Linking Legacies: Realising the Potential of the Rothamsted Long-Term Agricultural Experiments

Richard Ostler, Nathalie Castells, Margaret Glendining, and Sarah Perryman

Abstract Long-term agricultural experiments are used to test the effects of different farm management practices on agricultural systems over time. The time-series data from these experiments is well suited to understanding factors affecting soil health and sustainable crop production and can play an important role for addressing the food security and environmental challenges facing society from climate change. The data from these experiments is unique and irreplaceable. We know from the Rothamsted experience that the datasets available are valued assets that can be used to address multiple scientific questions, and the reuse and impact of the data can be increased by making the data accessible to the wider community. However, to do this requires active data stewardship. Long-term experiments are also available as research infrastructures, meaning external researchers can generate new datasets, additional to the routine data collected for an experiment. The publication of the FAIR data principles has provided an opportunity for us to re-evaluate what active data stewardship means for realising the potential of the data from our long-term experiments. In this paper we discuss our approach to FAIR data adoption, and the challenges for refactoring and describing existing legacy data and defining meaningful linkages between datasets.

1 Introduction

Long-term agricultural experiments (LTEs) are used to test the effects of different treatments on the sustainability of crop production and resilience of soil health (Dyke, 1974; Johnston & Poulton, 2018). The data collected from LTEs provide irreplaceable time-series while physical samples provide unique snap shots in time for future research. Since the experiments operate over extended periods of time,

R. Ostler (✉) · N. Castells · M. Glendining · S. Perryman
Computational and Analytical Sciences, Rothamsted Research, Hertfordshire, UK
e-mail: richard.ostler@rothamsted.ac.uk; nathalie.castells@rothamsted.ac.uk; margaret.glendining@rothamsted.ac.uk; sarah.perryman@rothamsted.ac.uk

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_7

they are well placed to monitor soil processes and responses to climate change. Samples and data can also be subjected to new and improved analytical, statistical and data science techniques, not imaginable or available when collected, such as gene sequencing and machine learning.

Overtime, cumulative additions or omissions of treatment factors can give rise to radically different soil environments between experiment plots resulting in contrasting levels of soil pH, nutrient availability, organic matter or biological activity. Researchers can take advantage of these conditions to use a long-term experiment as a living laboratory. In-field observations and measurements can be made, and, provided they do not interfere with the continuity of the experiment design and data collection, short-term interventions, which take advantage of quantitative differences between plots, may be introduced. Alternatively soils from these plots can be sampled for use as growing media for laboratory and glasshouse experiments.

However, LTEs are costly to run, and experiment managers must demonstrate their cost effectiveness to funders as measured by their continuing impact on science and agriculture. A good return on funder investment can be achieved if an experiment can serve more than one research objective, beyond its original purpose (Johnston & Poulton, 2018), and this can be realised by re-using LTE data and integrating it with other data. For the experiment manager this means a plurality of routine and relevant crop, agronomic and environmental observations are made, detailed experiment management records kept (Dyke, 1974) and appropriate procedures for experiment and sample access are in place. For the data curator the responsibility is to ensure an experiment and its data are sufficiently well stewarded to remain findable, accessible, interoperable and re-usable over time (Wilkinson et al., 2016). However, a major challenge for LTE data curators is a lack of widely adopted standards for managing data across the global LTE community.

The data from long-term experiments should therefore have an important role in understanding agricultural systems and the challenges they face. However, realising this potential means making the data from long-term experiments findable, accessible, and re-usable to the wider scientific community. This potential can be magnified by making the data both machine actionable and interoperable with data from other experiments, both short and long-term, and other datasets. Only then can the data from these unique experiments be used to help address challenges of food security, soil health and climate change adaptation.

This paper uses the Rothamsted Long-term Experiments and Electronic Rothamsted Archive (e-RA) to examine some of the challenges facing experiment data curators who manage LTE data and how publication of the FAIR Data Principles (Wilkinson et al., 2016) has stimulated a reappraisal of LTE data provision using e-RA.

