Online Research @ Cardiff

This is an Open Access document downloaded from ORCA, Cardiff University's institutional repository: https://orca.cardiff.ac.uk/101571/

This is the author's version of a work that was submitted to / accepted for publication.

Citation for final published version:

Leonelli, Sabina, Davey, Robert P., Arnaud, Elizabeth, Parry, Geraint and Bastow, Ruth 2017. Data management and best practice for plant science. Nature Plants 3 (6) , 17086. 10.1038/nplants.2017.86 file

Publishers page: http://dx.doi.org/10.1038/nplants.2017.86 <http://dx.doi.org/10.1038/nplants.2017.86>

Please note:
Changes made as a result of publishing processes such as copy-editing, formatting and page numbers may not be reflected in this version. For the definitive version of this publication, please refer to the published source. You are advised to consult the publisher's version if you wish to cite this paper.

This version is being made available in accordance with publisher policies. See http://orca.cf.ac.uk/policies.html for usage policies. Copyright and moral rights for publications made available in ORCA are retained by the copyright holders.

information services
gwasanaethau gwybodaeth

Paper published under the title "Data Management and Best Practice in Plant Science" in Nature Plants, June 2017

Best Practice in Data Management: The New Frontier for Plant Science

Sabina Leonelli[1,2], Robert P. Davey[3], Elizabeth Arnaud[4], Geraint Parry[5], Ruth Bastow[5,6]*

1- Department of Sociology, Philosophy and Anthropology & Exeter Centre for the Study of the Life Sciences, Byrne House, Exeter University, Exeter EX4 4PJ, UK
2- School of Humanities, University of Adelaide, Adelaide 5005, Australia
3- The Earlham Institute, Norwich Research Park, Norwich, NR4 7UG, UK
4- Bioversity International, Parc Scientifique Agropolis II, 34397 Montpellier Cedex 5, France
5- GARNet, School of Biosciences, Cardiff University, Cardiff, CF10 3AX, UK
6- Global Plant Council, 1a Bow Lane, London EC4M 9EE, UK

* Corresponding author: s.leonelli@exeter.ac.uk

Introduction: Making Open Data Work for Research

Policies that encourage Open Data are gaining support around the globe, and are now enforced by prominent funders and publishers including the Nature group and the Gates foundation [1-3]. This has enormous potential for transforming the production, management, and dissemination of scientific knowledge, particularly by enabling researchers to interrogate and re-use existing datasets [4, 5]. However, realising this potential is not straightforward [6]. As in other scientific domains, plant scientists across the world generate data of various types (quantitative, qualitative, text, computed values) in diverse formats and in ever-increasing abundance. Arguably the analysis of field-level agriculture might require the integration of as broad a variety of data as in any other discipline. This will range from the multiple growth parameters of individual plants, through field-level analysis of hydrodynamics and gas exchange to soil and climate level data that impacts regional and global growth rates. These datasets are of tremendous importance for research on crop production, food security, climate change resilience, nutrition for health, and sustainable agriculture.

Making such data widely available requires appropriate infrastructures and tools to organise, annotate, integrate, and share data, without which researchers cannot easily enrich existing data resources nor can they discover and reuse data.

