Federation University ResearchOnline

https://researchonline.federation.edu.au

Copyright Notice

This is the published version of the following article:

Wills, Benjamin & Parker, Julie & Robinson, N. & Wong, Megan. (2020). Improving the FAIRness of Australia's grains research sector data.

Which has been published in final form at:

https://www.agronomyaustraliaproceedings.org/images/sampledata/2019/2019ASA_Robinson_Nathan_241.pdf

See this record in Federation ResearchOnline at:
http://researchonline.federation.edu.au/vital/access/HandleResolver/1959.17/183329

Improving the FAIRness of Australia's grains research sector data

Ben Wills[1,2], Julie Parker[1], Nathan Robinson[1], Megan Wong[1]

[1] Centre for eResearch and Digital Innovation, Federation University Australia, Ballarat, Vic, 3350, Australia. www.cerdi.edu.au
[2] Corresponding author: b.wills@federation.edu.au

Abstract

Across Australia's arable landscapes, thousands of crop trials have been conducted to improve the profitability and sustainability of Australian grain production. Although there have been significant steps to make knowledge gained from trials available to users, there is the potential to further support the development of next generation data models and knowledge products by integrating trials from disparate sources by adhering to FAIR principles of data management. That is, making data: *findable*, *accessible*, *interoperable* and *reusable*. This research explores whether Online Farm Trials increase the FAIRness of agricultural grains trial datasets through a comparison of the trial data capture and handling practices of organisations whose datasets are not discoverable through Online Farm Trials (OFT) (*N* = 50) with the FAIRness of the datasets discoverable through OFT. The findings demonstrate that OFT is helping to make the results of Australia's grains trials more FAIR to the users of trial data, and suggests a number of improvements to the FAIRness of trial datasets, foremost through the use of machine-readable metadata.

Key Words

Private and public data, vocabulary, DOI, eScience, publishing controls

Introduction

If Australia is to meet the food requirements of its own population over coming decades, as well as helping satisfy the projected 70 percent increase in global food demand by 2050, agricultural production systems must become both more productive and more environmentally sustainable (Lawrence *et al.* 2013; Linehan and Thorpe 2012). As information technology takes an increasingly important role in agricultural operations, addressing these challenges requires not only the capture of increased amounts of data, but also the integration of data from different sources to generate actionable information (Antle *et al.* 2017). Salient data for Australian grain producers comes from both on-farm and off-farm sources, with an increasing array of remote and on-farm sensing technology producing streams of data with the capacity to complement data emanating from more traditional data creators (e.g. organisations conducting grains research trials). In addition, there is a wide range of data streams and databases relating to environmental conditions including, soil, water health and climate which are often assembled and curated by government agencies.

The productivity potential of increased data capture through on-farm wireless sensor networks and telecommunications systems has been heralded for some time, and the applicability of 'smart farm' technologies continues to expand across a range of production sectors (Wark *et al.* 2007). However, the harmonisation of on-farm and off-farm data sources to underpin detailed modelling and predictive decision support applications (including the Australian grains sector) is still in its relative infancy (Antle *et al.* 2017). A significant hurdle associated with combining relevant data for use within producer-ready applications to support productivity improvements is that relevant data are often stored in multiple formats, in multiple locations, by multiple organisations, encompassing both public and private sector entities. According to Antle *et al.* (2017), the type of knowledge products needed to drive gains in sustainable food production require the combination of privately and publicly held datasets within a 'pre-competitive' space that is accessible by developers of knowledge products. In the context of the Australian grains research sector, the Online Farm Trials database makes public and private trial datasets discoverable in a pre-competitive space.
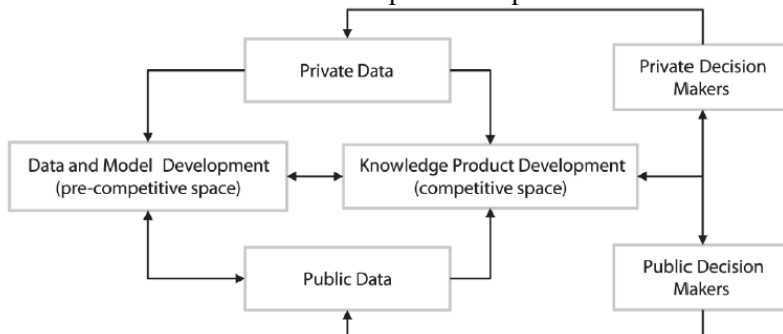