2 Long-Term Experiments at Rothamsted

Between 1843 and 1856 Sir John Bennet Lawes and his scientific collaborator Joseph Henry Gilbert established a series of experiments, including Broadbalk and Park Grass, at the Rothamsted Estate, Hertfordshire, UK (2019). The experiments aimed to test the effects of different organic and inorganic fertilisers on yields for a range of cereal and root crops and hay. Of the nine experiments established between 1843 and 1856, known as the "Classicals", five are still running; these are Broadbalk Wheat, Park Grass Hay, Hoosfield Barley, Exhaustion Land and Garden Clover. Since then at least 40 other long-term experiments, running for at least 10 years, have tested a diverse range of treatment factors including crop rotations, cultivation, manuring, pest and disease control and liming.

2.1 Data and Samples: Lawes and Gilbert's Enduring Legacy

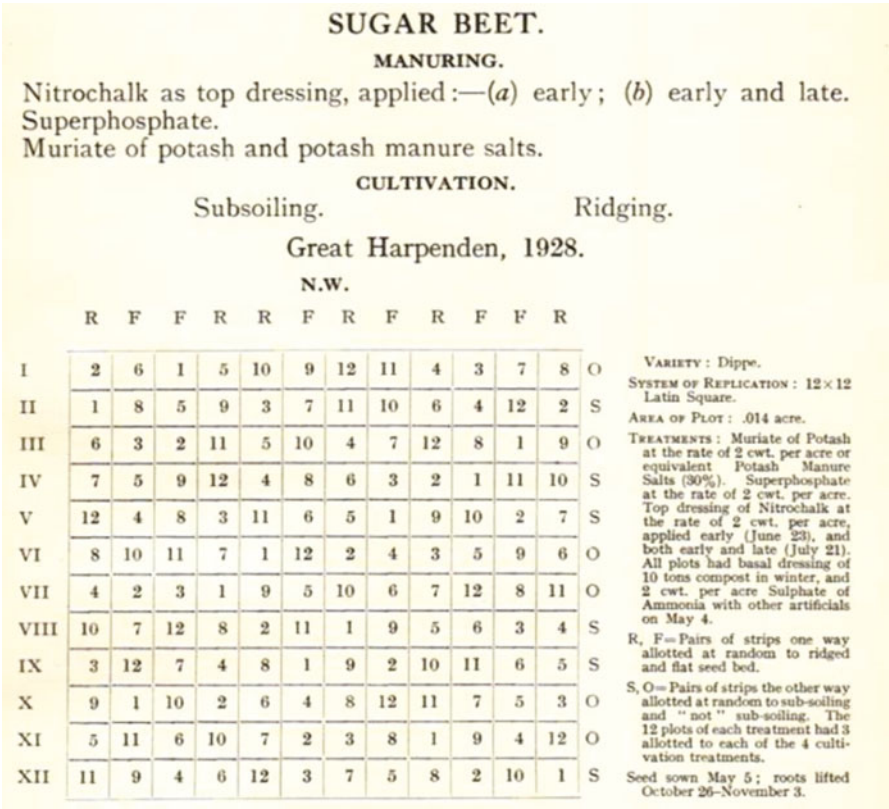
Early on, Lawes and Gilbert recognised the potential value of the data archives and physical sample collections for future scientists. In 1864 Lawes noted the rapid progress in soil science during his own time and speculated what further knowledge future progress could yield (Lawes & Gilbert, 1864). Their foresight in establishing a tradition of data collection and sample archiving has continued to benefit generations of scientists to the present day.

Lawes and Gilbert didn't just keep records and file them away. From as early as 1862 data and results were published, initially as 'Memoranda' (1862) and later as 'Supplements' then 'Yield books'. By 1927 (1928) the Supplements were publishing data and results alongside experiment documentation, recognisable today as structured metadata (Fig. 1), including experiment design; treatment factors and factor levels; plot plans; plot areas; crop varieties; and agronomic events. Later Yield Books added objectives; notes; investigator names; previous cropping; and plot dimensions.

