



Making experimental data tables in the life sciences more FAIR: a pragmatic approach

Daniel Jacob, Romain David, Sophie Aubin, Yves Gibon

► To cite this version:

Daniel Jacob, Romain David, Sophie Aubin, Yves Gibon. Making experimental data tables in the life sciences more FAIR: a pragmatic approach. GigaScience, 2020, 9 (12), pp.giaa144. 10.1093/giga-science/giaa144 . hal-02883355v4

HAL Id: hal-02883355

<https://hal.inrae.fr/hal-02883355v4>

Submitted on 7 Jan 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.





L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

COMMENTARY

Making experimental data tables in the life sciences more FAIR: a pragmatic approach

Daniel Jacob ^{1,2,*}, Romain David ^{3,4}, Sophie Aubin ⁵ and Yves Gibon ^{1,2}

¹INRAE, Université de Bordeaux, UMR BFP, 71 av E Bourlaux, 33140 Villenave d'Ornon, France;

²PMB-Metabolome, INRAE, 2018. Bordeaux Metabolome Facility, MetaboHUB, 33140 Villenave d'Ornon, France;

³INRAE, Montpellier SupAgro, Université de Montpellier, UMR MISTEA, 2 place Pierre Viala, 34060 Montpellier

Cedex 2, France; ⁴European Research Infrastructure on Highly Pathogenic Agents (ERINHA-AISBL), 101 rue de

Tolbiac, 75013 Paris, France and ⁵INRAE, DipSO, 42 rue Georges Morel, 49070 Beaucouzé, France

*Correspondence address. Daniel Jacob, E-mail: daniel.jacob@inrae.fr  <http://orcid.org/0000-0002-6687-7169>

Abstract

Making data compliant with the FAIR Data principles (Findable, Accessible, Interoperable, Reusable) is still a challenge for many researchers, who are not sure which criteria should be met first and how. Illustrated with experimental data tables associated with a Design of Experiments, we propose an approach that can serve as a model for research data management that allows researchers to disseminate their data by satisfying the main FAIR criteria without insurmountable efforts. More importantly, this approach aims to facilitate the FAIR compliance process by providing researchers with tools to improve their data management practices.

Keywords: research data management; FAIR Data principles; FAIR assessment; experimental data tables

Background

The publication of research data according to the FAIR principles [1] has become a major challenge with the aim of integrating them into the overall research process (e.g., the European Commission explicitly mentions FAIR principles as a mandatory reference [2]). However, implementing these principles is not so easy and requires changes in data management practices. According to Jacobsen et al. [3], the FAIR principles can be seen as a consolidation of good data management practices to extend management to the notion of machine reuse of data. Thus, it seems appropriate to use virtuous principles as far upstream as possible from the data rather than trying to comply with FAIR principles downstream. This is the starting point of the approach that we propose in order to integrate these principles into the practices of researchers.

Set the Scene

Let us take a concrete example from the plant biology domain. A study on the metabolism of tomato fruits [4] involved growing several hundred tomato plants in greenhouses. This multifactorial experiment (stages of development and type of treatment applied to each group of plants) generated a dozen large experimental data tables. Part of the data was acquired in the greenhouse manually using spreadsheets, while another part of the data comes from biochemical and metabolomics analyses, carried out on the thousand samples taken and returned in the form of data tables weeks or even months later.

The use of spreadsheets is therefore central here because it is a tool that researchers master well. However, manual handling is required to link together experimental data tables from several analytical techniques, according to samples. Such repeated data handling, a potential source of errors, can compromise the

Received: 29 June 2020; Revised: 3 November 2020; Accepted: 17 November 2020

© The Author(s) 2020. Published by Oxford University Press GigaScience. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

consistency of data, which must be managed throughout the study by ensuring that each analysis is well linked to its sample. We needed to review our data management practices. The question remained as to how to motivate and convince researchers to change their practice a little.

