From Farm to FAIR: The Trials of Linking and Sharing Wheat Research Data

Christopher John Rawlings and Robert P. Davey

Abstract This paper describes progress towards an integrated data framework that supports the sharing of data from the Designing Future Wheat (DFW) strategic research programme funded by the UK BBSRC. DFW is a 5 year project (https://designingfuturewheat.org.uk/) that spans eight research institutes and universities, and aims to deliver pre-breeding germplasm to breeders to improve and increase the genetic diversity of their breeding programs. DFW is committed to making its data open to the wider research community by adopting FAIR data sharing approaches. It is also a good example of a data-intensive strategic research programme which follows a cyclical Field-to-Lab-to-Field approach that is representative of much contemporary and multidisciplinary crop science research. However, even with dedicated funding to develop crop data research infrastructures within DFW, we found that there are many challenges that require pragmatic and flexible ways to enable them to interoperate. We present key DFW data resources as a case study to assess progress and discuss these challenges with a view to developing infrastructure that exposes metadata-rich datasets and that meets FAIR principles.

1 Background to Designing Future Wheat

The Designing Future Wheat (DFW) project is a strategic research programme funded by the UK BBSRC that spans eight research institutes and universities with aims to deliver pre-breeding germplasm to breeders to improve and increase the genetic diversity of their breeding programs. The DFW partners are John Innes Centre, Rothamsted Research, Earlham Institute, the Quadram Institute, the

C. J. Rawlings (✉)
Rothamsted Research, Harpenden, UK
e-mail: chris.rawlings@rothamsted.ac.uk

R. P. Davey
Earlham Institute, Norwich, UK
e-mail: robert.davey@earlham.ac.uk

© 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_6

European Bioinformatics Institute, the National Institute for Agricultural Botany and the Universities of Nottingham and Bristol. DFW was originally funded for 5 years (2017–2021) but has been extended a further year due to the impact of the COVID pandemic (https://designingfuturewheat.org.uk/). DFW builds on the success of an earlier BBSRC-funded cross-institute strategic collaboration – the Wheat Improvement Strategic Programme (WISP) which ran from 2011 to 2017 (http://www.cerealsdb.uk.net/cerealgenomics/WISP/). The genesis of both WISP and DFW were responses by BBSRC to an independent review of its funding of Crop Science by Prof. Chris Gilligan in 2005 which recommended, among other things, that BBSRC should increase and focus its funding on crops rather than solely model species. This was aimed at encouraging better coordination of research in the grasses and small grain cereals research community. A key outcome of the WISP program was to develop high throughput phenotyping techniques for use across large field experiments with thousands of plots; the success of this has resulted in the generation of large datasets, emphasizing the need for efficient data collation, storage and sharing platforms.

In addition to developing pre-breeding germplasm, an important aspect of DFW is an explicit declaration that results are made available free of intellectual property restrictions. This continues the principle adopted in previous BBSRC-funded wheat research initiatives such as WISP. A key output from DFW are panels of novel germplasm (seeds) that can be freely incorporated into other academic research or commercial breeding programmes by crop breeding companies that participate as collaborators, i.e. "pre-competitive breeding", or pre-breeding. Complementing this pre-breeding research in DFW is a wide-ranging and underpinning programme of genetics, genomics and trait biology research, including the generation of new genomics resources. The same principles of openness in the pre-breeding research applies to data from the wider project and DFW is committed to making data open to the whole research community by adopting FAIR data sharing approaches. DFW is therefore a good example of a data-intensive strategic research programme which follows a Field-to-Lab-to-Field approach that is representative of much contemporary and multidisciplinary crop science research.

2 Challenges and Approaches for Data Management

We will highlight a range of challenges and approaches to creating a consistent and reusable integrated strategy for common heterogeneous agricultural datasets. These challenges include the variety of trial designs, difficulties harmonising environmental data across remote sites, keeping up to date with data generation tools, technologies and formats (e.g. sensors, drones), and monitoring research outputs.

2.1 Characterising the Origins of Genetic Material

All crop improvement programmes focus on using diversity inherent in germplasm (seed) collections to access and exploit potentially beneficial traits. Measures of that diversity across clades, species or lines are used to describe and construct core collections. The ancestry and provenance of a seed provides a biological context crucial to integration. A well-managed genebank with readily available and good quality passport information (Food and Agriculture Organization of the United Nations, 2018) and minimal marker sets for establishing variation (e.g. SNPs) that identify germplasm are intrinsic to integration of data from derived samples. Active quality control becomes part of the story to ensure study design and experimentation have a firm basis to power biological interpretation.