Figure 1: Linkage between pre-competitive space and knowledge product development (Antle *et al.* 2017, p. 257)

The FAIR principals of data management

Good data management facilitates discovery and innovation over a longer period of time than would be possible without it (Dallmeier-Tiessen *et al.* 2016; Wilkinson *et al.* 2016). As a result, funding bodies are increasingly specifying research data management and stewardship plans as a precondition of funding. The outcome of effective data management and stewardship is high quality digital assets that facilitate downstream discovery, access and reuse – including in a machine-to-machine capacity. The FAIR principles attempt to clarify what constitutes '*good data management*' (Wilkinson *et al.* 2016). First developed during a workshop of the Lorenz Centre in the Netherlands in 2015, the FAIR principles cover four key areas of data management: *findability*, *accessibility*, *interoperability* and *reusability*.

Table 1: FAIR principles of data management

Findable	Accessible
F1. (meta)data are assigned a globally unique and eternally persistent identifier.	A1 (meta)data are retrievable by their identifier using a standardized communications protocol.
F2. data are described with rich metadata.	A1.1 the protocol is open, free, and universally implementable.
F3. (meta)data are registered or indexed in a searchable resource.	A1.2 the protocol allows for an authentication and authorisation procedure, where necessary.
F4. metadata specify the data identifier.	A2 metadata are accessible, even when the data are no longer available.
Interoperable	Reusable
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.	R1. meta(data) have a plurality of accurate and relevant attributes.
I2. (meta)data use vocabularies that follow FAIR principles.	R1.1. (meta)data are released with a clear and accessible data usage license.
I3. (meta)data include qualified references to other (meta)data.	R1.2. (meta)data are associated with their provenance.
	R1.3. (meta)data meet domain-relevant community standards.

(Forsström *et al.* 2017, p. 2)

These principles are intended to apply not only to traditional 'data', but also to 'algorithms, tools and workflows that led to that data' (Wilkinson *et al.* 2016, p.1). FAIR has received international recognition as a framework to ensure data has maximum use and reuse, especially for machine-to-machine applications.

Online Farm Trials as a pre-competitive space.

Online Farm Trials (OFT) is a web-based resource that enables the discovery of trial research data and information from across Australia. It is a Grains Research and Development Corporation (GRDC) initiative that has been developed and delivered by the Centre for eResearch and Digital Innovation (CeRDI) at Federation University Australia. OFT currently contains more than 10,000 trials (83% of which are publicly available) from 77 different organisations. Since its inception in 2013, OFT has been developed collaboratively with input from grower groups, research organisations and domain industry experts. OFT allows individuals and organisations conducting grains trial research to store and share findings by uploading, managing and publishing past and current trial information using the self-service OFT Administration Centre. Within OFT, grains trials information from both public sector entities (state government agricultural departments) and research laboratories are combined with trials from for-profit and not-for-profit organisations, such as agronomic service providers and grower groups. To add context to this data, OFT also seamlessly integrates soil (Soil and Landscape Grid of Australia: Grundy *et al*. 2015) and climate data from the BOM and the Scientific Information for Landowners (SILO) dataset.

Users can view, analyse and export grains research data as well as compare trials based upon historical, geographic and crop-specific search filters. Growers, farm advisors, researchers and other interested parties can view and analyse the data stored in OFT at no cost, and subsequently use it to make more informed crop-management decisions. Examples of users accessing the OFT system in a pre-competitive space include:

- An advisor is interested in potential benefits of subsoil amelioration for physio-chemical constraints on a client's farm and uses OFT to determine financial merits of applying such treatments;
- A researcher is interested in identifying pre-existing trials on a topic to establish if this research has already been undertaken and is relevant to their agroecological zone of interest.