The knowledge about an LTE also extends into an experiment narrative which describes its history in terms of the events and decisions that have shaped an experiment from inception through to either its termination or the present. This narrative provides crucial information for researchers which is often critical to appropriately interpret and re-use experiment data over time.

2.2 The Long-Term Experiments National Capability

Today the ongoing importance of the Rothamsted Long-term Experiments is recognised and funded by a Biotechnology and Biological Sciences Research Council National Capability Grant. The grant objective is to provide long-term



SUGAR BEET.

MANURING.

Nitrochalk as top dressing, applied:—(*a*) early; (*b*) early and late.
Superphosphate.
Muriate of potash and potash manure salts.

CULTIVATION.

Subsoiling. Ridging.

Great Harpenden, 1928.

N.W.

	R	F	F	R	R	F	R	F	R	F	F	R	
I	2	6	1	5	10	9	12	11	4	3	7	8	O
II	1	8	5	9	3	7	11	10	6	4	12	2	S
III	6	3	2	11	5	10	4	7	12	8	1	9	O
IV	7	5	9	12	4	8	6	3	2	1	11	10	S
V	12	4	8	3	11	6	5	1	9	10	2	7	S
VI	8	10	11	7	1	12	2	4	3	5	9	6	O
VII	4	2	3	1	9	5	10	6	7	12	8	11	O
VIII	10	7	12	8	2	11	1	9	5	6	3	4	S
IX	3	12	7	4	8	1	9	2	10	11	6	5	S
X	9	1	10	2	6	4	8	12	11	7	5	3	O
XI	5	11	6	10	7	2	3	8	1	9	4	12	O
XII	11	9	4	6	12	3	7	5	8	2	10	1	S

VARIETY: Dippe.

SYSTEM OF REPLICATION: 12×12 Latin Square.

AREA OF PLOT: .014 acre.

TREATMENTS: Muriate of Potash at the rate of 2 cwt. per acre or equivalent Potash Manure Salts (30%). Superphosphate at the rate of 2 cwt. per acre. Top dressing of Nitrochalk at the rate of 2 cwt. per acre, applied early (June 23), and both early and late (July 21). All plots had basal dressing of 10 tons compost in winter, and 2 cwt. per acre Sulphate of Ammonia with other artificials on May 4.

R, F = Pairs of strips one way allotted at random to ridged and flat seed bed.

S, O = Pairs of strips the other way allotted at random to sub-soiling and "not" sub-soiling. The 12 plots of each treatment had 3 allotted to each of the 4 cultivation treatments.

Seed sown May 5; roots lifted October 26–November 3.

Fig. 1 An early example of structured field experiment metadata published in the Rothamsted Reports, 1927–1928

experiments and their data for the benefit of the UK bioscience community and to facilitate internationally excellent research in the field of food security and agroecological research. The experiment plots, data and samples are available for scientific research and the grant supports this by funding ongoing maintenance of the experiments, routine data collection and data stewardship, the Rothamsted Electronic Archive (e-RA) and Rothamsted Sample Archive (2019).

2.2.1 e-RA Data Curators

The primary roles of the e-RA Data Curators are data stewardship, data provision and servicing data requests. Data provision includes both appending new experiment data to existing datasets, preserving the backlog of legacy LTE data into accessible research data and maintaining metadata and supplementary documents.

Servicing data requests typically involves advising researchers on the appropriate plots and treatments to use for their research question. It may also include working up new or bespoke datasets for a request. The curators can also help foster new collaborations by match-making researchers with Rothamsted scientists having similar interests. For requests where the curators provide significant support they would expect to be included as co-authors in any publications, otherwise standard acknowledgement text, rather than a citation would be provided plus notification for any publications using the data requested.

A further role of the Curators is to collate impact metrics for reporting to BBSRC. This includes information about data requests, website traffic and data downloads. The information collected for a data access request is detailed, but necessary to show which sectors are using the data, and how and where the LTE National Capability is supporting other BBSRC research.

