


SCIENTIFIC DATA

OPEN
COMMENT

Experiment design driven FAIRification of omics data matrices, an exemplar

Philippe Rocca-Serra  & Susanna-Assunta Sansone 

We outline a principled approach to data FAIRification rooted in the notions of experimental design, and whose main intent is to clarify the semantics of data matrices. Using two related metabolomics datasets associated to journal articles, we perform retrospective data and metadata curation and re-annotation, using community, open, interoperability standards. The results are semantically-anchored data matrices, deposited in public archives, which are readable by software agents for data-level queries, and which can support the reproducibility and reuse of the data underpinning the publications.

The scientific, economic and societal impact of research depends upon access to and reuse of the methods and data generated in every publication, and the utility of genomic research depends increasingly upon access to appropriately curated phenotypic data¹. For scalable, effective and trustworthy data-driven science, we need to ensure that data are Findable, Accessible, Interoperable and Reusable (FAIR) by humans as well as by machines. Since their publication in 2016², the FAIR Principles of data management and stewardship have become pervasive in discussions, policies and implementations in and around technological and social infrastructure for research data^{3,4}.

The principles put specific emphasis on enhancing the ability of machines to find, access and process data, in addition to supporting their interoperability and reuse by individuals.

Responding to the invitation to make a rose dataset FAIRer⁵, we present a principled approach to data FAIRification, which focuses on the clarity of the statistical results. We showcase its application using the work by Raymond *et al.*⁶ on a targeted metabolite profiling study of strain-related chemical signatures of the rose fragrance. Our starting point was the human-understandable data provided by the authors as a supplementary table. Using community, open, interoperability standards, available from FAIRsharing⁷ (<https://fairsharing.org>), we performed the retrospective curation and re-annotation of the data matrices, disambiguating them using the experimental design information. To assess inter-experiments agreement, we applied the same procedure to a second data source⁸, which is an early work of the same group on the same varieties of rose and plant parts. The results are served in an open syntax and fully documented, as well as executable data science project, with jupyter notebooks. In the following sections, we detail the process needed to transform typical supplementary tables into machine readable information to enable data-level queries, and support the reproducibility and reuse of the data.

The FAIRification Process

In the first article⁶, the “Supplementary Data Table 3” is a spreadsheet that collates the mean concentrations of 61 molecular compounds measured in different parts of the rose flower, in six distinct genotypes. Whilst this table is understandable to a human audience, it is not particularly suited to consumption by software agents, falling short on several of the FAIR Principles. In Fig. 1, we summarize the steps undertaken to make this dataset FAIRer. To address findability and accessibility of our work, we uploaded all relevant files to the Zenodo repository and assigned these artefacts an open license (CC-BY 4.0). Using the Digital Object Identifier (DOI) minted by Zenodo, the initial spreadsheet table⁹ and the associated FAIRified outputs^{10,11} are more discoverable and also formally citable. To address the interoperability and reusability of the data, we applied several steps. We exposed the semantics hidden behind the column headings in the spreadsheet by identifying the main types of tabulated entities and their relationships; we marked-up all entities with persistent resolvable identifiers to enhance dataset connectivity. We then regularized the matrix using a well-established syntax; and lastly, we performed a conversion to Linked Data.

Oxford e-Research Centre, Department of Engineering Science, University of Oxford, 7 Keble Road, Oxford, OX1 3QG, United Kingdom. email: philippe.rocca-serra@oerc.ox.ac.uk; susanna-assunta.sansone@oerc.ox.ac.uk