The aim of this paper is to assess how the OFT database is contributing to making Australia's grains research data more findable, accessible, interoperable and re-usable. The assessment is based upon recent data and information audits, and the open and free access governing principles of the system.

Methodology

In seeking to better understand the *findability*, *accessibility*, *interoperability* and *reusability* of Australia's grains trails research data, two primary data collection activities were undertaken: an external data audit; and an internal data audit. The external data audit (human research ethics project: 17-132A; Federation University) used an online survey tool, complimented with telephone interviews, to gather information on the data capture and handling practices of 50 Australian grains research organisations who store limited or no datasets on OFT. The internal data audit examined the characteristics of the datasets held within the OFT database. Results of these audits were considered within a semi-quantitative context to the FAIR principles to determine what, if any, impact OFT is having on making Australian grains trials research more FAIR.

Results
Findable
OFT increases trial dataset findability through a domain-specific repository with a web address (URL), with users discovering trial datasets via a searchable user interface. Trial contributors to OFT are required to complete five compulsory metadata fields to create a 'trial project' entry, supporting users to find trials of interest within the system. In contrast, 85% of respondents to the external data audit reported storing trial datasets and reports on internal organisational computer systems. OFT is not a complete catalogue of Australian grains trials with some organisations declining to add data to the system due to reasons including commercial sensitivity and concerns about the time (resourcing) required to enter data.
Few external trial organisations reported assigning durable unique identifiers to their trial datasets, although 15% of respondents (primarily those from the university sector and government departments) reported that they published their trial results in peer-reviewed journals that assign a Digital Object Identifier (DOI). Whilst in OFT each trial dataset is assigned a unique 'trial ID', findability of each trial dataset would be increased with a globally unique and persistent DOI. OFT increases the findability of trial datasets via searchable basic dataset metadata contributed by the trial contributors. OFT is also findable through repositories including Research Data Australia, however, discoverability of each individual dataset would be enhanced through employing a standard machine-readable metadata schema with DOI.

Accessible
OFT (www.farmtrials.com.au) is a free, open access resource that makes extensive use of open source software. Anyone can search for and obtain information from OFT, although data entry and publishing controls are administered by the contributing research organisation. Datasets and their metadata within OFT are fully accessible to the public, however the dataset contributor can set controls over when trial information is published, and which aspects of the trial report are publicly available. Access is via file download from a non-standard API. In these ways, OFT increases the accessibility of trial datasets since most organisations surveyed for the external data audit reported imposing restrictions on dataset access, with data being made available through individual contact with the dataset provider. Incorporation of access conditions for each trail dataset into a standardised metadata schema would further improve the accessibility of trial data in OFT. Most organisations surveyed for the external data audit reported imposing restrictions on access to data, with 30% of organisations reporting that they make trials information freely available without restrictions. Common restrictions include limiting access to fee paying members only, or, negotiating access on a case-by-case basis. A small number of respondents reported that their trials were available via interoperable web services (8%), but the majority reported access was governed by direct negotiation (40%).

Interoperable
The trial datasets available through OFT are available in an open file format and structured in a machine-readable way (in .csv). This improves the potential for interoperability of trial datasets (including machine-to-machine). Whilst 8% of external respondents did report making their data available via 'interoperable web services', most respondents reported that data was stored in propriety software, potentially unstructured (e.g. Microsoft Excel, Word and .pdf). OFT enhances the meaning of the data to the end-user and increases the interoperability between each trial in OFT through the data provider applying a local OFT 'vocabulary' to the data at the data input stage (such as 'treatment type'). Utilising expert governed vocabularies with persistent global identifiers (such as permanent URIs for 'subjects/keywords' in the metadata fields) would further assist the datasets in OFT to be read and understood by machines as well as humans.

