

Plant Science Data Integration, from Building Community Standards to Defining a Consistent Data Lifecycle



Cyril Pommier, Frederik Coppens, Hanna Ćwiek-Kupczyńska, Daniel Faria, Sebastian Beier, Célia Miguel, Célia Michotey, Flora D'Anna, Stuart Owen, and Kristina Gruden

Abstract FAIR (Findable, Accessible, Interoperable, Reusable) data principles for plant research build upon experience from other life science domains such as genomics. But plant specificities, e.g. plant-environment interactions or phenotypes, require tailored solutions. Major global players have joined forces to answer that challenge with the Minimal Information About a Plant Phenotyping Experiment

C. Pommier (✉) · C. Michotey
Université Paris-Saclay, INRAE, BioinfOmics, Plant Bioinformatics Facility, Versailles, France

Université Paris-Saclay, INRAE, URGI, Versailles, France
e-mail: cyril.pommier@inrae.fr

F. Coppens · F. D'Anna
Department of Plant Biotechnology and Bioinformatics, Ghent University, Ghent, Belgium
VIB Center for Plant Systems Biology, Ghent, Belgium

H. Ćwiek-Kupczyńska
Institute of Plant Genetics, Polish Academy of Sciences, Poznań, Poland

D. Faria
LASIGE, Faculdade de Ciências, Universidade de Lisboa, Lisbon, Portugal
BioData.pt, Instituto Gulbenkian de Ciência, Oeiras, Portugal

S. Beier
Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) Gatersleben, Stadt Seeland, Germany

C. Miguel
Instituto de Biologia e Tecnologia Experimental, Oeiras, Portugal
Biosystems and Integrative Sciences Institute, Faculdade de Ciências, Universidade de Lisboa, Lisbon, Portugal

S. Owen
Department of Computer Science, University of Manchester, Manchester, UK

K. Gruden
Department of Biotechnology and Systems Biology, National Institute of Biology, Ljubljana, Slovenia

© The Author(s) 2023

H. F. Williamson, S. Leonelli (eds.), *Towards Responsible Plant Data Linkage: Data Challenges for Agricultural Research and Development*,
https://doi.org/10.1007/978-3-031-13276-6_8

(MIAPPE, www.miappe.org) that handles general metadata organization and its companion web service API, the Breeding API (www.brapi.org). Both rely on two established data standards, the MultiCrop Passport Descriptors (MCPD) for identification of plant genetic resources and the Crop Ontology (www.croponontology.org) for trait documentation. Researcher communities' coordination and collaborative approaches have enabled the success and adoption of MIAPPE and led to a general data lifecycle description by ELIXIR Plant Sciences Community to identify gaps and needed developments. A priority has been placed on addressing the "first mile" of data publishing, i.e. the gathering and documentation of data by the researcher, which enables relevant data findability and reusability. Here we describe the existing ecosystem of tools and standards for plant scientists as well as their history, including their convergence through the use of MIAPPE for describing genotyping datasets.

1 Introduction

Plant research communities have invested a lot of effort not only in increasing biological knowledge of the plant realm but also in enabling greater sustainability of plant production. Indeed, climate change and the increasing world population have led to agricultural considerations being identified in 12 of the 17 United Nation Sustainable Development Goals (<https://sdgs.un.org/goals>). As a consequence, more and more data is being produced and we are now facing a range of Big Data challenges, as described in the 4Vs of Big Data (Velocity, Variability, Volume, Veracity) (De Mauro et al., 2016). Volume and Velocity may be less of an issue for plant sciences in comparison to other scientific fields such as astronomy for instance, but Variability is especially challenging both because of the genomic complexity of plants (polyploidy, for example) and because of the heterogeneous nature of plant phenomics. The latter encompasses all the observations and measures that can be made on a precisely identified plant material in a characterized environment. This very general definition of phenomics (Watt et al., 2020) includes diverse types of properties and variables measured at different physical (Tardieu et al., 2017) and temporal scales, ranging from field observation of plant populations to molecular cell characterizations, and for some research communities includes metabolomics or gene expression. The acquisition of these data is conducted in various experimental facilities like greenhouses, fields, phenotyping networks, or natural sites. It can be done using many different devices from manual measurements to high throughput means. The resulting complex and heterogeneous datasets include all the environment and phenotypic variable values at each relevant scale (plant, micro plot, and so on) and very importantly the identification of the phenotyped germplasm, i.e., the plant material being experimented upon. In addition, there are often relationships between levels (i.e. physical scales such as microplot or plant individual and organs) inside datasets and between different datasets. The resulting rich wealth of data is

usually formatted in a very heterogeneous manner and is difficult to integrate manually or automatically.