In a healthcare research setting, metadata integration is becoming essential to bring together a patient's record with information about a cell line, a disease state, a set of phenotypes, etc., to understand patient characteristics and clinical outcomes. The main driver for standardisation was the development of nomenclature (e.g. Read Codes (digital.nhs.uk/services/terminology-and-classifications/read-codes) and SNOMED (Spiers et al., 2017)) to harmonise clinical information in electronic medical records. Biomedical ontologies captured terms that may have been specific to a disease or body part, which subsequently necessitated the development of increasingly overlapping domain ontologies, e.g. UBERON (Mungall et al., 2012), to facilitate integration of data which increasingly needed to be at the systems level.

Enabling biological integration through access to consistently described and integratable data sources is not a new concept (Berti-Equille, 2001), but has remained essential due to the "data deluge" and increasingly multi-disciplinary science. Care has to be taken that technical integration is also paired with biological integration, i.e. bringing datasets together can be technically feasible, but whether that is appropriate to a biological question is often down to availability, quality, and richness of metadata of individual datasets (Börnigen et al., 2015).

Genebanks can be integrated into the downstream data management process, leading to FAIRer data infrastructure solutions for crops (Lapatas et al., 2015). Therefore, genebanks and the researchers that use them have a key role to play in making sure germplasm is fit for purpose with a view to designing field trials based on representative genetic diversity, analogous to designing a clinical trial based on representative demographics.

2.2 Describing the Variety of Field Trial Experimental Data

Designing Future Wheat is a crop research programme with a major pre-breeding component conducted in collaboration between UK research institutes and with a consortium of commercial wheat breeders and seed companies. While the design of the pre-breeding trials follow industry standard multi-site germplasm evaluation, the

trials at each partner research site are designed to answer more specific questions and are hence more varied. This might comprise singular experimental outcomes, such as assessments of a particular treatment, but each of these experiments is carried out by different research scientists. This data is recorded and held in spreadsheets and/or bespoke databases, each with their differing formats between research organisations. As such, the data produced is not standardised between trials and is therefore not easily linked. To do so requires coordinated effort across the whole programme in order to understand what data is being produced, how it can be put into context using common metadata, and how it might be relevant to other trials and experiments (see Challenges for Data Linkage).

Therefore, an important objective of our project data management strategy has been to bring all the DFW trial datasets into one data repository annotated with the necessary metadata to support both findability and interoperability. We were not able to identify an established metadata standard which adequately describes field experiment design, plot layouts and related spatial information and so we have established our own and incorporated it into our central data management tools. Our data ingestion framework allows field trial data managers to upload their designs and link to trait measurements. This has also resulted in identification of new metadata terms that are required for describing elements of these trials but are not part of the commonly used ontologies (see "Ontology richness and standardisation of trait names").

2.3 Harmonising Environmental and Management Data

Agricultural research is often conducted to evaluate how the interactions between the crop genotype (G), environmental (E) and management (M) factors influence the behaviour of a particular trait (e.g. crop yield or flowering time), often called GxExM studies. Alternatively, a trial may evaluate the response of a particular trait to a treatment (e.g. amount of fertiliser applied) or change in management (e.g. use growth suppressors). The difference between a management or treatment effect depends on the nature of the research question. Management and treatments are aspects of a trial that can be controlled. Environmental factors, on the other hand, are those aspects of a trial that can't be controlled, such as levels of pest infestation, rainfall, temperature or level of sunlight.

When planning field trials and associated measurement protocols, the key decisions will be the selection of the genotypes (varieties), standardisation of trait measurements and any treatments, but also what environmental monitoring and management measures will be used in the trial. In the DFW project, we are working on a single crop (wheat) and our consortium has collaborated for many years. The wheat community has also been developing shared standards for trait measurements for some years (e.g. Dzale Yeumo et al., 2017). Therefore, in many ways we have less of a challenge in this area than if were we building a data sharing infrastructure for multiple crops.

In the case of our project trials, most of the harmonisation of data and methods have been agreed in the trial design stage. Where work is still ongoing is the collection of the associated environment data and management data alongside the core trait measurements and the management information in common forms so that the datasets we hold are complete and satisfy the requirements of re-usability.

2.4 Ontology Richness and Standardisation of Trait Names

Consistent descriptions of experimental processes and measurements of observed entities is vital to ensure cross-compatibility of datasets. The generation of agricultural data comes in many forms and from a wide range of instruments and methods (see Table 1 "*Complexities of data types based on their collection and analysis profiles*" for example). Without some form of internal consensus on standards for data description and sharing, attempting to compare results across high variable spaces such as these is at worst impossible and at best will require a great deal of manual curation. The description of these datasets is dependent on the availability of ontologies that comprehensively cover the domain, based on the richness of the terms within them.