To Promote Good Practices, Provide Services

Efforts have been undertaken for several years to propose format standards that would make it possible to disseminate data according to FAIR principles and in particular experimental data tables (e.g., ISA-TAB [5]). In our approach, the emphasis has been mainly on the integration of FAIR principles from the beginning of the data's life, i.e., as soon as they are acquired. Thus, we have focused on the structural metadata related to the experimental data in the spreadsheets, i.e., how they are organized so that we can more easily exploit them. The objective of our approach called ODAM (Open Data for Access and Mining) is to make this upstream capture an advantage to facilitate data analysis and therefore an incentive to perform this metadata capture. In relying on the tool that researchers know best, i.e., spreadsheets, it was nevertheless necessary to remedy its drawbacks and in particular the lack of constraints in the structuring of data. From this perspective, we propose a data structuring similar to data dictionaries that is easy to implement by the researchers themselves (Fig. 1, Additional File 1). Structural metadata (e.g., links between data tables) are described, together with unambiguous definitions of all internal elements (e.g., column definitions along with their semantic definition), through links to accessible definitions, such as community-approved ontologies where possible, as recommended by Jacobsen et al [3]. But for good practices to be adopted, researchers must take advantage of them. Thus, we propose tools that greatly facilitate the combination and merging of datasets according to a common attribute (identifiers), allowing the analysis of several types of variables according to different parameters without any tedious manipulation of the data, offering researchers an appreciable time savings by avoiding repetitive and tedious tasks. Besides, depending on the needs and skills of researchers, data can be used in a wide variety of ways (Fig. 2).

The advantage of this approach is manifold. It allows the data to be structured in such a way that the researcher (i) can proceed step by step as the data become available and (ii) can easily exploit them with tools immediately afterwards. In doing so, FAIRification of data is carried out in order to handle data more efficiently and not just to publish it. Thus, it is the FAIRification of data that is integrated by design into the data processing workflow. So, in our approach data FAIRification is closely related to data management, avoiding a retroactive process that would require more time, costs, and computer skills [6, 7]. In addition, from the structured metadata it becomes possible to convert them directly into a standard format. In our case, we chose the "Frictionless datapackage" (<https://frictionlessdata.io/>), a community-based, open interoperability standard (Figs 2 and 3). Slightly adapted to our needs, e.g., by specifying the category of each attribute, the data can thus be disseminated according to an open schema that greatly facilitates the reuse of data by machines. The choice of the data repository to disseminate its data formatted in this way is quite open because a clear separation is established between structural metadata on the one hand, and descriptive metadata depending on the type of repository, on the other hand. However, on the basis of metadata files this does not prevent the development of converters

to other formats such as the complex data model (e.g., ISA-TAB [5]) in order to include the data in existing standards-compliant data infrastructures (e.g., SEEK data management platform [8]).

Publication of Data

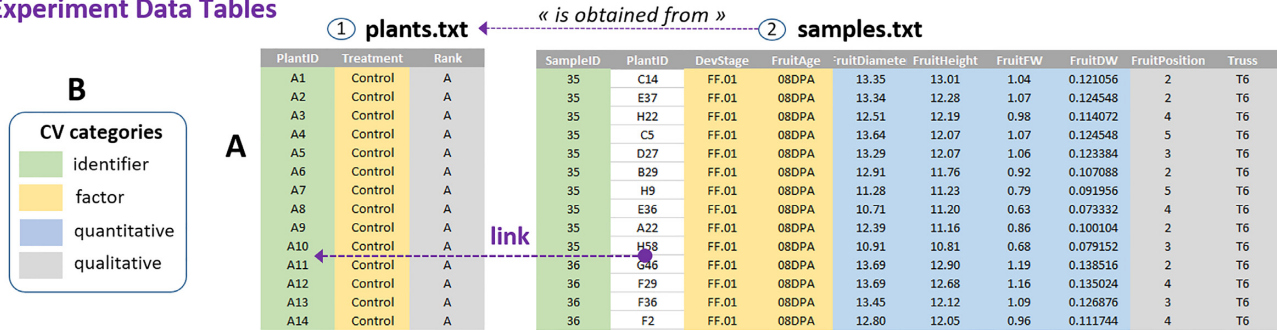
When it comes to publishing their results, researchers usually provide the minimum required to support their claims [10]. This generally results in a loss of data (quantitative aspect) and information (qualitative aspect) compared with the totality of the data acquired during the study. We believe that our approach should facilitate the dissemination of the complete dataset because the work has been done upstream, and that when needed the data can be deposited in an appropriate repository quickly enough without having to perform data archaeology, while at the same time meeting the essential criteria of the FAIR principles is guaranteed (Additional Files 2–4). In addition, the expected benefits are numerous, including exploiting the full potential of the datasets, improving the reproducibility and reliability of the data, but also increasing visibility and citation due to the reuse of the data by both humans and machines.