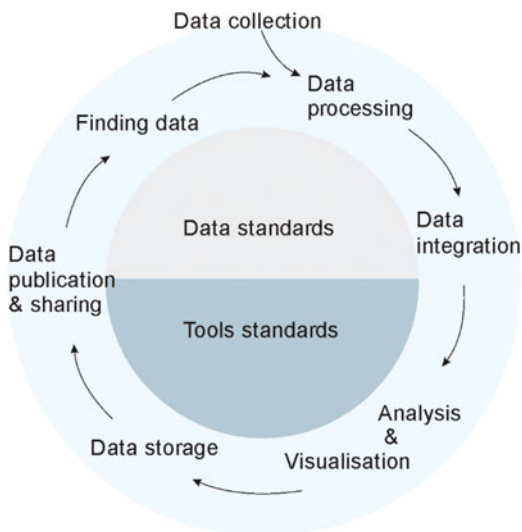
It is therefore a necessity not only to produce scientific data – from genetic and genomic to phenomic and environmental and up to systems biology – but also to develop the means for managing it, integrating it and therefore analysing it at high throughput dimensions, not only for model species such as *Arabidopsis* but also for crops and trees. This management is therefore a direct application of the FAIR (Findable, Accessible, Interoperable, Reusable) data principles (Wilkinson et al., 2016), especially crucial in regard to such a complex data life cycle. In the present paper we will describe the plant research data lifecycle and data management challenges as well as the solutions developed by different communities over the past years. This is followed by a focus on the ‘first mile challenge’ and some considerations on findability of data across distributed data repositories.

2 The Plant Data Life Cycle

The plant sciences community of the European Infrastructure for life sciences ELIXIR (Harrow et al., 2021) has been structured in recent years by funded activities such as the ELIXIR EXCELERATE Horizon 2020 project as well as its collaboration with the EMPHASIS European Infrastructure for Plant Phenotyping. With a switch from big structuring projects to a funding model that relies on the coordination of many smaller projects, including national projects such as ELIXIR implementation studies or European Open Science Cloud (EOSC) demonstrations, it has been necessary to create a roadmap to organize and coordinate the necessary activities. This roadmap (Pommier et al., 2021) needed first a general description of its objectives, through the definition of a data life cycle underlying the needs of plant science.

The building of this roadmap has been possible thanks to many years of collaborations within and between different groups with tangible results, including the groups responsible for the Minimum Information About a Plant Phenotyping Experiment (MIAPPE) standard (Papoutsoglou et al., 2020); BrAPI, the plant breeding API (Selby et al., 2019); the International Wheat Information System (WheatIS); and the transPLANT genomic infrastructure, among others. This facilitated a community of ideas that drafted the life cycle and the roadmap. We ensured the openness of the community both by welcoming new members and institutions and by setting up formal collaborations between ELIXIR and other groups such as EMPHASIS in particular. The structure of this community, *i.e.* a European infrastructure that relies on a network of national nodes supported and encouraged at each national level, has further pushed forward the activities. These activities are both bottom-up, with concrete use cases and demonstration datasets used as the basis for discussions by the persons in charge of actually running data related activities, and top-down, with strategic decisions made by principal investigators and node representatives to increase collaboration between communities and infrastructures. The existence of

Fig. 1 Schema of the data lifecycle, that begins with data gathering, then preparation for analysis through processing and integration, finally publication and sharing for knowledge extraction. Note that the integration encompasses both statistical data integration and normalization as well as data linking and mapping



this bottom-up approach, mobilising concrete elements and utilising real datasets to demonstrate the validity of the data standards and the data life cycle elements, has been instrumental to trigger interest and collaboration within the ELIXIR plant community. Last but not least, the fact that the elements of the roadmap and our objectives have been included as deliverables of many projects ensures that they will be realised.