An example of this is how trait measurements are recorded across sites (Pérez-Harguindeguy et al., 2013). Manual scoring of traits within field or greenhouse settings is commonplace and a route to understanding growth, development, and heredity often based on external factors. Traits tend to be semi-descriptive, where a measurement will be inferred but specifics of how that measurement was made are not, e.g. "plant height" is a common trait to measure, but says nothing of the process, constraints, or units when measuring. Therefore, we have produced a trait measurement catalogue which brings together the project-level consensus on terms used when measuring traits, based on the Crop Ontology (Shrestha et al., 2012) http://www.cropontology.org/ontology/CO_321/Wheat. Where traits and measurements are not harmonised with user experience, we feed back these potential additions and improvements to the ontology as appropriate.

2.5 Data Curation Tools and Techniques

Producing FAIR data requires active management through curation. Tools and services that help users meet FAIR data requirements without needing extensive manual intervention are often lacking. When they are provided, they often are geared towards data managers and curators themselves rather than tackling the issue of solving standardisation of datasets at the point where they are generated by a researcher.

Within DFW, using our trait measurement catalogue, users who need to record and submit trait measurements into our centralised data services can use a dedicated

Table 1 Complexities of data types based on their collection and analysis profiles

Data class/origin	Characterisation	Challenge/progress
HTP and UAV imaging	High volume, relatively low complexity	Wide range of different sensors/ instruments Useful data comes only after bespoke data analysis pipelines – often research projects. Provenance tracking in analysis pipelines unsolved problem
Pangenomic datasets Wheat 10+ www.10 wheatgenomes. com	Medium to high volume, high complexity Problems with *de novo* vs "lift over" annotations etc	Large polyploid genome sizes. Data processing pipelines compute intensive and in research projects. Visualisation is challenging, but tools are beginning to emerge.
Single cell genomics	High volume, high complexity	Metadata capture and processing across 1000s–100000s cells is complex. Data processing pipelines can be compute intensive, and software development is still required within research projects
Field trial datasets	Low volume, medium complexity "Integrative" already – images, traits, geolocation, unstructured metadata, low standardisation across sites	Solutions vary across sites and with different crops. Community awareness of the importance of standardisation is patchy. Trial / experimental design metadata has no accepted standard.
Diversity set genotyping	Medium volume – and complexity is a factor of the SNP density	Generally well developed standards and data pipelines exist in crop genetics labs.
Epigenetic datasets	Medium to high, high complexity (interpretation) Often need other datasets to contextualise	Similar to crop pan-genome datasets. Data pipelines are research projects in progress.
LTP and "physical" phenotyping -architectural traits	Low throughput, low complexity Manual technologies lead to human-centric problems	These are the traditional datasets used in crop genetics and plant breeding. Range of standards in place (e.g. for breeders) and some good community agreements in place. In general, problems relate to formalisation of metadata and data quality issues at point of recording. Portable devices offer big improvements in field recording.
Chemical phenotyping	Low/medium volume, low complexity Quantitative bulked datasets (typically plot based)	Complexity will depend on the compositional analysis required. Issues will be about tying together analyses from multiple instrumentation used

web service to search for consistent terminology *and* measurement properties to use in their collection strategy. For example, users can select a number of traits they wish to measure, and our system will produce an organised spreadsheet that they can fill out and then directly submit into our data repository. This reduces: the need for user support in terms of ontology use; incompatibility of generated datasets; friction experienced when trying to reformat datasets for deposition in repositories.

3 Challenges for Data Linkage

A central challenge of large projects such as DFW is the feasibility of real world management and coordination across large, varied data generation technologies. Researchers are increasingly reliant on a greater number of more varied datasets housed in multiple locations, and the integration of field data with large genomic and phenotypic datasets is needed to push forward the understanding of the relationship between genotype, phenotype and the environment. Approaches need to involve understanding the characterisation of these datasets into estimates of their volume and complexity which, when coupled to the availability of standards and curation tools, often forms the main barriers to data sharing and integration.

There are experimental technologies, e.g. high content phenotype data (phenomics) using state of the art imaging technologies and UAVs, that are being applied to the same germplasm used in the rest of the project for genomics studies. Integration of metadata-rich field trial, trait measurement, imaging and genotyping will offer new ways of predicting GxExM relationships using automated methods such as statistical inference, machine vision and AI.