2.2.2 The Electronic Rothamsted Archive and Data Provision 2013–2020

Development of the electronic Rothamsted Archive, commonly known as e-RA, started in 1991 and its evolution to the launch of a public website in 2013 (referred to as e-RA 2013) is documented by Perryman et al. (2018). In 2021 a new version of e-RA (e-RA 2021) was released, and the changes made are discussed in the following sections.

e-RA provides detailed information about the long-term experiments and meteorological stations, and either direct or request access to LTE and weather data. The site also hosts the Rothamsted Document Archive (e-RADoc) http://www.era.rothamsted.ac.uk/eradoc/, which contains scanned copies of historical documents including the Memoranda, Annual Reports, Yield Books, Guides, Farm Maps and Experiment Plans, making these printed resources available online.

The LTE Data are stored securely in the e-RA database, implemented in Microsoft SQL Server. Prior to e-RA 2021, researchers accessed datasets held in the e-RA database by submitting a data access request agreement, stating the scientific basis for the request and datasets required, to the e-RA Data Curators. The researcher would either be given password access to requested datasets via the online Data Extraction Tool, renamed e-RA Data for e-RA 2021, or the e-RA curators would compile a bespoke dataset. While useful for experienced users, e-RA Data has limited functionality, allowing users to filter, sort, and download subsets of data from defined tabular datasets. It does not allow dataset (table) joins and downloaded data are provided without accompanying metadata such as an identifier, query parameters, experiment name, plot treatment details or column definitions; bespoke datasets extracted by the e-RA curators would be provided with supplementary documentation, but not published with a DOI.

This data access process allowed collection of usage data for impact reporting to funders and to control the release of data as a safeguard against misinterpretation or misrepresentation of the experiments.

Since 2016 aggregated 'Open Access' datasets have been freely available for download. These datasets are published with a DOI and accompanying metadata following the DataCite Schema (Group, 2019) recommendations. Unlike e-RA Data which provides access to annual plot data, the Open Access datasets provide an overview of key findings and trends or changes in the data and are typically averaged over several years or plots. For example the Broadbalk Mean long-term winter grain yields dataset (2017) uses 10 year means for selected plots to illustrate differences between fertilizer treatments and cropping system alongside the introductions of new agricultural technologies.

2.2.3 Data Reuse and Impact

The Long-term experiments are a well-used resource. In the first 5 years of e-RA 2013's public launch, there were approximately 400 requests for long-term experiment data and between 2011 and 2020 an average of 24 publications per year (updated from (Perryman et al., 2018)).

2.2.4 The Rothamsted Sample Archive

The Rothamsted Sample Archive holds over 300,000 soil, grain, herbage, fertiliser and organic manure samples from the long-term experiments, dating back to 1843. The samples are a unique resource freely available to scientists across the globe, and around 15–20 requests are received annually. The Sample Archive has been used to investigate diverse subjects ranging from the effects atmospheric pollution on agriculture (Fan et al., 2008) to evolutionary trends in pesticide resistance (Hawkins et al., 2014) and wheat grain quality traits (Mariem et al., 2020). The Sample Archive is not currently searchable online.

2.3 Sources of Long-Term Experiment Data

Long-term experiment datasets are created and/or added to in one of three ways:

1. Routine data creation by the LTE National Capability
2. Non-routine data creation by researchers external to the LTE National Capability
3. Digital preservation of legacy data by the e-RA Data Curators

2.3.1 Routine Data

The long-term experiments collect routine data for yields and yield traits, management data, soil chemistry and botanical (weed) diversity. Data management is a mature process with data collection and ingestion workflows, analytical methods, and quality assurance documented by internal standard operating procedures (Perryman et al., 2018). However, there is scope for modernisation to better reflect new practices, for example, creating data that is 'born FAIR' rather than making FAIR at a later stage.

2.3.2 Non-routine Data

The long-term experiments and sample archive can be used as a living laboratory resource by researchers external to the National Capability. This provides opportunities for new data creation and together these externally created datasets represent a highly heterogeneous collection, ranging from tabular observations to imagery and sequence data. Non-routine data are generated via three routes including:

1. In field observations and surveys using either manual assessments or sensor technologies (Edwards & Lofty, 1982; Morris, 1992).
2. Soil and vegetation laboratory analyses, using either archived samples or fresh samples collected from experiment plots (Hawkins et al., 2014).
3. Using soil collected from experiment plots as a growing medium for pot and laboratory experiments which are analysed to generate additional datasets (Neal et al., 2020).