The activities described within the Roadmap are designed with the goal of enabling successful data handling across the complete data life cycle (Fig. 1). As a community, we will focus on tailoring data findability both at the level of describing the generated datasets in order to be discoverable as well as via developing tools for data retrieval. We will also define pipelines for efficient data pre-processing, integration, analysis and visualisation to enable successful biological interpretation of results. For the last part of the data life cycle we will work on data storage in accordance with different standards and describe it using appropriate vocabularies and ontologies. This will enable publication of plant data and scientific papers in accordance with FAIR principles.

3 Plant Data Management Challenges

Plant research communities handle different types of data, some of which are shared with other realms, like genomics, genetics and systems biology, while others are very specific, like phenotyping and plant-environment interactions. Existing data standards and management practices have offered practical solutions for genomics and genetics, but phenotyping needed a whole new framework. Several communities

have built (separately or jointly) their solutions through three types of data standards: Semantic, Structural and Technical.

Semantic standards provide means for data description. They encompass controlled vocabularies, with term names and definitions, possibly organised as ontologies through the addition of semantic linking. For plant phenotyping, 10 years ago the Crop Ontology (Shrestha et al., 2012) formalized the Trait-Method-Scale model that has subsequently been embedded in MIAPPE. This model is aligned with the practices and approaches commonly applied by agronomists and phenotyping researchers, especially in terms of the terminology, organisation and range of descriptors for documenting observed and measured variables. Such semantic standards make researchers' life easier by providing a description framework and common vocabularies. They are therefore mostly driven by biologists. However, they don't answer the problem of data organisation.

Structure standards allow the organisation of datasets through schemas of metadata descriptors, *i.e.*, sets of fields, possibly grouped hierarchically, including mandatory and recommended information. They allow the description of data and defining of the interrelations between the different data files that gather all the measures and analyses done (something that is especially challenging in multilocal and/or multiannual experimental networks). The MIAPPE standard is a good example of such a standard. It takes advantage of the Investigation, Study, Assay (ISA) (Sansone et al., 2012) approach and encompasses elements taken from the Crop Ontology and the MultiCrop Passport Descriptor (MCPD) format, which is the reference for identifying and describing plant genetic resources and varieties within international genebanks (Alercia et al., 2015). In light of current research technologies and the increasing amounts of data being produced, those standards should allow one to organise data and metadata in a machine actionable way. At the same time, they must also be usable and still explicit enough to be effectively adopted by plant researchers. Therefore, they are built through close collaboration between computer scientists, biologists and agronomists who organise the data and provide their semantic description to ensure long term understandability and reusability.

Technical standards address interoperability challenges and data exchange between databases, tools and analysis environments. They include, for instance, web service APIs such as the Breeding API, a web service specification implementing MIAPPE. These standards are thus mostly computer scientist driven.

4 Plant Data Standards

The history of the Minimum Information About a Plant Phenotyping Experiment standard, MIAPPE, shows how a solution designed for a focused use case, plant phenotyping experiments, can be extended to other data types and in particular to genetic variation. The goal of MIAPPE (Papoutsoglou et al., 2020) is to support researchers in explaining plant phenotypes, *i.e.* the observable results of

the growth of plants in specific environmental conditions. To disentangle genotype-environment interactions and identify the biological mechanisms leading to specific phenotypes, a good description of the biological material, environmental conditions and observed variables is needed. These constitute the cornerstones of plant phenotyping experiments, and thus also comprise the pivots of phenotyping experiment description in MIAPPE.

The description of biological material in a plant experiment encompasses characteristics of the studied genotypes, such as their origin (the source of the seeds/plant parts, their pedigree) or taxonomic classification. The environment description comprises geographical locations, environment type and growth conditions, and additional treatments applied. In addition, the design of the experiment needs to be included (e.g., spatial and hierarchical arrangements of observation units, temporal arrangement of actions and events). Such a description of a plant experiment, if sufficiently detailed, allows us to understand the genetics and environmental factors that interact and produce particular phenotypes.