TABLE 1 – General tools for data management: typology and exemplars

Type of tool	*Function*	*Examples of relevance to the plant community*
Open Lab Books	Digital and shareable version of traditional lab book	RSpace [URL: https://www.researchspace.com/]
Generic Open Data Repositories	General storage for many different data types	Figshare [figshare.com] DataVerse [dataverse.org]
Specific Databases	Fine-grained datasets which require subject-specific metadata	TAIR [www.arabidopsis.org] BAR [bar.utoronto.ca] iHub [http://www.ionomicshub.org/home/PiiMS]
Data Portal	Aggregating and providing visibility for various databases and resources	Araport [araport.org] Biosharing [biosharing.org/] Agroportal [agroportal.lirmm.fr/]
Bio-Ontologies	Keywords for the annotation, ordering and retrieval of data	Plant Ontology [planteome.org] Crop Ontology [cropontology.org]
Metadata Standards	Standardise experimental data collection	Minimal Information on Biological and Biomedical Investigations (MIBBI) [biosharing.org/standards/] Minimal Information About a Microarray Experiment (MIAME) [fged.org/projects/miame/] Minimal Information About Plant Phenotyping Experiments (MIAPPE) [cropnet.pl/phenotypes/?page_id=15]
Identifiers for Research Materials	Annotation and retrieval of research materials on which experiments were originally performed	Germplasm Resource Information Network – Global [www.grin-global.org/] Multi-Crop Passport Descriptors [bioversityinternational.org/e-library/publications/detail/faobioversity-multi-crop-passport-descriptors-v2-mcpd-v2/] Genesys [genesys-pgr.org]
Informatics Standards	Software tools helping to format, store and visualise data	Breeding API [docs.brapi.apiary.io/#] InterMINE [intermine.org]
Data Annotation Pipelines	Annotation of data from generation to re-use	Integrated Breeding Platform [integratedbreeding.net/] CropStore [cropstoredb.org/description.php] eDal [edal.ipk-gatersleben.de/]

Guidelines of Good Practice	Articulation of data management principles and actions fostering data re-use	FAIR Data [force11.org/group/fairgroup/fairprinciples] Wheat Data Interoperability Guidelines [rd-alliance.org/group/working-and-interest-group-chairs-wheat-data-interoperability-wg/outcomes/wheat-data]

Much has been written about the importance of funding and developing digital data repositories, data classification tools such as computational ontologies, and software that would enable the efficient retrieval and analysis of data available online [7]. What remains less explored is what scientists themselves can do to improve the ways in which they manage their own data for subsequent reuse. It is widely acknowledged that implementing Open Data guidelines involves a substantial shift in the resources and attention needed by researchers to handle the data produced in any one of their projects, and that this requires advance planning and strategic decisions by researchers, their institutions and publishing venues [8]. However, there is often a disconnect between the type of management plan that is appropriate on the institutional level and what is actually useful and effective for an individual researcher when collecting and analysing data. This generates confusion around how exactly to improve data management practices, who to ask for help, and what skills are needed to be able to curate and share their data in ways that fit the emerging landscape of Open and Big Data analysis.

This is troubling given the difficulties encountered by researchers in their day-to-day dealings with data. A common experience is being asked to submit a body of work, comprising one or more datasets, to various repositories where each requires a different and laborious annotation process. It is easy to get lost in a large range of metadata, ontologies and thesauri, especially since each type of standard and storage options is tailored to a different goal. For instance, generic data repositories are able to house many data types, and yet tend to be promoted to institutes or funders rather than researchers, and thus might not respond well to field-specific needs. Specific repositories can be so detailed in their requirements for data curation that they scare away individual users, who do not always have the resources and capacity to be able to learn the relevant skills.

Additionally, researchers typically consider data to be sensitive research outputs that can easily be misused or misinterpreted when taken out of context. As a result, many scientists may not trust global repositories that have no direct and personal connection to their own work. Whilst the cost of integration of a public dataset into their own analyses might be made cheaper by good open data practices, the cost of validating the trustworthiness of a public dataset may be too much to bear. Moreover, scientists have doubts about which repositories and software tools will survive in the longer term and whether they are worth investing time and effort to adopt. This uncertainty is well-justified, since identifying who will manage and support the long-term sustainability of data infrastructures in the face of exponential data generation, and how this can be achieved, remains a huge challenge. Efforts to enhance the future viability of data infrastructures include finding business models that are minimally dependent on support from funding bodies. A number of popular databases have opted for cost recovery and income generation mechanisms such as subscription

membership, private funding, or support by publishers, higher education institutions, and/or libraries [9-12]. This is a legitimate response, since governmental budgets for research are limited, prone to political vagaries, and vulnerable to economic trends. However, existing business models fall short of delivering data resources that are freely and widely accessible as well as well-maintained, regularly updated and sustainable.