Additionally, different strategies are needed for integration of each class of data which adds to complications and a large time cost when attempting to standardise – sometimes the plot is the reference unit, sometimes gene name, sometimes genotype, sometimes sample from plot. There are a wide range of different "primary" keys needed and the challenge is getting standardisations and adherence to naming and identifying schemes. This is especially important when retro-fitting technology to an ongoing project that is being run by multiple organisations and may have large legacy datasets.

There are other challenges when developing data management infrastructure for large research consortia, such as DFW. These emerge from the independence of research leaders within the consortium to set their research agenda while remaining within an agreed programmatic or strategic framework. Understandably, these leaders also focus on their science specialisms among the participating research groups and within the programme. Such large consortia therefore create a landscape of research, with research groups functioning as islands of outputs that contribute to the bigger picture. Providing the support to capture all outputs and making necessary connections (usually retrospectively) between them, for example a detailed finding about crop genetics and a trait brought into pre-breeding material, is not immediately possible without manual processes involving curation.

The process of collecting common data about field trials in DFW has highlighted the gap in tooling for data management of phenotyping from field experiments and the need to coordinate and share these data across the consortium to facilitate collaboration. Other data domains have recognised repositories, e.g. for wheat genome and transcriptome sequences, genetic variation (SNP) data, etc. There is also active development of plant phenotype metadata standards (Papoutsoglou et al., 2020) but as yet no dedicated publicly accessible repository. In future there will be a need to integrate other types of data that provide *data waypoints* in the landscape of crop science.

The following three research papers from DFW illustrate the variety of data and where there has been partial success (within the project) to link these data together in the DFW repositories.

In a time course study a high throughput plant phenotyping platform was used to extract plant height information (an important agronomic trait in wheat) using computer vision methods (Lyra et al., 2020). This was an investigation of new statistical genetic methods to address some of the challenges of high throughput datasets from time course studies. The data came from 197 wheat lines grown over two seasons, where 22–26 time points were measured by laser scanning, and plant height was extracted from the subsequent point-cloud data. Statistical genetics analysis permitted identification of persistent and transient QTLs. Genotype data came from a SNP array.

In Shorinola et al. (2019) mutant lines from a wheat tilling population were sequenced. Grain size and root phenotyping methods were used to explore the genetic links between root and grain development. The data processing steps involved extraction of quantitative measures of root growth based on imaging techniques.

In a time-course study of senescence in wheat (Borrill et al., 2019) the experimental focus is on a single wheat variety (bobwhite). The plants were phenotyped for chlorophyll content and grain moisture content. RNA expression data was generated from different tissues and this was used to infer gene regulatory networks. Through integration with public datasets from *Arabidopsis* and the public wheat genome sequence, gene function annotations were transferred to propose the genetic control mechanisms that underlie senescence.

All these papers are based on material that has been grown in a field or greenhouse and demonstrate a large scale generation and reuse of data across the variety of data types and resources:

- Phenotyping (single trait)
- Omics (multiple traits/genes/transcripts)
- Genetics (multiple varieties/QTL)

Our approach in DFW to providing a higher-level view of the data landscape has been to further develop the KnetMiner system. KnetMiner integrates information extracted from public databases and literature resources as well as data services from other wheat information resources to create a comprehensive knowledge graph of wheat information (Hassani-Pak et al., 2021). It has also been possible to

interoperate with some of the large-scale DFW data services into this resource (e.g. gene co-expression networks, (Ramírez-González et al., 2018). Furthermore, the KnetMiner team is experimenting with knowledge-level integration by means of "lightweight ontologies", such as Bioschemas (Gray et al., 2017). Knetminer for wheat (knetminer.com/Triticum_aestivum/) therefore provides an open-access wheat resource to DFW participants and the wider community as presented in published papers and datasets in public repositories. However, KnetMiner provides linkage at the knowledge level, and it does not support integration and linkage down to the individual datasets and measurements from those studies.

It is increasingly clear that researchers are producing data with future reusability in mind. However, this does not implicitly make the task of data linkage easier in terms of the complexities of data types and their contexts. Within one experiment, the challenge of linkage is somewhat manageable. To interlink trials and experiments across multiple organisations is very difficult to accomplish in retrospect, even within a single coordinated programme.

4 How to Do It – Data Stewardship Strategy and Infrastructure

As we have described above, in DFW we aim to support integration and sharing of data for our multi-faceted, multi-year, cross-institute programme comprising a portfolio of different experiments, and make research outputs visible both internally and to a range of external stakeholders. To do this, we needed to implement multiple strategic layers of physical, virtual, and coordination infrastructure.