Observations are carried out for individual observation units at desired time points, and in phenotyping assays they typically involve the measurements of macroscopic plant traits (anatomy, yield, physiology, etc.) and environmental variables (actual environmental conditions). However, these are not the only measurement types that can be observed. More and more frequently, notably in systems biology approaches, the same plant experiments are the source of samples for other assays, such as microscopic measurements and multi-omics studies. Development of new technologies and gradually decreasing costs make this type of analyses more affordable and allow plant researchers to repeat the assays in multiple environments and time points. These omics data need to be properly placed in the context of the whole plant experiment. Thus, MIAPPE constitutes a solid foundation not only to integrate different types of data but also to build bridges between communities.

This notion of integration of different data types is broad and encompasses multiple things. Indeed, for a biostatistician or a data scientist, integrating means normalizing the data, reducing it to make its analysis and understanding possible. But from a data management point of view, integrating datasets is about finding links and common keys among several datasets. Since plant science relies heavily on the integration between phenomic, environmental and omics data, it is necessary to link them using common pivot objects (Pommier et al., 2019), *i.e.* common keys. The work of the ELIXIR implementation study FONDUE sets up such common keys between databases and datasets. The concepts and descriptions from MIAPPE are used to describe the environment and the plant material in genotyping experiments published in EMBL-EBI data repositories, especially Biosamples, the European Nucleotide Archive (ENA) and the European Variation Archive (EVA). This allows both findability and interoperability with other data repositories. A BioSamples checklist, based on a reduced set of MIAPPE metadata, has been developed to describe the samples more precisely and uniformly. In addition, the FONDUE project is developing further recommendations for new metadata information in the header of genotyping files (Danecek et al., 2011), such as the BioSamples sample identifiers.

5 Plant Data Standards History, Use and Adoption

The community management and gathering around the data standards described above has been highly collaborative. The history of the use by ELIXIR of the Crop Ontology, initiated by Bioversity International of the CGIAR, is a good example of how we decided to join forces to avoid creating new data standards when possible. Giving up existing institution-specific standards in favor of international ones, *e.g.* Crop Ontology, allowed to avoid unnecessary competition. Furthermore, through common workshops, like the PhenoHarmonIS series of conferences, and projects, such as ELIXIR EXCELERATE, new communities have been invited to actively contribute to the Crop Ontology, either through dedicated sub-ontologies or at the level of the ontology's formal concepts.

The building of MIAPPE has been even more community intensive. It has been built through close collaboration among European research groups, initially between European Union FP7 projects transPLANT and the European Plant Phenotyping Network (EPPN), together with the CGIAR, later followed by European infrastructures ELIXIR, EMPHASIS. Eventually, MIAPPE became an open project, and multiple plant researchers were invited to contribute to its development and to adopt and introduce the standard in their communities. To collect as broad a range of feedback as possible, open collaboration and requests for comments were organised and advertised in plant-focused events (conferences, webinars, mailing lists). In parallel, MIAPPE has been kept up-to-date with other external activities (BrAPI and Crop Ontology, among others) and projects. Prioritizing the directions for development and drafting new versions were led by dedicated working groups in connection with the MIAPPE steering committee. Currently, MIAPPE governance is minimal and pragmatic, with the steering committee in charge of discussing and organizing decisions around the evolution of MIAPPE specifications, an issue tracker on github to follow evolution requests, and a website and mailing lists for announcements. An important part of the current life of MIAPPE is about outreach, promotion and further adoption of the standard. This is done through webinars, training or workshops and is handled by any member of the MIAPPE community. Coordination of decision making is done by the six members of the steering committee. Any group can propose an addition to MIAPPE, formalize it, and bring it to the committee, which will organize the adequate consultation. Therefore, the evolution of MIAPPE relies on the willingness of self-elected workgroups that will hammer down all the details of their propositions through meetings and workshops. It is noteworthy that some of the most important evolutions of MIAPPE were made thanks to several EU projects that funded a group of people to work on a given problem, and were required to deliver a working solution. The available time of members of the focus groups has been critical here. As a consequence, the decision-making process takes a lot of time but ensures both the quality of the evolutions and improvements, which are tested with real datasets, as well as the fairness of the decisions, to ensure any stakeholder constraints will be taken into account.