Reusable
Trial entries in OFT have been entered into the system under the terms of a uniform disclaimer and copyright statement that provide guidance on how the stored data can be reused by third parties. In this way, OFT

improves the reusability of trial datasets. The majority of organisations involved in the external data audit make trials information available via publications, trial booklets and pdf documents hosted on their individual organisational websites. Whilst this information is likely to be protected by copyright, the conditions under which third parties can reuse the information may not be stipulated and may vary between organisations where it is the responsibility of potential re-users to contact them directly and negotiate the terms of reuse. Applying a standard licence (e.g. Creative Commons) with the licence deed and DOI for citation of the dataset encoded in the metadata record in a machine-readable format, would further increase reusability of each trial dataset through OFT.

Discussion and conclusions

Our research suggests that in the absence of OFT, most grains trail research datasets would remain undiscoverable to the broader grains research community, remaining locally stored with individual research organisations in proprietary, often unstructured data formats, and accessible largely only through individual agreement with the data custodian. OFT is making Australia's grains trials research datasets significantly more *findable* and *accessible* by bringing a significant and growing number of research trials from both public and private research organisations across Australia together online. This information resource is publicly accessible via a URL and has a searchable interface using 'local' OFT vocabularies with the ability to download data in a standard format under stipulated conditions. Steps are required to ensure that metadata records for each trial dataset are machine readable and the content compliant with recognised standard schema acceptable to the agricultural research domain, e.g. www.ands.org.au/guides/metadata-working. Beyond the current scope and user needs/user cases of OFT, it is recognised that describing the data itself with machine readable vocabularies and delivering this well described data via a standardised service would greatly enhance the interoperability of trial data made available through OFT and further assist in breaking down barriers for data discovery and re-use for more deep and broadly integrative agricultural eScience. As highlighted in this paper, further improvements to findability, accessibility and re-usability of trial datasets through OFT can be enhanced with improvements to dataset metadata (for example through employing standard metadata schema and incorporating DOIs, standard licencing, provenance information and vocabularies). Improving the interoperability of the OFT data for seamless integration with other data sources presents a challenge due to the variation of trials and nature of the data collected, how the data is described, how the data is stored (for example Excel or database structures) and individual provider institutional data management policies and capacity. Increasing the FAIRness of data along the machine-readability continuum, particularly through interoperability, remains a great challenge yet presents great opportunity for OFT and the broader grains and agricultural research sector.

References

Antle JM, Basso BO, Conant RT, Godfray C, Jones JW, Herrero M, Howitt RE, Keating BA, Munoz-Carpena R, Rosenzweig C, Tittonell P and Wheele TR (2017). Towards a new generation of agricultural system models, data, and knowledge products: model design, improvement and implementation. Agric. Syst. 155: 255–250.

Dallmeier-Tiessen S, Darby R, Gitmans K, Lambert S, Matthews B, Mele S, Suhonen J and Wilson M. (2014). Enabling sharing and reuse of scientific data. New Review of Information Networking 19: 16–43.

Forsström P, Haapio H and Passera S (2017) FAIR design jam: a case study on co-creating communication about FAIR data principles. In E Schweighofer *et al*. (eds) Trends and communities of legal informatics. Proceedings of the 20th International Legal Informatics Symposium IRIS 2017. pp. 433–440. (Österreichische Computer Gesellschaft: Wien, Germany).

Grundy MJ, Rossel RAV, Searle RD, Wilson PL, Chen C and Gregory LJ (2015). Soil and Landscape Grid of Australia. Soil Research 53(8), 835-844.

Lawrence G, Richards C and Lyons K (2013). Food security in Australia in an era of neoliberalism, productivism and climate change. Journal of Rural Studies 29: 30–39.

Linehan V, Thorpe S, Andrews N, Kim Y and Beaini F (2012). Food demand to 2050: opportunities for Australian agriculture, ABARES conference Canberra, March.

Wark T, Corke P, Sikka P, Klingbeil L, Guo Y, Crossman C, Valencia P, Swain D and Bishop-Hurley G (2007). Transforming agriculture through pervasive wireless sensor networks. *IEEE Pervasive Computing*, 6(2): 50–57.

Wilkinson MD, Dumontier M, Aalbersberg IJ, *et al.* (2016). The FAIR guiding principles for scientific data management and stewardship. Scientific Data 3:160018: 1–9.