Furthermore, a lever of motivation would be the recognition of this effort to publish data according to the FAIR principles. Unfortunately, researchers who devote time and expertise to activities like data curation are not currently rewarded by traditional career progression metrics. We believe that this should change in the future, and crediting and rewarding mechanisms are the subject of the Research Data Alliance SHaring Rewards and Credit Interest Group [9].

Summary of the Proposed Approach and Beyond

1. The goal is to provide researchers with services that are truly useful, time-saving, and efficient. Care must also be taken not to deprive them of their know-how or trap them in turnkey solutions that prevent any opportunity of testing several hypotheses or scenarios. Rather, we need to open the data to a whole ecosystem of software possibilities.
2. Because researchers have the best control and understanding of their data, they are in the best position to annotate it. It is therefore advisable to help them as much as possible in this process, by offering them protocols and methods adapted to their IT skills, an area that is not their core business. In particular, vocabulary dictionaries corresponding to their domain and frequently used should be provided to standardize annotations as much as possible.
3. Behind a dataset, there is often an involved team with a wide range of skill levels. We must take into account their way of working, their work habits; so instead of wanting to change their habits completely, we must rather adapt them in a way that is beneficial to them and make them actors in the process of making data FAIR compliant. Probably the best FAIR training is the one based on their data. But it is necessary to capitalize on good practices in written protocols, and referenced into data management plans.
4. Researchers must be involved in the process of making data FAIR compliant by providing them with tools for assessing their practices so that they can progressively improve them in stages, having properly integrated each of the criteria implemented. Especially, these assessment tools should highlight all the steps where small actions can make the data significantly more FAIR. They will thus be more inclined to