The building of BrAPI took a similar path. Indeed, in the BrAPI community, initiated by the CGIAR and Cornell University, several groups from the European infrastructures and the CGIAR brought forward their use cases with the set of necessary specifications and web service calls. Here again, like in MIAPPE, those theoretical models were both tested with real data and gathered in a consensus specification during dedicated hackathons. These one week events, occurring twice a year, were organised both with the dedication of the BrAPI community members, who took turns organising them, and thanks to the Gates Foundation, who initiated BrAPI by funding those events for 3 years. Here again, like with MIAPPE, a need for coordinated governance arose quickly. The proposition was made to have a dedicated full-time coordinator, located at Cornell University. This structure proved highly successful for two important reasons. First, the coordinator is dedicated full time to the technical aspects, hence ensuring the consistency of the versions and the release cycle. Second, he does very important community management by organising the discussions on the mailing lists and github, by offering support, by organising further hackathons and everything needed to keep BrAPI members involved. One of the most important things in his activity, like in MIAPPE, is to ensure equity by being neutral, hence ensuring that the needs of all stakeholders – *e.g.* the CGIAR, ELIXIR, EMPHASIS, and national institutes – are taken into account, and that BrAPI doesn't get driven mainly by the needs of its most active members. For that purpose, a review board has been created.

The dissemination of data standards toward their end users, biologists and computer scientists, can be achieved through training, publications, workshop standard registries and open science policies and projects. Training, oral communication and valorisation have already been discussed and are actively used within ELIXIR and EMPHASIS communities. The WheatIS of the Wheat Initiative of the G20 has built standard recommendations to guide wheat researchers (Yeumo et al., 2017). It is indeed important to offer researchers the possibility to share recommended standards lists, carefully tailored for their domain, within the plethora of available standards and good practices. There are two types of recommendations that can be used. The first type are general recommendations targeting mainly computer scientists but which can also be used by researchers and principal investigators. The main example is FAIRsharing (The FAIRsharing Community et al., 2019) which can also be used to build collections dedicated to some communities, such as the WheatIS Data Interoperability Guidelines (<https://fairsharing.org/collection/WheatDataInteroperabilityGuidelines>). Those can furthermore be hosted and promoted in dedicated community registries (<http://wheatis.org/DataStandards.php>) in order to foster good practices within one-stop community web portals.

There are also activities that promote the use of standards in the frame of general data management and stewardship practices, such as the ELIXIR Research Data Management (RDM) toolkit developed within ELIXIR CONVERGE (<https://rdmkit.elixir-europe.org>). It contains, among other things, guides for using specialised tool assemblies like the Plant Genomics Assembly (https://rdmkit.elixir-europe.org/plant_genomics_assembly.html) in order to promote the use of

community standards, and supplements this with prominent data management tools and best practices. The RDM Toolkit is an interesting example of how data sharing practices can be brought to researchers, at a minimal and therefore affordable cost of invested time, and by putting its content directly in the hands of the researchers that will ultimately use it.

The example of the INRAE (National Research Institute for Agriculture, Food and Environment, France) forest tree community is interesting because it shows how the data lifecycle can be updated to cope with the constraints of data acquisition and data sharing through the use of the plant data standards described here. An automated data flow was set up to synchronize data shared by multiple information systems. Data produced by research teams are managed in local experimental information systems, each platform having its own. The system facilitates the daily management of those data (raw, analysed or inventory data). Publication and sharing is not done at this local level, however, but through a global information system using an automated workflow. This makes it possible to improve the datasets' visibility and interoperability in order to share this knowledge and enhance data quality, reuse and enrichment.

6 The First Mile Challenge

The first step of the data life cycle is to gather and organize the data needed to answer a given scientific question. This can be achieved by documenting the data during experimentation or by adding and organizing the necessary metadata to existing datasets. For phenomics, this documentation process relies on dedicated tools and laboratory information management systems (LIMS) such as PIPPA (Coppens et al., 2017), Breedbase (Fernandez-Pozo et al., 2015; Agbona et al., this volume) or PHIS (Neveu et al., 2018). For omics data, equivalent LIMS systems exist to run an experiment. But while those systems are commonly available with high throughput experimental platforms, there is a need for another tool both for classical experimentation management and for managing the data obtained from integration and reduction of the experimental datasets. The solution came from a joint activity between ELIXIR, EOSC and EMPHASIS to add MIAPPE to the FAIRDOM-SEEK (Wolstencroft et al., 2017) data management system.