When confronted with so much uncertainty, many researchers end up developing their own unique controlled vocabularies and ontologies, metadata formats, software and storage facilities. Consider the case of plant trait data. These are typically generated by diverse, costly and time-consuming experiments, and thus can hugely benefit from increased data sharing and integration [13]. However, trait data are collected from a range of different experimental locations from the field to the greenhouse, in controlled environmental chambers, and in laboratories. Also, the data are generated by individual researchers in a variety of ways, ranging from manual measurements to remote sensing, and can be recorded by hand, in electronic lab/fieldbook or automatically captured. Given the range and breadth of variables involved and the difficulties to identify and apply standards, plant trait data are typically stored in an individual's computer, often in loosely formatted spreadsheets – or, at best, in a local database.

This produces a fragmented and confusing landscape of resources that are often disconnected from other relevant initiatives, and require expensive and highly specialised curation in order to be made compatible and interoperable with other datasets and international databases. The lack of compatibility and links between plant trait databases is particularly problematic when attempting to associate plant data with climate or environmental data of relevance to studying gene-environment interactions. In order to support spatial and temporal modelling of trait and crop performance, plant trait data must be associated with measures of the environment, alongside crop management and agronomic practice. These aspects are key for understanding complex multi-variate theories, such as the impact of climate change on crop production, and yet such data integration is impossible to achieve without using common standards and databases. The lack of standardisation and dialogue between data producers also generates semantic obstacles to the development of interoperable phenotypic databases. Researchers from different parts of the world typically describe plant traits differently, depending on geographical location, culture and language [14]. This makes it hard to identify which datasets pertain to the same trait. Furthermore, if experimental findings are not documented via internationally approved germplasm identifiers, they can be impossible to link back to their original sources.

What Solutions Exist for Plant Scientists?

Projects need to find and apply a common pool of standard variables, bio-ontologies, and controlled vocabularies to support their data management strategies, creating semantic links between available data and enabling the analytical interoperability of datasets.

TEXTBOX 1 – Bio-Ontologies for the Plant Sciences
An ontology is a formal representation of a domain knowledge for the purpose of annotating data, where key concepts are defined, as well as the relationships that exist between those concepts. The most popular ontology within the life sciences is the Gene Ontology, developed by a consortium administered by a central office at the European Bioinformatics Institute to manage the dissemination and retrieval of gene product information across several model species. For plant science, the most used ontologies include the Plant Ontology, Trait Ontology, Crop Ontology, Environment Ontology, Agronomy Ontology and Taxonomy Ontology.

To expand data discovery and data integration abilities, scientists need over-arching ontologies where concepts are independent of the species. This is now starting to be addressed by the set of reference plant ontologies of the Planteome [15]. This is a long term undertaking requiring the engagement of many projects around the world that already share this type of data, such as the Breeding Management System of Integrated Breeding Platform [16], the Next Generation Breeding Databases of the CGIAR Root, Tubers and Banana Research Programme [17], the Global Agricultural Trial Repository [18] the Agricultural Model Intercomparison and Improvement Project [19], and the CGIAR Big Data Platform project [20].

Several initiatives are providing co-ordinated frameworks to foster the integration of diverse datasets, comparative analysis across species, as well as computational reasoning that will infer knowledge. For instance, the Crop Ontology [14, 21] project works with communities of practice centred on a specific crop to develop and curate list of controlled vocabularies for crop traits along with an associated measurement method and scale. This approach provides a common schema for researchers to systematically quantify and describe plant phenotypes that will support future data integration, interpretation, and analysis. The emphasis is on participatory development and for curation to accurately and proactively meet user needs. Furthermore, Crop Ontology facilitates data re-use by intersecting with existing data annotation pipelines such as the Integrated Breeding Platform [16], and is accessible in the emerging agronomic ontology repository 'Agroportal' using the NCBO Bioportal technology [22].