Unlike routine data collection, which adds to the LTE time series, non-routine data collection events normally cover a subset of plots and treatment factors for a limited time and are not required to follow prescribed data collection methods.

Before accessing either an experiment or sample archive, researchers are required to submit a scientific justification, however, there is no requirement to provide a data management plan to demonstrate how the data will be collected, managed, and published.

2.3.3 Legacy Data

Rothamsted has conducted many Long-term Experiments over the decades, but much of the data collected is inaccessible or in need of preservation. An ongoing task for the Curators is to mobilise these data, however, this can be a slow process requiring data transcription, and, finding and checking source documents such as experiment plans. If legacy data are not being requested, the potential value of the data may be unclear, and the effort required to recover it difficult to justify.

2.4 Complementary Data: Environmental Monitoring Activities

Environmental monitoring at Rothamsted began in 1853 when Lawes and Gilbert started recording meteorological observations to better understand variations in yield due to weather. Since then the variety, velocity, and volume of additional environmental variables available has been extended through technological innovations and participation in long-term environmental monitoring networks. Together these provide important complementary datasets.

Rothamsted is part of the UK Environmental Change Network (http://www.ecn.ac.uk/) which records biodiversity data and atmospheric, water and soil chemistry data, and the UK Cosmic-ray Soil Moisture Monitoring Network (UK-COSMOS) (https://cosmos.ceh.ac.uk/).

In 1964 Rothamsted established the Rothamsted Insect Survey (https://insectsurvey.com/), a national network of light traps and later suction traps, for recording moth and aphid distributions. Light traps operate at all four sites and suction traps at Rothamsted and Brooms Barn.

In 2019 permanent soil moisture sensors were added to selected Park Grass plots.

3 Challenges for Long-Term Experiment Data Stewardship

Lawes and Gilbert left a remarkable data legacy but providing continuing access to reusable data remains a challenge. The previous section provided the context for LTEs and in this section we elaborate on the data challenges facing them.

When the Elliot 401 computer was introduced to Rothamsted in 1954 data management entered a new digital age. Since then the technologies and practices for managing and accessing data have evolved rapidly. Just as archive samples can be reanalysed in ways previously unimagined, today data can be published, linked, chunked, and reused, all as a machine actionable resource. But getting data to this state requires specialist data science skills and effort and, just because a computer can link data, it doesn't mean it always should. Understanding how to provide, interpret and integrate LTE data with confidence remains imperative if it is to be used to generate meaningful knowledge continuing impact through re-use.

The FAIR data principles are being widely adopted across research institutions in the agricultural sector and promoted by communities such as Elixir (https://elixir-europe.org/system/files/elixir_statement_on_fair_data_management.pdf) and outputs of the Research Data Alliance Agricultural Data Interest Group such as the Wheat Data Interoperability Group Guidelines (https://ist.blogs.inrae.fr/wdi/) and Agrisemantics Working Group 39 hints guide (Brandon Whitehead and Aubin, 2019). Devare et al. in their chapter highlight the new responsibility of data curators now extends to wider data governance, including active data stewardship to adopt these new standards and guidelines to support wide access and re-use.

Adopting the FAIR principles is a challenge which cannot be ignored, and, in the case of LTEs, raises multiple issues. There are issues of choice and agency ranging from technical decisions around standards adoption to determining how far the responsibility to steward LTE data runs. LTEs can have complex histories and understanding this narrative alongside various sources of variability that affect the interpretation of data is essential. Reducing barriers to data access, interoperability and, ultimately seamless data linkage, while retaining oversight of how data is used for funder reporting is a significant challenge. Further challenges exist for how to understand the potential value of currently inaccessible legacy data then mobilise them and how to ensure externally generated data are retained as part of the experiment narrative alongside routine LTE data.