From a community point of view, this collaboration went smoothly thanks to the quality of the existing software, the willingness of all partners and the existence of an accepted and published standard, MIAPPE. Indeed, there hasn't been any extended discussion on the selection of metadata and fields, something that commonly occurs in those types of development projects; MIAPPE was simply selected and implemented in FAIRDOM-SEEK. The fact that both MIAPPE and FAIRDOM-SEEK share the common backbone of the ISA tools helped a lot.

7 The Findability Challenge for Dispersed Community Data

Data discovery, i.e. the ability for researchers to find any dataset suitable for their scientific questions, is a very active domain nowadays, with many solutions. Indeed, to find data, one can either (i) query all relevant data repositories one after the other, (ii) use one or many data discovery web portals, or (iii) use general search engines such as google data search. The first solution isn't appropriate in our era of big data. The third approach might be too general, hence lacking specificity without the help of dedicated markups such as bioschemas.org. The second approach leads to building community global portals to cross data repository boundaries. The plant community has in particular built two of them: the WheatIS (<http://wheatis.org/Search.php>) and FAIDARE (<https://urgi.versailles.inrae.fr/aidare/>).

The WheatIS (Sen et al., 2020) is an interesting example from a community management point of view. Indeed, it showed that the success of this portal relies on several key points: (i) keeping data distributed in a global federation rather than gathering all of it in a single global data repository, (ii) sharing at least one critical need, (iii) having clear leadership that ensures mutual benefit and relies on engaged people from several institutions, and (iv) gathering experts and making it easy to join the data federation through technical simplicity.

FAIDARE goes one step beyond this; first, by extending the WheatIS species range to more crops and, second, by enabling MIAPPE and BrAPI data standards. This brings two main benefits: Including in the FAIDARE federation new BrAPI data repositories at no additional cost and providing refined search based on the MIAPPE/BrAPI metadata. The key to success of FAIDARE relies on its ability to extend the data federation it indexes by merging the BrAPI network with the WheatIS network, and later with the addition of Bioschemas.org sources.

8 Conclusion

Sharing data to ensure its useful reuse is complex and poses major technical, social and scientific challenges. The present paper has shown how some international communities, through their social interactions, managed to build technical solutions to enable FAIR data management throughout the data lifecycle. We have seen in particular that the most time-consuming challenge is the community management, not only to formalize the standards, but also to build adoption and train users in the long term. This can be done in a sustainable way if first adopters become trainers too, through 'train the trainers' initiatives, and ensure adoption by other research networks. The solutions presented in this paper are already helping a lot, but more work is needed to allow easy data management to be realised. This will occur through technical improvements and their sharing, adaptation and adoption, and a lot of activities are already ongoing in this regard. But some social and scientific aspects

have not really been discussed yet, in particular the criterion to select data that needs to be shared for the future. Indeed, from raw data to computed, reduced data, there are huge volumes of data to be stored, much more than the expected storage capabilities. We know that to enable reproducibility we should aim at sharing raw data, but that is often where the highest volume lies. The research community is therefore waiting for debates and guidance to make the right choices on such future issues.