Making plant trait data fit for modelling and analysis includes the harmonisation and quality control of annotations from data capture to publishing. This typically requires extensive manual labour, including extraction of data from published papers, as well as semi-or fully-automated annotation pipelines that can work over existing material, such as electronic field books. Projects such as the Collaborative Open Plant Omics (COPO) [23], GnpIS [24] and the Agroportal project [22] are developing user-oriented services to simplify annotations of plant data using ontologies, thus making it easier for researchers on the front end to identify and use data management resources of relevance to their work. Making sure that data are annotated and shared in accordance with international standards is labour-intensive, and yet provides great value. For instance, datasets can be integrated to assess plant responses to various environments and stresses to uncover favourable traits for future breeding programmes targeted at productivity, sustainability, and resilience of crop varieties and agricultural systems.

TEXTBOX 2 – Developing international standards: ELIXIR and COPO
A good example of international collaboration on data management and related tools is the pan-European ELIXIR project [25], which engages policy-makers, funders and infrastructure coordinators across national borders in order to build open interoperable frameworks of software and data. ELIXIR is in the process of establishing "node resources", that is projects that meet the criteria for interoperable open science. One such projects is the Collaborative Open Plant Omics (COPO), devoted to the development of web-based interfaces for plant scientists to describe multi-omics datasets according to relevant domain metadata standards, submit these data to public repositories, and subsequently analyse, publish, and track their outputs through the research lifecycle [23]. To these aims, COPO uses already established technologies such as the ISA metadata tracking framework [26-27] and the computational infrastructure for life scientists CyVerse [28], runs tailored training workshops for end users, and supports the ORCiD researcher identification system [29], which allows published data to be linked to a user, their publications, their projects and grants. This way of working guarantees interoperability with other ongoing data management efforts at the international level, while also enhancing the visibility and long-term sustainability of the data being stored and disseminated through such tools, as well as empowering researchers to achieve recognition and credit for their data sharing efforts.

What Researchers Can Do

Scientific institutions, funding bodies, and publishers should play a leading role in incentivising data sharing in a standardised format, requiring academic institutions to commit resources to data curation and management, and committing resources towards establishing relevant training and rewards. At the same time, there are several steps that researchers themselves can take to support a scientifically fruitful transition to Open Data:

- *Use data management plans (DMPs) in grant proposals and technical documents as opportunities to streamline your work and increase international visibility*

A good DMP should set out the use of standard formats and reproducible ways of recording experimental procedures. It should also be a concrete commitment to adhering to best practice, as far as possible, with information on supported data types and formats, licences for data and software, and how these will be delivered throughout a given project. DMPs provide an opportunity to carefully consider and scope out the programme of work in advance of the project start date and to strategically think about how to minimize data management time during the project and generate datasets and information that can be used both during the project but also after its completion. The Digital Curation Centre (DCC) has produced a catalogue of resources that can assist researchers in constructing and making the most of their DMPs [30]. Once in place, publishing DMPs in a suitable publicly available space fosters the sharing of best practice.

- *Involve data managers, data curators and computer scientists within your own institution and community in ongoing research and future plans*

Computational sciences and biology need to move forward together to yield data infrastructures and associated analytical software that are simultaneously user-friendly yet resilient in the face of vertiginous computational demands. Scientists' needs, requirements and use cases must be communicated and frequently discussed with those developing data tools and resources. Without regular interactions between these groups, much of what is required on both sides may be lost in translation. It is essential that computational scientists, data managers and data curators are not just viewed as a support service but valued as researchers and collaborators in their own right, are integrated into project development and are provided with opportunities to meet biologists.

- *Critically evaluate which type of data curation you use, and why*

A minimal level of long-term support and manual curation for data infrastructures will be essential as new technologies are developed, novel data are created, new insights are added to the existing knowledge base, and unique experimental information is recorded. These activities are very time-consuming, so it is vital to be strategic. Just as they do whenever learning to use a given laboratory instruments, researchers should investigate whether existing data management tools can serve their needs, and adopt such tools whenever possible. Aggregation services such as Biosharing and Agroportal can help to identify available resources, including databases and metadata standards. Tracking progress against the DMP can also be helpful.