FAIR Data for Humans and Machines

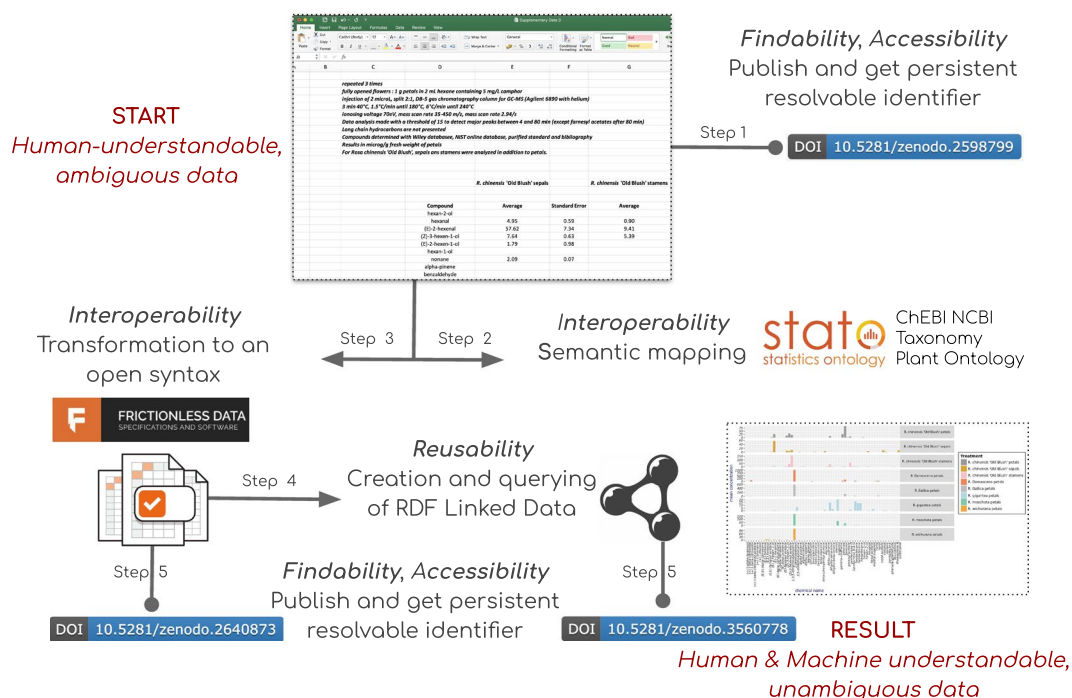


Fig. 1 Summary and overview of the steps undertaken to make the targeted metabolite profiles of the Rose scent datasets FAIRer, and how each one address one of more of the FAIR elements. Step 1: made the initial spreadsheet table discoverable and citable, assigning an open license (CC-BY 4.0). Step 2: regularized the three dimensions of the matrix (which represent the metabolites, the treatments, and the quantitation type), by unpacking the information held in the column header, replacing free text with ontology terms, disambiguating the experimental design and clarifying the measurement performed. Step 3: used the Frictionless Tabular Data Package to describe the table headers in JSON format, documenting the transformation in a jupyter notebook. Step 4: performed a conversion to Linked Data, plotting the metabolite measurements using visualization libraries. Step 5: made the FAIRified outputs discoverable and citable, assigning an open license (CC-BY 4.0).

Semantic mapping. Conceptually, matrices of results are data cubes, or hypercubes, where the information is organized according to three (or more) dimensions, respectively. To be FAIR, the entities these dimensions hold must be unambiguously reported. In our case, the matrix defines three dimensions representing: i) the metabolites (or, more generally, molecular entities), ii) the ‘treatments (or experimental conditions and corresponding biomaterials and bioassays), and iii) the quantitation type (or generally, measurements). In the original spreadsheet table, the metabolites were reported using free text, which, in most instances, matched their common chemical name. Such practice is known to cause issues due to its imprecision. To address this point, we accessed CHEBI¹² programmatically through its LibChebi library¹³ (<https://github.com/libChEBI/libChEBIpy>) and retrieved relevant CHEBI identifiers and InChI strings¹⁴, which are the community-developed, non-proprietary textual identifiers for chemical entities. Metabolites names were thus augmented with these unambiguous codes.

As a next step, we ‘unpacked’ the information held in the column header and semantically anchored it. Starting with the biological materials, originally denoted with composed terms such as “*R. chinensis* ‘Old Blush’ sepals”, we disambiguated the taxonomic name of the cultivar from the anatomical part using terms and identifiers from the NCBI Taxonomy¹⁵ and Plant Ontology¹⁶ respectively. The use of established and community-driven terminologies, which are used by a number of public resources, offer higher potential for data discovery and interoperability.

This initial semantic mapping, however, is just the first step towards FAIRer data. The values reported in these columns headers are in fact the *factor levels* of 2 *independent variables*, which need to be properly marked up and made explicit. To better explain this point, we ask the reader to consider the overall design of this experiment and the hypothesis being tested. These biological materials have been selected on purpose, “by design”, to allow a comparison between parts of the plant, and across cultivars in the same tissue type (sepals in this instance). Therefore, basing data reporting on study design concepts provides a *principled way* for organizing the description of such data matrices. Emerging terminologies, such as the STATistics Ontology (STATO; <https://doi.org/10.25504/FAIRsharing.na5xp>), are essential to unambiguously express and semantically type these notions. In our case, a *factorial design* was recognized, with two *independent variables* identified, namely the rose variety and the organism part, which are both categorical variables with six and three discrete *factor levels*, respectively. In such a context, 18 theoretical *factor combinations* are possible, as determined by the result of the cross product (cartesian product) for the two sets of variables levels, with each combination identifying a possible *statistical treatment*. Since only eight out of eighteen are reported, we conclude this is a *fractional factorial design*.