4.1 FAIR Data Sharing within DFW

When the DFW project was being developed, it was agreed that sharing data **within** the project would be essential and that the different work streams would need to coordinate efforts and share best practice. It was also important that data management and sharing was not seen as the sole responsibility of the informatics teams, but one that was shared with the plant scientists too. Coordination was needed to inform the development of the data sharing tools and ensure interoperability between the specialist bioinformatic and genomic data resources that individual partners were developing during the project.

So, the DFW Data Coordination Task Force (DCTF) was established to bring together a multidisciplinary group with representation from all parts of the project including crop breeders, UAV drone phenotyping experts, and software developers. Members were selected from across the programme's plant science teams to act as local experts on the data sharing activities. As such, DCTF members help and

encourage individual scientists to come forward with data and get the support they need to annotate it with the necessary metadata and submit to data repositories. Coordination among DCTF members has been achieved through video calls held every 2 months that focus on prioritisation of collaborative elements of the programme, including data resource development that requires input from biologists and statisticians, data generators, data analysis and visualisation experts, and information system software engineers. Regular hackathons have brought developers together with plant scientists to refine data submission and data sharing technologies and to make rapid progress on software and web site interoperability. Hackathons were initially face to face working meetings which led to continued joined-up activities for periods of time after the event. During the COVID-19 pandemic we have used collaboration tools (Microsoft Teams) to run Hackathons with no obvious reduction in their effectiveness in terms of engagement during the meeting or in followup discussions. Hackathon topics have been nominated either by DCTF members or through wider calls for topics from project work-package leaders. More recently, a hackathon to discuss the potential implication of new wheat pangenome data sets (e.g. Walkowiak et al., 2020) for the project was organised with input from the broader wheat and pangenome research community.

Individually, the DCTF members and many of the wider DFW team are active participants in larger national and international communities with shared interests in wheat and wider grass and cereals research and the translation to crop improvement, e.g. the Monogram network (www.monogram.ac.uk), The International Wheat Initiative (www.wheatinitiative.org) and The Wheat Genetic Improvement Network (www.wgin.org.uk). Many of the DCTF are also members of the ELIXIR plant sciences bioinformatics and data infrastructure community (http://elixir-europe.org/communities/plant-sciences). This allows skills, expertise and knowledge to be shared more effectively through a formalised network of peers with a clear remit.

4.2 Compute, Storage, People, Skills

Research Data Management (RDM) lifecycles (Higgins & Others, 2012) are not only concerned with the human-level aspects of collecting, managing, analysing and sharing results, but also the technical aspects where Research Infrastructures (RIs) now play a huge part in the modern high-throughput crop research arena. Traditional research outputs of publications, software and data are increasingly underpinned by a fabric of digital infrastructure, intrinsically woven into how RDM is carried out. UKRI recently produced a landmark report on the status and future recommendations for UK digital infrastructures across the UKRI family of Research Councils and HEIs, and part of this report was concerned with this "ferrying" of data in and out of life science RIs – termed "data stewardship".

As part of DFW, our data stewardship strategy is a formal part of our programme, where we use investments made into digital infrastructure at each of the partner sites in order to facilitate the wider integration of research data through effective

coordination. For example, the DFW Data Portal houses a large amount of pre-publication data and acts as a long term storage area for datasets typically not suited to public repositories. This repository comprises an infrastructure of virtual servers and data storage architectures that is provided by CyVerse UK cloud, running within the EI National Capability for e-Infrastructure (Earlham Institute, 2018).

The underpinning infrastructure is invisible to end users via a typical "as-a-Service" architecture (www.intel.co.uk/content/www/uk/en/cloud-computing/as-a-service.html), providing data access APIs, hosted websites, and analytical platforms. This enables us to rapidly develop new tools, share new data, and explore technical solutions collaboratively across the project and beyond.

4.3 Benefiting from Open Source Tools

Our approach relies on open source platforms, both in terms of our own developments but also when using off-the-shelf solutions, e.g. CKAN (ckan.org) for our published outputs, including papers and supplementary datasets. The software and data resources developed within the project (see Table 2) are heavily reliant on open source software (e.g. Neo4J, REACT, PostGreSQL), typically with free academic license agreements. This allows us to maximise the cost-effectiveness of our investments in research, and retain the ability to adapt and modify tools and methods as appropriate. A key challenge is keeping up with new technologies whilst also being able to interoperate in a backwards-compatible manner with previous resources or those developed by other groups. Open source tools give flexibility to try and adapt new ways of working with data and metadata without needing paid software, proprietary codebases or formal licensing agreements.