References

- Alercia, A. et al. (2015). *FAO/Bioversity multi-crop passport descriptors V.2.1 [MCPD V.2.1]. DataCite*. <https://doi.org/10.13140/rg.2.1.4280.2001>
- Coppens, F., et al. (2017). Unlocking the potential of plant phenotyping data through integration and data-driven approaches. *Current Opinion in Systems Biology*, 4, 58–63. <https://doi.org/10.1016/j.coisb.2017.07.002>
- Danecek, P., et al. (2011). The variant call format and VCFtools. *Bioinformatics*, 27(15), 2156–2158. <https://doi.org/10.1093/bioinformatics/btr330>
- De Mauro, A., et al. (2016). A formal definition of big data based on its essential features. *Library Review*, 65(3), 122–135. <https://doi.org/10.1108/LR-06-2015-0061>
- Fernandez-Pozo, N., et al. (2015). The sol genomics network (SGN) – From genotype to phenotype to breeding. *Nucleic Acids Research*, 43(D1), D1036–D1041. <https://doi.org/10.1093/nar/gku1195>
- Harrow, J., et al. (2021). ELIXIR-EXCELERATE: Establishing Europe’s data infrastructure for the life science research of the future. *The EMBO Journal*. <https://doi.org/10.15252/embj.2020107409>
- Neveu, P., et al. (2018). Dealing with multi-source and multi-scale information in plant phenomics: The ontology-driven phenotyping hybrid information system. *New Phytologist*. <https://doi.org/10.1111/nph.15385>
- Papoutsoglou, E. A., et al. (2020). Enabling reusability of plant phenomic datasets with MIAPPE 1.1. *New Phytologist*, 227(1), 260–273. <https://doi.org/10.1111/nph.16544>
- Pommier, C., et al. (2019). Applying FAIR principles to plant phenotypic data management in GnpIS. *Plant Phenomics*, 2019, 1–15. <https://doi.org/10.34133/2019/1671403>
- Pommier, C., et al. (2021). *ELIXIR plant sciences 2020–2023 roadmap*. F1000 Research Limited. <https://doi.org/10.7490/F1000RESEARCH.1118482.1>
- Sansone, S.-A., et al. (2012). Toward interoperable bioscience data. *Nature Genetics*, 44(2), 121–126. <https://doi.org/10.1038/ng.1054>
- Selby, P., Abbeles, R., Backlund, J. E., Basterrechea Salido, M., Bauchet, G., Benites-Alfaro, O. E., Birkett, C., Calaminos, V. C., Carceller, P., Cornut, G., Vasques Costa, B., Edwards, J.D., Finkers, R., Yanxin Gao, S., Ghaffar, M., Glaser, P., Guignon, V., Hok, P., Kilian, A., König, P., Lagare, J. E. B., Lange, M., Laporte, M. A., Larmande, P., LeBauer, D. S., Lyon, D. A., Marshall, D. S., Matthews, D., Milne, I., Mistry, N., Morales, N., Mueller, L.A., Neveu, P., Papoutsoglou, E., Pearce, B., Perez-Masias, I., Pommier, C., Ramírez-González, R. H., Rathore, A., Raquel, A. M., Raubach, S., Rife, T., Robbins, K., Rouard, M., Sarma, C., Scholz, U., Sempéré, G., Shaw, P. D., Simon, R., Soldevilla, N., Stephen, G., Sun, Q., Tovar, C., Uszynski, G., & Verouden, M. (2019). BrAPI consortium: BrAPI-an application programming interface for plant breeding applications. *Bioinformatics*, 35(20):4147–4155. <https://doi.org/10.1093/bioinformatics/btz190>. PMID: 30903186; PMCID: PMC6792114.

- Sen, T. Z., et al. (2020). Building a successful international research community through data sharing: The case of the wheat information system (WheatIS). *F1000Research*, 9, 536. <https://doi.org/10.12688/f1000research.23525.1>
- Shrestha, R., et al. (2012). Bridging the phenotypic and genetic data useful for integrated breeding through a data annotation using the crop ontology developed by the crop communities of practice. *Frontiers in Physiology*, 3. <https://doi.org/10.3389/fphys.2012.00326>
- Tardieu, F., et al. (2017). Plant phenomics, from sensors to knowledge. *Current Biology*, 27(15), R770–R783. <https://doi.org/10.1016/j.cub.2017.05.055>
- Watt, M., et al. (2020). Phenotyping: New windows into the plant for breeders. *Annual Review of Plant Biology*, 71(1), 689–712. <https://doi.org/10.1146/annurev-arplant-042916-041124>
- Wilkinson, M. D., et al. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3, 160018. <https://doi.org/10.1038/sdata.2016.18>
- Wolstencroft, K., et al. (2017). FAIRDOMHub: A repository and collaboration environment for sharing systems biology research. *Nucleic Acids Research*, 45(D1), D404–D407. <https://doi.org/10.1093/nar/gkw1032>
- Yeumo, E. D., et al. (2017). Developing data interoperability using standards: A wheat community use case. *F1000Research*, 6, 1843. <https://doi.org/10.12688/f1000research.12234.1>
- The FAIRsharing Community, et al. (2019). FAIRsharing as a community approach to standards, repositories and policies. *Nature Biotechnology*, 37(4), 358–367. <https://doi.org/10.1038/s41587-019-0080-8>

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), 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 chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter'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.