3.1 Navigating Experiment Narratives

There is a long-held fear, founded in experience (Stroud, 2018), of data misinterpretation and misrepresentation which stems from a view that LTEs are inherently complex and therefore require expert interpretation.

However, while it is true the experiments are complex, this is something scientists can deal with, but only if they have the necessary information to support interpretation, so rather than using this as a reason to raise barriers, the data should be sufficiently well described to withstand and challenge deliberate cherry picking of data to present false narratives. Maintaining generational records for an LTE in terms of the experimental and agronomic decision making, methodological changes and external events that impact interpretation and uses of LTE data forms the LTE narrative. This narrative is not only crucial for using the data, but also provides consistency by giving Curators and researchers a reference point, and ensures knowledge is not lost as LTE staff move on.

For example, Macholdt et al. (2020) in an analysis of yield stability on Broadbalk, used this experiment narrative to explain why certain plot years data are excluded from the analysis. A further example is changes in phosphorous applications on the Broadbalk Wheat and Hoosfield Barley Experiments where phosphorous has been withheld on selected plots since 2001 and 2003 respectively, but for different reasons. On Broadbalk, phosphorous is non-limiting and being withheld as a management decision to allow plots to reduce to more agronomically realistic levels when phosphorous applications will resume. By contrast on Hoosfield Barley, phosphorous is being withheld as a treatment decision to study residual effects on yield. In the Broadbalk case withholding phosphorous should not impact the continuity of data over time, but in the Hoosfield LTE, a new boundary condition is being introduced that does affect this continuity and how data can be analysed.

Since 1906 experiment narratives have been published in a series of 12 Guides, updated on an irregular basis, the most recently published update in 2019 (2019). e-RA 2013 consisted of publicly accessible HTML files with free text descriptions and supplementary files provided for each LTE.

e-RA 2021 has improved the earlier version by adopting the Global Long-term Agricultural Experiments Network (GLTEN) schema to replace free text with structured and consistent experiment descriptions. Launched in 2018, the GLTEN is a community of LTE researchers which aims to improve the visibility and use of these experiments. An early output was the GLTEN schema (https://github.com/GLTEN) which describes LTEs across six themes (Box 1,) using a semantically rich and structured format.

Box 1: Metadata Themes Captured by the GLTEN Schema

1. Experiment objectives
2. Experiment design: experimental factors, factor levels and factor level combinations; plot layouts; replication; cropping system and crop rotations.
3. Administration: ownership; management; contacts; site access; sample access; data access; funding.
4. Environmental characterisation: geo-location; elevation, slope and aspect; climate; baseline and manipulated soil properties; landscape.
5. Routine data collection.
6. Research outputs: datasets; publications; supporting documentation.

An intentional property of the GLTEN schema is the Experiment Design Period which supports capturing narrative knowledge in a more structured way. Design Periods are temporally bounded and capture significant changes or transition points for an experiment, including changes to objectives, experiment factors, design, methods, management or cropping. Within a design period all properties including experiment factors and cropping can be temporally bounded to denote minor changes. In the earlier Hoosfield Barley example, the decision to withhold phosphorous to study residual effects would mark the start of a new design period.

Despite the detail provided in the GLTEN schema, it isn't comprehensive, so to plug these gaps the schema provides a structure for referencing outputs.

3.2 Sources of Variability

LTEs operate over extended periods of time and so are subject to changes. For example Glendining and Poulton discuss the problems associated with changes to sampling protocols and analysis methods for interpreting soil organic matter (Glendining & Poulton, 1996). For the Park Grass Hay experiment, in 1960 a change to the harvest method was introduced which resulted in increased dry matter yields caused by fewer yield losses during harvest. Consequently, reported yields before and after 1960 are not directly comparable. To address this a conversion factor for post 1960 yield data has been determined to allow comparisons.

The role of the Curators is to understand and manage these changes to maintain consistency and comparability of routine data over time. The experiment narrative aims to capture and explain these changes as an aid to using the data, however this typically only extends to methodological change and extreme events.

