this promote transparency and reproducibility, it also demonstrates the value of databases and repositories to funders.

#### **Deposit Your Germplasms**

For many model species, there are two types of germplasm repositories. The first type are those for genetic laboratory stocks (with mutations, insertions, etc.), which are usually communitybased. For example, the Arabidopsis Biological Resource Consortium (ARBC) accepts Arabidopsis seeds (and DNA stocks), and the Maize Genetics Cooperation accepts maize seeds. The second type are those that accept and maintain different accessions of "non-mutant" lines (land races, inbreds, naturally occurring variations, etc.). These are generally large world-wide repositories. For example, the USDA's National Plant Germplasm System (NPGS; www.ars-grin.gov) holds over 15 000 plant species and over 500 000 different accessions. If you publish a whole-genome sequence of any plant species, please deposit sibling seeds in an appropriate stock center and link the genome sequence to the given accession number. If there is a stock center that accepts and distributes the type of germplasms you are studying, you should deposit them into the stock center and include the stock IDs in your publication. Sharing the stocks ensures that others can reproduce your experiments and also means you spend less time fielding and fulfilling requests for materials.

#### **Concluding Remarks**

The above suggestions will help authors to increase the impact of their work by publishing research that is more accessible and reusable. If the data in published papers are more machine readable, they are more discoverable and easier to analyze. If the methods are clearly laid out and the data and reagents are made available, then it is easier for others to reproduce the research. While the emphasis here is on publication, practicing good data hygiene at all stages of experimentation will make it easier for you to manage the data and prepare your research for submission to journals and repositories. We suggest you plan your data management strategy at the start of every experiment, think about the metadata you need to capture and how you will do that, and adhere to your plan. When the time comes to write up your results, it will be much easier to make it FAIR.

#### FUNDING

TAIR is funded by individual, institutional, national, and corporate nonprofit subscriptions. MaizeGDB is funded by the USDA-ARS (grant no. 5030-21000-068-00D).

#### **Molecular Plant**

#### Editorial

#### **AUTHOR CONTRIBUTIONS**

L.R. and L.H. wrote the draft and prepared the table and figure. M.F., B.H., and S.L edited the text.

#### ACKNOWLEDGMENTS

No conflict of interest declared.

Received: July 25, 2018 Revised: July 25, 2018 Accepted: July 30, 2018 Published: August 1, 2018

> Leonore Reiser<sup>1,5,\*</sup> Lisa Harper<sup>2,5,\*</sup> Michael Freeling<sup>3</sup> Bin Han<sup>4</sup> Sheng Luan<sup>3</sup>

<sup>1</sup>The Arabidopsis Information Resource/Phoenix Bioinformatics, Fremont, CA 94538, USA

<sup>2</sup>Maize Genome Database/United States Department of Agriculture, Albany, CA 94710, USA

<sup>3</sup>Department of Plant Biology, University of California, Berkeley, CA 94720, USA <sup>4</sup>National Center for Gene Research, CAS Center for Excellence in Molecular Plant Sciences, Institute of Plant Physiology and Ecology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200233, China <sup>5</sup>These authors contributed equally to this article. \*Correspondence: Leonore Reiser (Ireiser@arabidopsis.org), Lisa Harper

(lisaharper@me.com) https://doi.org/10.1016/j.molp.2018.07.005

#### REFERENCES

Leonelli, S., Davey, R.P., Arnaud, E., Parry, G., and Bastow, R. (2017). Data management and best practice for plant science. Nat. Plants 3:17086.

Schiermeier, Q. (2018). Data management made simple. Nature 555:403–405.

Vasilevsky, N.A., Minnier, J., Haendel, M.A., and Champieux, R.E. (2017). Reproducible and reusable research: are journal data sharing policies meeting the mark? PeerJ 5:e3208.

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3:160018.

Ziemann, M., Eren, Y., and El-Osta, A. (2016). Gene name errors are widespread in the scientific literature. Genome Biol. 17:177.