We did consider open source solutions that exist for technical integration of crop improvement data, e.g. BreeDBase (http://breedbase.org), Germinate (https://germinateplatform.github.io/get-germinate/; Lee et al., 2005), etc. but currently they do not implicitly provide the curation, QC, filtering and importing steps needed to help reach a suitable quality level for biological integration across a wide range of data types. They also assume a mature data and metadata specification. In a project

Table 2 List of main DFW funded data resources to date

CerealsDB	www.cerealsdb.uk.net
KnetMiner	https://knetminer.com/Triticum_aestivum
Wheat Expression Browser	www.wheat-expression.com
Wheat Germplasm Resource	www.seedstor.ac.uk
Ensembl Plants – Wheat	plants.ensembl.org/Triticum_aestivum
DFW Field Trials	grassroots.tools/dfw
DFW Data Portal	opendata.earlham.ac.uk/wheat
DFW Digital Repository	ckan.grassroots.tools

such as DFW which represents a large legacy and heterogeneous data landscape, we needed to build resources that provided a route of least resistance to harmonisation but also included APIs to remain interoperable.

4.4 FAIR Publication Strategies

An important aspect of our commitment to FAIR principles is to ensure that research publications and supplementary data sets are linked and that the data are in usable formats (e.g. not simply referenced in a published PDF). To achieve this we make use of the CKAN digital repository framework and capture all DFW publications and associated data as explicitly listed resources alongside a publication, allowing users to search for and access supplementary data, data files, code on GitHub and other outputs in one place. CKAN is in use by many research institutions, and as such has an active development and support community, and has a fully featured API for integration and programmatic access.

All the data resources and other information are linked from the main project web site (designingfuturewheat.org.uk).

4.5 Meeting Community Obligations

Large research projects have to balance project level obligations and those from stakeholders. For example, DFW reports regularly to funders, researchers, breeders, farmers, and policymakers in food security, nutrition, national farming, etc. These reports contain evidence of our data resources and their use in the community. Openness and transparency is essential to maintain effective communication in a complex landscape of commercial and academic interests. In DFW, FAIRness of our data is a keystone to contribute effectively to the pre-competitive aspects of wheat research in the UK and beyond. We aim to deliver the benefits of publicly funded wheat research with the least barriers to access as possible, from data to seeds to research outputs:

Data Availability is a project-level commitment, which has to be agreed from the beginning and necessary behaviours of all participants reinforced by the project management team. From the outset, the funders of DFW (BBSRC-UKRI) mandated that the project should release data to maximise public benefit. All other data generated in the project is also expected to make it as quickly as possible into the public domain, conforming to FAIR data sharing principles. The stakeholders in the best position to exploit data from the project are other researchers and crop breeders.

Germplasm Availability A key resource being developed by DFW are the Breeders Toolkits – pre-breeding germplasm that are evaluated by a pre-competitive community of wheat breeders associated with the project for potential use in their own

(competitive) breeding programmes. After a 2 year embargo, the breeders are also obliged to return their assessment data to the project in order to share with the wider community. The germplasm is publicly available to any user outside the consortium as well (via https://www.seedstor.ac.uk/).

Research output Availability The BBSRC-UKRI have paid close attention throughout the project to the research outputs being generated and the biannual reporting to them includes information on publications, datasets and software outputs. All members of the project are expected to report into a central (Google spreadsheet) on a range of research outputs and this sheet is also made available to BBSRC. Collecting this project-wide information in a simple and transparent way has proved to be extremely useful to show both project progress and commitment to openness. In particular it is the reference source for all research publications for the project for all partners which has provided the basis for a publications portal using CKAN. It has also had the beneficial side-effect of creating confidence in the project with the funder who can easily follow the result of their investment and use this evidence internally and with the government to demonstrate the value of a major investment in UK crop research.

The delivery of these resources does not implicitly improve data linkage in and of itself. Community obligations in this sense are related to the deployment of services and infrastructures that adhere to community standards, such as implementing the Breeding API (BrAPI) (Selby et al., 2019) on top of a data resource or the use of agreed controlled vocabularies, data formats, etc. To ensure these APIs work suitably, the data itself needs to be prepared and described adequately which is not simply a technical task, but a sociological one. This requires community acceptance for the need for standardisation and having resources to comply with agreed standards and protocols, but these efforts are typically not explicitly funded through research grants or programmes. To avoid siloed information and abandoned data warehouses, data linkage requires dedicated funding and resources to bring together both the people and the technology to deliver fit-for-purpose tools and services that demonstrate strong and useful interoperability. Future sustainability then demands openness across these decentralised but interconnected data resources.