Environmental variability over time is more difficult to describe and may only be understood through analysis of the data and raises the importance of linking LTE data to complimentary covariate observations. For example, using the Rothamsted temperature record the mean air temperature is known to be 1.1 °C higher than the 1878–1987 mean with the 10 warmest years on record occurring in the last 17 years and increases greatest in the autumn and winter months, and in night time temperatures (Perryman et al., 2020). Recent work using the Broadbalk and Hoosfield LTEs has demonstrated the importance of including weather temporal variation for crop yields (Addy et al., 2020).

3.3 Adopting FAIR

The FAIR principles provide a benchmark for assessing LTE data stewardship. e-RA 2013 data provision, when measured against FAIR only partially satisfied some of the principles, and important areas could be identified where the principles are not met (Table 1).

e-RA 2021, provides access to a new class of curated dataset, referred to as LTE Standard datasets which are developed following FAIR data principles and feature data standards use, data packaging, DOI assignment and a simplified dataset registration process.

3.3.1 Standard Long-Term Experiment Datasets

The LTE Standard datasets are intended to provide comprehensive and usable datasets as an alternative to the summary Open Access datasets and e-RA Data. The Open Access datasets, while providing a useful overview, present data at a coarse scale with limited utility for research. e-RA Data by contrast provides data at the resolution of plot years, but while this resolution is clearly more useful for research, the architecture of e-RA Data means datasets are provided without an identifier or context making it uncitable and difficult to interpret. This is of course a problem for any dataset, but in the case of long-term experiments the problem is exacerbated by the often complex experiment narrative.

The LTE Standard datasets aim to address the limitations of coarse open access data and non-FAIR compliant e-RA Data downloads by re-packaging the data in the e-RA's SQL Server database (which Open Access datasets summarise and e-RA Data queries) both in line with FAIR data principles and by excluding certain data.

The aim of LTE Standard datasets is to provide data with metadata and supporting information to allow researchers to independently reuse the data with confidence.

Table 1 Comparison of the e-RA 2013 data provision methods against the FAIR principles

FAIR principle	DET	Open access
Findable	Downloaded datasets are not provided with a DOI (F1), therefore cannot meet F3 and are not provided with metadata (F2). Experiment descriptions are available from the website but are not explicitly related to downloaded datasets. Data provided by the e-RA Data Curators would be provided with appropriate supporting documentation. DET is not registered in a searchable resource (F4)	Datasets are provided with a DOI (F1) and include metadata (F2), the identifier describing the resource (F3) and are registered in a searchable and indexable resource (F4), namely the Rothamsted Data Repository and DataCite Search
Accessible	DET uses a query interface to parameterise a dataset for download. The query used cannot be saved nor is a snapshot of the data downloaded retained with an identifier assigned, therefore it is not possible to retrieve a DET derived dataset by an identifier (A1) and since DET derived datasets are not linked to experiment metadata, there is no explicitly identified metadata to link to (A2)	Datasets have a DOI and are therefore retrievable by their identifier over the internet (A1). The datasets are supported by a landing page which is accessible even if the data is not (A2)
Interoperable	Datasets do not use any formal knowledge representation or controlled vocabularies and do not have qualified references to other resources, therefore do not meet any of the interoperability principles	Datasets are described using the DataCite Schema (I1), but do not use relevant vocabularies (I2). The datasets may have relationships to other resources, formally defined in the DataCite Schema (I3)
Reusable	Users are required to agree to a data access policy stating the conditions of use (R1.1) before access is granted	Datasets are provided with accurate and relevant attributes including a Creative Commons licence (R1.1), and provenance (R1.2), however, they do not use community relevant standards (R1.3)

One expectation for the LTE Standard datasets is to reduce the time spent by the e-RA Curators servicing data requests and free them to spend more time mobilising additional legacy datasets and supporting other researchers to manage LTE data. Nevertheless, supporting researchers to use LTE data will likely remain a core activity.

Data Exclusion

The LTE Standard datasets aim to provide comprehensive and well described subsets of LTE data which are internally consistent over time and treatment factors. This

means some data may be excluded based on a set of four criteria (Box 2). Excluded data can still be requested, but with the caveat that it must be used with caution and may need additional support to use.