Experiment Data Tables



Structural metadata

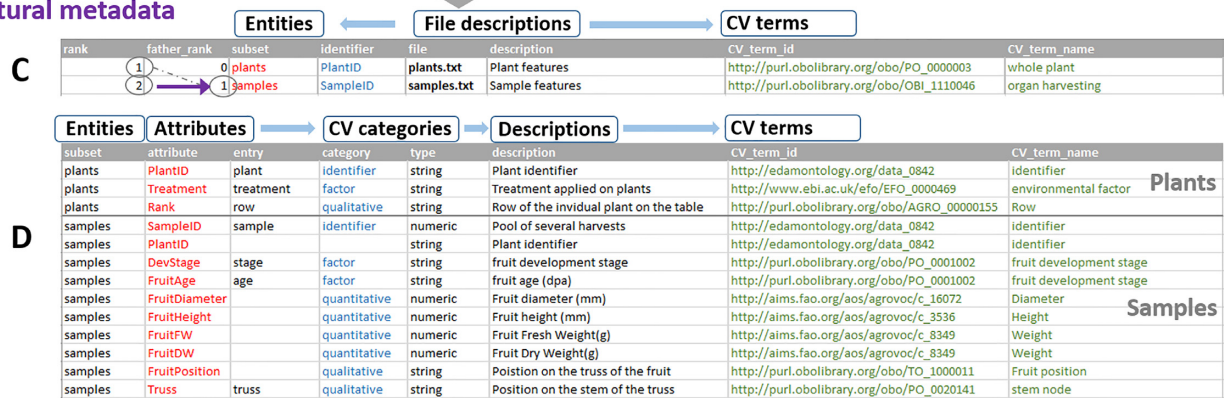


Figure 1: ODOM (Open Data for Access and Mining) is an experiment data table management system based on good data management practices concerning data structuring and the description of structural metadata. Indeed, the strong point of the approach is to define metadata in depth, i.e., at the level of the data themselves (i.e., metadata at column level such as factors, variables, and so forth) and not only as a “hat” on the dataset. Thus, having structural metadata allows datasets to achieve a higher level of interoperability and greatly facilitates functional interconnection and analysis in a broader context. (A) To simplify, we consider here the first 2 tables of data from the experiment, namely, the individuals (plants.txt) followed by the samples (samples.txt). The data must be well organized, i.e., each variable forms a column, each observation forms a row, and each table is relative to an entity, i.e., the same type of observational unit (e.g., plants, samples), and a file must contain only 1 data table. Because all experimental data tables were generated in an experiment associated with a particular experimental design, the data tables were acquired sequentially as the experiment progressed. A link must exist between each of them, generally defined by identifiers. In our example, each sample is linked to the plant from which it comes. (B) Furthermore, whatever the type of experiment, it requires an experimental design involving individuals, samples, or whatever, as the main objects of study and producing several tables of experimental data. It also involves the observation of dependent variables resulting from the effects of certain controlled independent variables (factors). In addition, the objects of study usually each have an “identifier,” and the variables can be quantitative or qualitative. Thus, each of the columns within a table (attributes) can be associated with 1 of the 4 categories: identifier, factor, quantitative, qualitative. Associating a category with each column greatly facilitates subsequent statistical analyses by machines. All structural metadata can be grouped in 2 specific files. (C) The first metadata file associates with each data table (subset) a key concept corresponding to the main entity of the data table. It also defines for each table the link with the table from which it comes (purple arrow). These links can be interpreted as “is obtained from.” (D) The second metadata file annotates each attribute (concept/variable) with minimal but relevant metadata, such as its category defined above, its description with its unit, the data type. In each of these 2 files (entities and attributes), it is possible to annotate each of the terms with unambiguous definitions (CV terms) through links to accessible (standard) definitions based on ontologies. The choice of ontologies is very domain-specific but nevertheless should preferably be based on those that follow the FAIR principles [3]. In the case of the FRIM experiment, we mainly used AgroPortal (<http://agroportal.lirmm.fr/>) and especially its “annotator” module, made efficient thanks to the alignment of ontologies. Because these ontological terms are not essential for statistical analysis, they can be omitted up to the publication stage. It should be noted that tools for adding ontology terms to Excel spreadsheets are still being developed for the ODOM software suite to facilitate this tedious task (<https://inrae.github.io/ODAM/todo>). Some tools such as RightField (<https://rightfield.org.uk/>), ISA-Tools [5], or Swate (<https://github.com/nfdi4plants/Swate>) offer interesting approaches and are certain to be good sources of inspiration. Given the knowledge that ontological terms are essential mainly for data dissemination, a connection with the ISA-TAB format for instance would make it possible to benefit from the tools already available for this type of task. In any case, established mainly by and for the scientists who produced the data, these structural metadata will later allow non-expert users to explore and visualize the data, thus offering a better guarantee of correct (re)use by those who did not produce them. See Data Preparation Protocol for ODOM Compliance for more details (Additional File 1).

integrate them into their practice with full knowledge of the facts.

- Concerning the data provenance: not only the authors, but above all the context, the methods of data acquisition and processing are crucial information for a good reuse. Unfortunately, this aspect is somewhat neglected if not absent. An effort still needs to be made in this direction in order to sensitize and motivate data producers. Finally, it should be mentioned that the license of the data must be appropriate for the reuse of the data.

Data Availability

All information, documents, data, and software concerning ODOM are accessible from Github (<https://inrae.github.io/ODAM/about/>).

Additional Files

Additional File 1. Data Preparation Protocol for ODOM Compliance. The purpose of this protocol is to describe all the steps involved in collecting, preparing, and annotating the data from an