5 Conclusions

There are other crop programmes internationally, so DFW is not unique in terms of coordinated efforts to bring about improved access to crop data, e.g. Sol Genomics, MaizeDB, Brassica Information Portal. Other national and international efforts have also been focused on wheat, e.g. Wheat Initiative WheatIS, Triticeae Toolbox (T3), CIMMYT. Indeed, DFW efforts interlink some of these existing resources, e.g. KnetMiner and T3, and uses API standards in some of its outputs to harmonise with the community such as BrAPI (Selby et al., 2019). The geographic and data-centric heterogeneity and diversity of all these resources has influenced and focused

DFW's strategy for data linkage in that we can learn from prior work and also drive the adoption of best practice and standards for FAIR data within the UK's national wheat programme to comply with the consensus of the broader community. The DFW strategy explicitly promotes both openness and FAIRness in an effort to support future wheat data access with the fewest barriers possible.

A strategy for data linkage should be set out at project start and all research outputs would be aligned to that strategy, supported by fully-featured production-level tools to manage datasets FAIRly. However, this is often infeasible due to the challenge of producing a cohesive data strategy without *a priori* oversight of all data types within a programme, and then ensuring compliance with that strategy from day one.

A "catch-22" situation arises where data managers cannot know all data outputs from the start of a project, and will not have all the required data management tools at their disposal, so cannot accurately model data linkage. Furthermore, even if these elements are prepared in advance, there is still the issue of the required manual curation of datasets to ensure that the data is modelled in such a way that biological interpretation is correctly maintained. This curation is often under-resourced compared to the generation of hypotheses, data, and publications. This situation is commonplace – we interact with many stakeholders within research grants, advisory and policy work, and strategic scientific programmes, and throughout all these scenarios we see the same points raised.

A solution would be to strictly control *all* data collection and sharing activities across all researchers, mapped to complete standards that can comprehensively represent all data types. This simply is not feasible for the majority of the data generated from the fast-moving world of omics-intensive agriculture. This also comes at a very large sociological cost, and would likely be rejected as overbearing by the community.

So, a solution of three parts remains:

- Constant ongoing supportive coordination to ensure a data management and linkage strategy is sufficient
  - We suggest the use of coordination committees to ensure inclusion of researchers across a scientific project in order to encourage data standardisation and sharing, and to address societal and technical changes in data management methods
- Proactive, potentially automated, management of well-known "standard" data types and studies
  - Whilst not at the forefront of scientific projects, formalised data management is absolutely required to ensure the automation of routine tasks, in turn promoting effective data reuse by the community at large
- Retrospective application of integrative methods to cater for the new, "known unknown", or less-well-standardised data types

- This is an exciting area for opportunities to develop inventive techniques to add structure to unstructured data, extracting information within scientific publications and datasets, thus future-proofing data for advances in linkage and analysis

Funders are increasingly motivated to look to facilitate these solutions, but key elements that we have highlighted are still lacking in the research landscape, including the adequate resourcing of sociological and technical research that would underpin their data management policies.

To summarise:

– Implementing FAIR principles within a project of the size of DFW is a significant undertaking
– There is a hidden cost of FAIR that is not often taken into consideration by funders and other stakeholders, leaving major FAIR data management tasks to *ad hoc* efforts by research staff rather than dedicated data stewards
– Openness leads to collaborations – in our DFW experience, communities are more willing to engage and share when our strategy for FAIR data is evidenced by open tools.
– Openness ensures that the project and its participants are visible as good collaborators and technical ecosystems benefit from open communication about issues, benefits, and functionality leading to better interoperability and future-proofing.
– Involvement is required across all stakeholders: field experiment managers, farm staff, greenhouse technicians, genebank managers, statisticians, experimentalists, molecular biologists, bioinformaticians, software developers, data managers, project managers and reporting coordinators, PIs, and funders. This is a *"Matrix of Responsibilities"* which takes time and effort to establish in a complex project.
– Increasing multidisciplinarity within a common strategic backdrop for standardisation leads to multiple teams working together to produce multiple datasets more effectively, and with better FAIRness.
– Planning for FAIR in a live research project has to adapt as experiments and methods evolve over time. Initiation often looks very different to milestone delivery so it is still too commonplace that FAIR considerations are only thought about at the time a project has finished and its outputs are being released.
– Transparency of project outputs has benefits across obligations to project-level and a wider stakeholder community.

Acknowledgements The authors gratefully acknowledge support from the UKRI-BBSRC Designing Future Wheat project (Rothamsted BB/P016855/1; Earlham Institute: BBS/E/T/000PR9783) as well as the strategic support funding to Rothamsted Research and the Earlham Institute.

The authors wish to thank the members of the Designing Future Wheat Data Coordination Task Force (designingfuturewheat.org.uk/dfw-data-coordination-taskforce/) for their support in the writing of this paper and for their work to make much of the progress described in this paper possible.