- *Carefully document metadata that conforms with existing standardisation projects.*

A key component of producing data that can be effectively reused is selection and management of the associated metadata, which can vary dramatically depending on which types of data are being curated. Projects like the Crop Ontology are making progress towards developing standards and guidelines for metadata annotation across many different data types of relevance to plant science. Minimal Information projects such as the Minimal Information About a Microarray Experiment (MIAME) [31-32] or the Minimal Information About a Plant Phenotyping Experiment (MIAPPE) [13] can also provide crucial guidance, and are increasingly built into data description pipelines. Increased interactions between researchers and data managers are needed to develop metadata standards that are fit for purpose for diverse data collected on different species across geographical and technological boundaries.

- *Participate in international efforts to develop good data management tools*

It is common for researchers approaching data management tools for the first time to find problems or some form of incompatibility with their own datasets. Given the diversity of data, methods and goals in question, this is hardly surprising. We recommend that instead of giving up on using existing resources at the first hurdle, researchers consider helping the developers of these tools to improve and expand their standards and services, so as to make them more widely applicable. As we argued above, this is preferable to developing one's own local database or standard. It is also

enormously useful to the scientists in charge of data management tools, who are often actively looking for feedback on their activities from users. Perhaps most importantly, it helps to establish communities of researchers that share similar needs and conceptions of best practice, and can compare and improve their methods and assumptions through mutual exchange. Establishing communities of practice that encompass a variety of stakeholders across geographical locations and research areas is central to making Open Data into a powerful tool for discovery. For example the Findable, Accessible, Interoperable and Reusable (FAIR) Guiding Principles [33] aim to set out best practice for open research data and metadata, including but not limited to the use of persistent identifiers for data description and retrieval, use of suitable authentication and authorisation for data services, and the adoption of community-accepted standards and open licenses. These principles are being implemented by working groups within the Research Data Alliance (RDA) [34] and the Global Open Data for Agriculture and Nutrition (GODAN) [35], which provide international venues for collaboration towards the articulation of good data management practices (such as for instance the Wheat data interoperability Guidelines produced by the Wheat Data interoperability working group [36]).

Conclusion: All Hands on Deck

Many of the suggestions above require changes in researchers' existing modes of work and the use of their collaborative networks. The potential results are worth the effort: Open Data makes research more transparent, accountable, and accessible, and implementing openness from the start of a research project keeps funders happy and ensures compliance with requirements from the start, rather than leaving researchers to struggle when reporting outcomes at the end of a grant. Open Science guidelines are not going away, and being able to report on all aspects of research data will become increasingly more important in the future. Critically, Open Data improves the ability of researchers to find and reuse datasets that could shape the direction of future research or prove key hypotheses. In increasingly uncertain times for research funding and global networking, these altruistic practices are sensibly cost-effective, and foster international and cross-disciplinary collaborations.

References

1- Nature Editorial. 2016. Announcement: Where are the data? **Nature** 537,138 doi:10.1038/537138a

2- European Commission, Directorate-General for Research and Innovation. 2016. Open Innovation, Open Science, Open to the World. [http://bookshop.europa.eu/en/open-innovation-open-science-open-to-the-world-pbKI0416263/]

3- Gates Foundation OA Policy. 2016 [http://www.gatesfoundation.org/How-We-Work/General-Information/Open-Access-Policy]

4- Science International. 2015. Open Data in a Big Data World.

[http://www.icsu.org/science-international/accord/open-data-in-a-big-data-world-short]

5- Leonelli, S. 2016. Data-Centric Biology: A Philosophical Study. Chicago University Press.