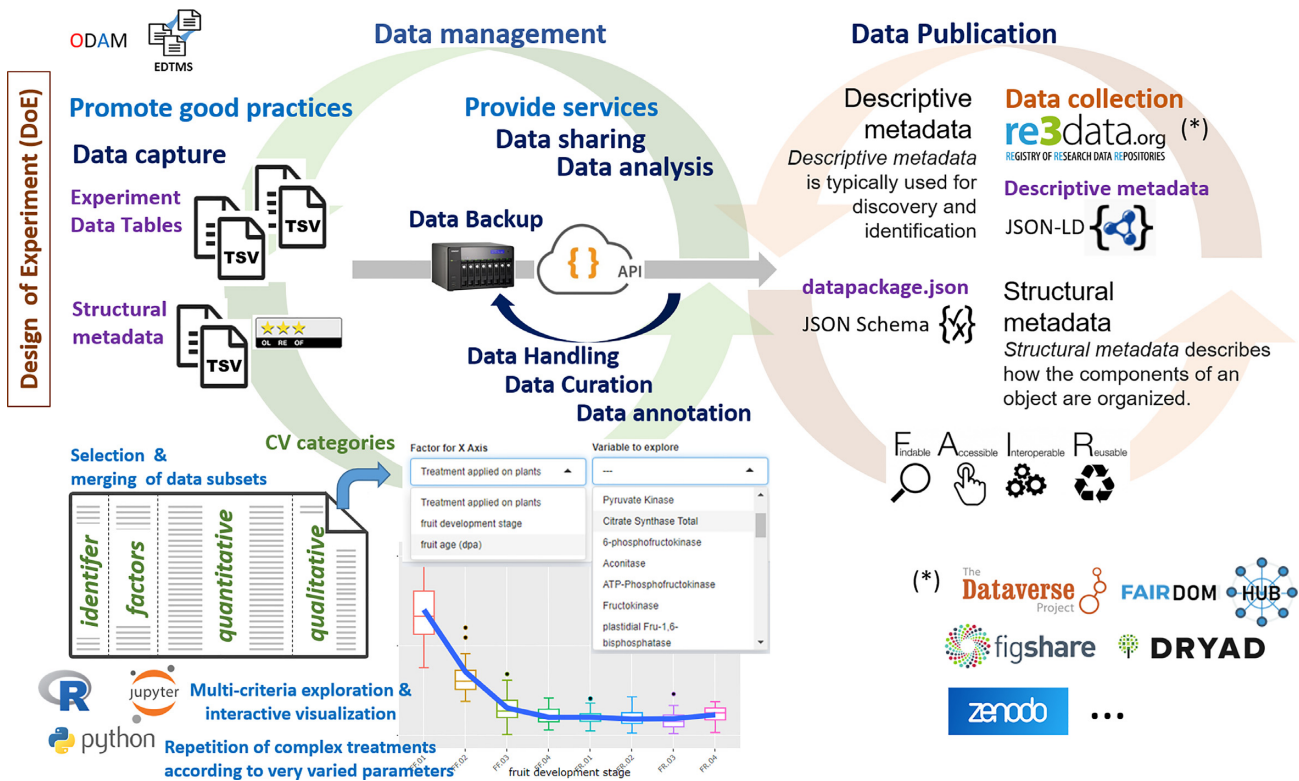


Figure 2: ODA software suite: Light blue indicates the engine of the approach; purple, the data and metadata provided by the user; and dark blue, the activities related to the life cycle of the data. The whole process is implemented primarily to make better use of data before their dissemination. The ODA software embeds an API layer that allows interoperability between the different tables and the applications that will be able to use them. With the help of this layer, it opens up a whole ecosystem of potential applications, depending on the user's needs but also on the user's skills in the proposed tools. From the set of data files (which are non-combined tables, each corresponding to a particular observational unit that we name an entity), the user can (i) visualize the data associated with their metadata according to several criteria and in a completely interactive way with the help of the data explorer; (ii) export in tabular form subsets selected according to his criteria with combined, merged data; and (iii) build and test his models more easily using a scripting language such as R, which allows different scenarios to be repeated according to a variety of parameters. All this is made possible thanks to the category as controlled vocabulary associated with each column, which facilitates statistical analysis by both humans and machines. Moreover, the first available data can be exploited as soon as the corresponding metadata have been captured without waiting until all the data are available. The benefit of this approach is that the "life of the data" is integrated into the scientific research process, according to good data management practices that meet the essential FAIR criteria. Then, distributed data are enriched by associating a structural metadata file called datapackage (<https://frictionlessdata.io/>), a simple container that serves as metadata aggregator based on JSON schema specifications, an open, community-based interoperability standard. This compact and hierarchically structured format proved to be suitable for integrating all of our structural metadata, thus placing the dataset in its experimental context, a key factor in making the data FAIR. Data generation according to this open schema is included in the proposed tools and does not require additional effort for the researcher. The definition of an explicit schema for structural metadata thus enables machines to better interpret the data for reuse. Indeed, exporting these metadata in datapackage format offers a great flexibility of using data via scripting languages such as R and Python on the basis of existing packages. Besides, this type of format allows a great variety in the choice of data repository because a distinct separation is established between structural metadata described in the datapackage format on the one hand, and descriptive metadata depending on the type of repository on the other hand. Preferably the chosen data repository should offer the ability to query and retrieve data using an API that conforms to the OpenAPI specification (<http://spec.openapis.org/oas/v3.0.3>) and meet the essential criteria of the FAIR principles. For example, the following data repositories registered in re3data.org (<http://re3data.org/>) can be cited without being exhaustive: Dataverse (<https://dataverse.org/>), Dryad (<https://datadryad.org/>), FAIRDOMHub (<https://fairdomhub.org/>), FigShare (<https://figshare.com/>), Zenodo (<https://zenodo.org/>). Finally, making data FAIR compliant can be considered from 2 points of view: (i) it is linked to the data life cycle by the annotations and curations made on the data themselves, and to the quantity and quality of the information associated with the data (e.g., protocols, publications, keywords); (ii) it can also be considered from the point of view of its data management practices, which must improve over time, which is precisely what the FAIR assessment grids attempt to measure, and more particularly the reproducibility and reusability of the data. See ODA Deployment and User's Guide for more details (<https://inrae.github.io/ODAM/>). EDTMS: experiment data table management system; TSV: tab-separated values.