References

Berti-Equille, L. (2001). Integration of biological data and quality-driven source negotiation. In *Conceptual modeling—ER 2001* (pp. 256–269). Springer.

Börnigen, D., Moon, Y. S., Rahnavard, G., et al. (2015). A reproducible approach to high-throughput biological data acquisition and integration. *PeerJ, 3*, e791.

Borrill, P., Harrington, S. A., Simmonds, J., & Uauy, C. (2019). Identification of transcription factors regulating senescence in wheat through gene regulatory network modelling. *Plant Physiology, 180*, 1740–1755.

Earlham Institute (2018) *National Capability in e-Infrastructure*. https://www.earlham.ac.uk/national-capability-e-infrastructure. Accessed 23 Feb 2021

Food and Agriculture Organization of the United Nations. (2018). *Genebank standards for plant genetic resources for food and agriculture*. Food & Agriculture Org.

Gray, A. J., Goble, C., & Jimenez, R. C. (2017). *The bioschemas community (2017) bioschemas: From potato salad to protein annotation*. ISWC 2017 Poster Proceedings.

Hassani-Pak, K., Singh, A., Brandizi, M., et al. (2021). KnetMiner: A comprehensive approach for supporting evidence-based gene discovery and complex trait analysis across species. *Plant Biotechnology Journal*. https://doi.org/10.1111/pbi.13583

Higgins, S., & Others. (2012). The lifecycle of data management. *Managing research data*, 17–45.

Lapatas, V., Stefanidakis, M., Jimenez, R. C., et al. (2015). Data integration in biological research: An overview. *Journal of Biological Research, 22*, 9.

Lee, J. M., Davenport, G. F., Marshall, D., Ellis, T. H. E., Ambrose, M. J., Dicks, J., van Hintum, T. J. L., & Flavell, A. J. (2005). GERMINATE. A generic database for integrating genotypic and phenotypic information for plant genetic resource collections. *Plant Physiology, 139*(2), 619631. https://doi.org/10.1104/pp.105.065201

Lyra, D. H., Virlet, N., Sadeghi-Tehran, P., et al. (2020). Functional QTL mapping and genomic prediction of canopy height in wheat measured using a robotic field phenotyping platform. *Journal of Experimental Botany, 71*, 1885–1898.

Mungall, C. J., Torniai, C., Gkoutos, G. V., et al. (2012). Uberon, an integrative multi-species anatomy ontology. *Genome Biology, 13*, R5.

Papoutsoglou, E. A., Faria, D., Arend, D., et al. (2020). Enabling reusability of plant phenomic datasets with MIAPPE 1.1. *The New Phytologist, 227*, 260–273.

Pérez-Harguindeguy, N., Díaz, S., Garnier, E., et al. (2013). New handbook for standardised measurement of plant functional traits worldwide. *Australian Journal of Botany, 61*, 167.

Ramírez-González, R. H., Borrill, P., Lang, D., et al. (2018). The transcriptional landscape of polyploid wheat. *Science, 361*. https://doi.org/10.1126/science.aar6089

Selby, P., Abbeloos, R., Backlund, J. E., et al. (2019). BrAPI—An application programming interface for plant breeding applications. *Bioinformatics, 35*, 4147–4155.

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 Physiology, 3*, 326. https://doi.org/10.3389/fphys.2012.00326

Shorinola, O., Kaye, R., Golan, G., et al. (2019). Genetic screening for mutants with altered seminal root numbers in Hexaploid wheat using a high-throughput root phenotyping platform. *G3, 9*, 2799–2809.

Spiers, I., Goulding, J., & Arrowsmith, I. (2017). Clinical terminologies in the NHS: SNOMED CT and dm+d. *British Journal of Pharmacy, 2*. https://doi.org/10.5920/bjpharm.2017.02

Walkowiak, S., Gao, L., Monat, C., et al. (2020). Multiple wheat genomes reveal global variation in modern breeding. *Nature, 588*, 277–283.

Yeumo, E. D., Alaux, M., Arnaud, E., Aubin, S., Baumann, U., Buche, P., Cooper, L., Ćwiek-Kupczyńska, H., Davey, R. P., Fulss, R. A., Jonquet, C., Laporte, M.-A., Larmande, P., Pommier, C., Protonotarios, V., Reverte, C., Shrestha, R., Subirats, I., Venkatesan, A., Whan, A., & Quesneville, H. (2017). Developing data interoperability using standards: A wheat community use case. *F1000Research, 6*, 184314407. https://doi.org/10.12688/f1000research.12234.2

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.