6- GARNet/Egenis Workshop. 2016. 'Integrating Large Data into Plant Science: From Big Data to Discovery'. [www.garnetcommunity.org.uk/sites/default/files/GARNetEgenis_Book.pdf]

7- Kitchin R, Collins S, Frost D. 2015. Funding models for Open Access digital data repositories, Online Information Review, Vol. 39 Iss: 5, pp.664 – 681 [http://www.emeraldinsight.com/doi/abs/10.1108/OIR-01-2015-0031]

8- EU Open Science Agenda. 2016. [ec.europa.eu/research/openscience/pdf/draft_european_open_science_agenda.pdf#view=fit&pagemode=none]

9- Gruber T. 2009. Ontology. Entry in The Encyclopedia of Database Systems. Liu L., Tamer Özsu M. (Eds.), Springer-Verlag.

10- Reiser L, Berardini TZ, Li D, Muller R, Strait EM, Li Q, Mezheritsky Y, Vetushko A, Huala E. 2016. Sustainable funding for biocuration: The Arabidopsis Information Resource (TAIR) as a case study of a subscription-based funding model. **Database (Oxford)** doi: 10.1093/database/baw018

11- Dryad [http://datadryad.org/]

12- Bastow R and Leonelli S. 2010. Sustainable digital infrastructure. *EMBO Reports,* 11(10): 730-735

13- Ćwiek-Kupczyńska H, Altmann T, Arend D, Arnaud E, Chen D, Cornut G, Fiorani F, Frohmberg W, Junker A, Klukas C, Lange M, Mazurek C, Nafissi A, Neveu P, van Oeveren J, Pommier C, Poorter H, Rocca-Serra P, Sansone SA, Scholz U, Schriek M, Seren U, Usadel B, Weise S, Kersey K, Krajewski P. 2016. Measures for interoperability of phenotypic data: minimum information requirements and formatting. **Plant Methods**, 12:44, DOI: 10.1186/s13007-016-0144-4

14- Shrestha R, Matteis L, Skofic M, Portugal A, McLaren G, Hyman G, Arnaud E. 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 Plant Physiology** v. 3 Article 326: doi: 10.3389/fphys.2012.00326

15- Planteome [planteome.org]

16- Integrated Breeding Platform [integratedbreeding.net/breeding-management-system]

17- CGIAR Roots Tubers and Bananas [http://www.rtb.cgiar.org/]

18- Global Agricultural Trial Repository [www.agtrials.org]

19- Agricultural Model Intercomparison and Improvement Project [http://www.agmip.org/]

20- CGIAR Big Data Platform [http://ciat.cgiar.org/global-partnerships/big-data-platform/]

21- Crop Ontology [cropontology.org]

22- Agroportal [agroportal.lirmm.fr/]

23- Collaborative Open Plant Omics [earlham.ac.uk/copo]

24- Genetic and Genomic Information System [https://urgi.versailles.inra.fr/gnpis/]

25- ELIXIR [https://www.elixir-europe.org/]

26- ISA metadata tracking framework [http://www.isa-tools.org]

27- Sansone SA *et al* 2012. Toward interoperable bioscience data. **Nature Genetics** 44, 121–126 doi:10.1038/ng.1054

28- CyVerse [http://www.cyverse.org/]

29- ORCiD Researcher Identification System [http://orcid.org/]

30- Digital Curation Centre (DCC) Data Management Plans [http://www.dcc.ac.uk/resources/data-management-plans]

31- Brazma A *et al* 2001. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. **Nature Genetics** 29 365-71 doi: 10.1038/ng1201-365

32- Minimal Information About a Microarray Experiment [http://fged.org/projects/miame/]

33- Wilkinson, MD *et al.* 2016. The FAIR Guiding Principles for scientific data management and stewardship. **Sci. Data** 3:160018 doi: 10.1038/sdata.2016.18

34- Research Data Alliance [https://www.rd-alliance.org/]

35- Global Open Data for Agriculture and Nutrition [http://www.godan.info]

36- Wheat Data Interoperability Recommendations. 2015. [https://www.rd-alliance.org/group/working-and-interest-group-chairs-wheat-data-interoperability-wg/outcomes/wheat-data]