Making such notions explicit clarifies and disambiguates the intent of the experimentalists and delivers clearer and more reusable datasets. However, not all metadata models are capable of representing such information with sufficient granularity. For example, a number of models implemented by major public genomics databases have no dedicated objects to represent *study factors* and their values; variables are implicitly declared as sample attributes, and therefore the database cannot explicitly enable queries on treatment groups and their sizes.

The last dimension of the data cube corresponds to the *quantitation types*: two measurements were identified for each of the experimental conditions: average and standard error. To anchor those in a semantic framework, we have also used STATO to replace the string *average* with the class *sample mean*, which corresponds more specifically to the notion of *arithmetic mean*, and the field header *standard error* with the class *standard error of the mean*. The *sample mean* measurement, as formally defined by the STATO class, should be reported with the size of the sample over which the calculation is performed. Although incomplete, some of this information was provided by the authors in the “Reporting Summary” file, part of the Supplementary Information available from Raymond *et al.*⁶. It seems that for each treatment, a single biological material was prepared and assayed three times on the same analytical platform. Therefore, the computed sample mean can only be used to estimate the variability of the measurement technique, not the biological variability. Once again, reporting such critical information accurately is essential to data analysts and statisticians for them to confidently apply methods and software agents to process the data, for instance, to run a workflow with parameters set for specific classic univariate analysis, such as 2-way ANOVA.

Open syntax. Having clarified the semantics, the next step was aimed at ensuring the long term preservation of data matrices. We used the Frictionless Tabular Data Package, a simple container format used to describe and package a collection of data (<https://frictionlessdata.io/data-packages>). The package provides a description of table headers using a JavaScript Object Notation (JSON) format, a popular open-standard representation, used to validate the tabular data themselves, provided alongside, as comma or tab separated values. The transformation is fully documented in our jupyter notebook (<https://github.com/proccaserra/rose2018ng-notebook/blob/master/notebooks/0-rose-metabolites-Python-data-handling.ipynb>).

Linked data. The last step of our FAIRification process is the creation of Linked Data, a method of publishing structured data so that it can be interlinked with other resources. The Resource Description Framework (RDF) is one of the key ingredients of Linked Data: it provides a generic graph-based data model for describing data that can be queried using the SPARQL language. The RDF representation, which relies on terms from OBO Foundry ontologies¹⁷, enables queries such as “Retrieve study predictor variables and their levels” and “What is the sample size used to compute the mean?”, therefore supporting study results review and assessment. As shown in the jupyter notebooks available as part of the code released for this work, the metabolite measurements themselves can be plotted using popular visualization libraries (Python plotnine or R ggplot2) from either a SPARQL query over the RDF representation or from the data package directly.

Integration and preservation. To further demonstrate the value of such study design driven data representation, we applied a similar FAIRification process to the supplementary material from an earlier work by the same group⁸. This also helped to assess inter-experiment agreement, as both studies used the same varieties of rose and plant parts. However, we had to modify our annotation pipeline to extract the metabolite profiles, not just from spreadsheets but also from PDF tables, adding an extra step to our process. Such additional work is quite common when FAIRifying data retrospectively. The results of this comparison, also released via Zenodo¹⁸, can be visually explored using the aforementioned graphic grammar compatible libraries. A Venn diagram and an Upset plot¹⁹ provide a visual overview of the metabolites shared between the two studies, and are available along with the executable code used to generate them. Lastly, we produced a study description file, in ISA-Tab format²⁰, which references the Tabular Data Packages representing the results held in data matrices. The ISA file is suitable for deposition to MetaboLights²¹, a public repository for metabolomics data recommended by several journals (<https://doi.org/10.25504/FAIRsharing.kkdpxe>).

The code and data associated with this project are archived in Zenodo, as detailed below:

- Rose scent FAIRification project code release²².
- Associated Material to “The Rosa genome provides new insights into the domestication of modern roses” publication⁹.
- GC-MS data from the ‘Rose Genome’ available as Frictionless Tabular Data Package¹⁰.
- RDF Linked Data representation of GC-MS data from the ‘Rose Genome’ article¹¹.
- Comparison of GC-MS datasets available as Frictionless Tabular Data Package¹⁸.

The future is FAIR data at the source. FAIRifying data retrospectively nevertheless remains limited and challenging. Data readiness needs to begin at the source. Our exemplar approach to make these rose metabolomics datasets FAIRer is very generic, applicable to most ‘omics’ datasets and should encourage experimentalists and data scientists to capture the intent of the experimental design prospectively, by regularizing and annotating the resulting matrices in a formal way. To enable data FAIRness, we will continue contributing to a number of international efforts that work to develop guidelines, tools and services (e.g. FAIREvaluator²³) for researchers and data stewards in the life sciences (such as the European FAIRplus, <https://fairplus-project.eu>, the USA National Institute of Health, <https://commonfund.nih.gov/dataecosystem>), and across-disciplines (including GO-FAIR, <https://www.go-fair.org>, and the Research Data Alliance, <https://www.rd-alliance.org>). We need new technological and social infrastructure, to transform the concept of data readiness into a powerful toolkit at the researchers’ fingertips, to realize FAIR data.