experiment associated with an experimental design (DoE) that will then allow the user to benefit from the services offered by ODA.

Additional File 2. FAIR evaluation of the FRIM1 dataset according to the 5-Star Data Rating Tool grid. It aims to perform an evaluation based on the FAIR principles as defined by Wilkinson et al. [1]. The main result is an overall rating, indicating the overall fairness of the dataset.

Additional File 3. FAIR assessment of the FRIM1 dataset according to the FDMM (FAIR Data Maturity Model) grid. This document

describes a maturity model for the FAIR assessment with indicators, priorities, and assessment methods, which are useful for standardizing assessment approaches to allow comparison of their results.

Additional File 4. FAIR assessment of the FRIM1 dataset according to the SHARC (Sharing Rewards and Credit) grid. This document allows the fairness of projects and associated human processes to be assessed, either by external evaluators or by the researchers themselves.

project funded by the French National Research Agency (ANR-09-SYSB-003) and the MetaboHUB project funded by the French National Research Agency (ANR-11-INBS-0010).

Authors' Contributions

Conceptualization: D.J.; data curation: D.J.; funding acquisition: Y.G.; methodology: D.J., R.D.; software: D.J.; writing—original draft: D.J., R.D.; writing—review and editing: All authors. All authors read and approved the final manuscript.

Acknowledgments

We thank Catherine Deborde (PMB-Metabolome, INRAE, MetaboHUB) for advice on the manuscript and for constructive reviews.

References

1. Wilkinson MD, Dumontier M, Aalbersberg IJJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 2016;3:160018.
2. European Commission Directorate General for Research and Innovation. Turning FAIR into reality, Final Report and Action Plan from the European Commission Expert Group on FAIR Data. 2018. https://ec.europa.eu/info/publications/turning-fair-reality_en. Accessed 4 June 2020.
3. Jacobsen A, de Miranda Azevedo R, Juty N, et al. FAIR Principles: interpretations and implementation considerations. *Data Intell* 2020;2, doi:10.1162/dint.r'00024.
4. Bénard C, Biais B, Ballias P, et al. FRIM - Fruit Integrative Modelling. Portail Data INRAE, V3. 2018. <https://doi.org/10.15454/95JUTK>.
5. Sansone S, Rocca-Serra P, Field D, et al. Toward interoperable bioscience data. *Nat Genet* 2012;44:121–6.
6. Rocca-Serra P, Sansone SA. Experiment design driven FAIRification of omics data matrices, an exemplar. *Sci Data* 2019;6:271.
7. European Commission, Directorate-General for Research and Innovation. Cost-Benefit analysis for FAIR research data - Cost of not having FAIR research data. 2018. <https://op.europa.eu/en/publication-detail/-/publication/d375368c-1a0a-11e9-8d04-01aa75ed71a1>. Accessed 4 June 2020.
8. Wolstencroft K, Owen S, Krebs O, et al. SEEK: A systems biology data and model management platform. *BMC Syst Biol* 2015;9:33.
9. David R, Mabile L, Specht A, et al. FAIRness literacy: the Achilles' heel of applying FAIR principles, *Data Science Journal*; 19 1:32. <https://doi.org/10.5334/dsj-2020-032>.
10. Leonelli S, Smirnoff N, Moor J. Making open data work for plant scientists. *J Exp Bot* 2013;64: 4109–17.