Received: 29 August 2019; Accepted: 31 October 2019;

Published online: 12 December 2019

References

1. Aronson, S. J. & Rehm, H. L. Building the foundation for genomics in precision medicine. *Nature* **526**, 336–342 (2015).
2. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).
3. Wise, J. *et al.* Implementation and relevance of FAIR data principles in biopharmaceutical R&D. *Drug Discov. Today* **24**, 933–938 (2019).
4. Directorate-General for Research and Innovation (European Commission). Turning FAIR into reality - Publications Office of the EU. *Turning FAIR into reality*, <https://doi.org/10.2777/54599> (2018).
5. A rose on the garden fair. *Nat. Genet.* **50**, 769 (2018).
6. Raymond, O. *et al.* The Rosa genome provides new insights into the domestication of modern roses. *Nat. Genet.* **50**, 772–777 (2018).
7. Sansone, S.-A. *et al.* FAIRsharing as a community approach to standards, repositories and policies. *Nat. Biotechnol.* **37**, 358–367 (2019).
8. Magnard, J.-L. *et al.* PLANT VOLATILES. Biosynthesis of monoterpene scent compounds in roses. *Science* **349**, 81–83 (2015).
9. Bendahmane, M. & Raymond, O. Supplementary Material to “The Rosa genome provides new insights into the domestication of modern roses” publication. *Zenodo*, <https://doi.org/10.5281/zenodo.2598799> (2018).
10. Rocca-Serra, P. & Sansone, S. A. Frictionless Tabular data package for GC-MS data from Rose Genome article published in Nature genetics, June, 2018. *Zenodo*, <https://doi.org/10.5281/zenodo.2640873> (2019).
11. Rocca-Serra, P. & Sansone, S. A. RDF Linked Data representation of GC-MS data from the “Rose Genome” article published in Nature genetics, June, 2018. *Zenodo*, <https://doi.org/10.5281/zenodo.3560778> (2019).
12. Hastings, J. *et al.* ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Res* **44**, D1214–9 (2016).
13. Swainston, N. *et al.* libChEBI: an API for accessing the ChEBI database. *J. Cheminform.* **8**, 11 (2016).
14. Heller, S. R., McNaught, A., Pletnev, I., Stein, S. & Tchekhovskoi, D. Inchi, the IUPAC international chemical identifier. *J. Cheminform.* **7**, 23 (2015).
15. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **44**, D7–19 (2016).
16. Cooper, L. & Jaiswal, P. The plant ontology: A tool for plant genomics. *Methods Mol. Biol* **1374**, 89–114 (2016).
17. Smith, B. *et al.* The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.* **25**, 1251–1255 (2007).
18. Rocca-Serra, P. & Sansone, S. A. Frictionless Tabular Data Package for GC-MS Rose scent profile data for Data published in Nature genetics, June, 2018 & Science, July 2015. *Zenodo*, <https://doi.org/10.5281/zenodo.2640919> (2019).
19. Conway, J. R., Lex, A. & Gehlenborg, N. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics* **33**, 2938–2940 (2017).
20. Rocca-Serra, P. *et al.* ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level. *Bioinformatics* **26**, 2354–2356 (2010).
21. Kale, N. S. *et al.* MetaboLights: An Open-Access Database Repository for Metabolomics Data. *Curr. Protoc. Bioinformatics* **53**, 14.13.1–18 (2016).
22. Rocca-Serra, P. proccaserra/rose2018ng-notebook: FAIRer rose, clarifying the semantics of data matrices. *Zenodo*, <https://doi.org/10.5281/zenodo.3560911> (2019).
23. Wilkinson, M. D. *et al.* Evaluating FAIR maturity through a scalable, automated, community-governed framework. *Sci. Data* **6**, 174 (2019).

Acknowledgements

P.R.S. and S.A.S. are funded by several grants, and the ones most relevant to this work are: H2020-EU.1.4.1.3 PhenoMeNal 654241, Wellcome Trust ISA-InterMine 208381/A/17/Z, IMI FAIRplus 802750, and H2020-EU.1.4.1.1. EOSC-Life 824087. S.A.-S. is also funded by the Oxford e-Research Centre, Department of Engineering Science of the University of Oxford.

Competing interests

S.A.S. is Honorary Academic Editor of Scientific Data and P.R.S. is a member of the Scientific Data Senior Editorial Board.

Additional information

Correspondence and requests for materials should be addressed to P.R.-S. or S.-A.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2019