Box 2: Exclusion Criteria for Standard Datasets

C1. There is insufficient documentation to support interpretation and re-use of the data.
C2. A plot does not have continuity of treatments over time.
C3. A plot deviates from planned treatment regimes
C4. The treatments do not have relevance or comparability to other treatments.

For example, in the Broadbalk experiment, between 1987 and 1990, plots were split for a comparison of the modern wheat cultivar Brimstone and the older Squareheads Master, grown until 1967. In 2015, the spring wheat Mulika was grown as a wet autumn and winter prevented sowing of winter wheat. Following these criteria an LTE Standard wheat yield dataset would exclude the Squareheads Masters plot data and Mulika data based on criteria C4 and C3 respectively.

Data excluded from one dataset might be included in another, for example, the Squareheads Master data could be included in a separate dataset comparing it with Brimstone for the periods 1987–1990.

Fair Data Adoption

Previously, when the e-RA Curators provided data for a request it would conventionally be provided as an annotated Excel file with additional supporting documentation such as cropping and treatment plans, but without reference to existing data standards or best practices. Adoption of FAIR data principles requires a new formalisation in how the e-RA Data Curators manage and present LTE data. Key steps in the adoption of FAIR data include organising tabular data as non-proprietary CSV files containing only column headings and observations, and with notes and style formatting removed; publishing with a DOI and making full use of the DataCite metadata schema to identify related documents and research outputs; providing conditions of use, licencing and recommended citation; providing a plurality of relevant metadata; dataset enrichment using semantic annotation. Supporting and training Curators to adopt FAIR as a best practice are also required.

Dataset Formats and Packaging

Within agricultural sciences, field experiment disciplines such as plant breeding and phenotyping tend to be used by specialist of researchers using established data

Table 2 Standard dataset package contents

README.md	Readme file in Markdown. Contains information extracted from the datapackage.json and DOI metadata reformatted into Markdown
README.html	The Markdown file reformatted as browser viewable HTML
datapackage.json	Metadata following the frictionless tabular data package specification. It describes the contents of the data package and table definitions for each CSV file
CSV data files	Tabular CSV files containing the data. The observation data tables use the Tidy data format
Excel file	A single Excel file where each worksheet is matches one of the CSV data files. For each worksheet, the first row is reserved for field names, subsequent rows must contain only data relating to the defining field. Additional text annotations and the use of text and cell formatting to convey information are disallowed
Other documents	Other supporting documents

standards and community best practices. Long-term experiments by contrast have broader appeal across a range of disciplines (Perryman et al., 2018) spanning soil and environmental sciences, metagenomics, agronomy, ecology, plant pathology and even social and economic sciences.

A challenge for the Curators is to provide data in accessible formats for a diverse range of potential users. In some cases, the data format is proscribed by existing communities of practice, for example soil metagenomics data would naturally fit with deposition to an existing genomics repository using data standards for sequence data. However, other users may have differing expectations and experiences for the same types of data. For example, an LTE time series of plant trait data, a plant phenotyping scientist might reasonably expect to access and use data in the ISA-TAB format (Sansone et al., 2012) using the MIAPPE standard, while an agricultural systems modeller might expect data conforming to the ICASA standard (White et al., 2013). Maintaining a plurality of different representations of the same data would be a costly and place an unnecessary curation burden on the e-RA Data Curators.

To address this, we have adopted the Frictionless Data Package Specification (https://specs.frictionlessdata.io/data-package/) and tabular data formatting following Tidy Data principles (Wickham, 2014) to provide a structured dataset accompanied by supporting information (Table 2). Frictionless provides a simple container format for describing CSV data using a standard schema. It has good tool support including R (https://github.com/frictionlessdata/datapackage-r) and Python (https://github.com/frictionlessdata/frictionless-py) libraries and so provides a directly usable format in two languages commonly used by research scientists for data manipulation and analysis. The Tidy Data format is a readily understandable structure commonly used for analysis where columns are variables and rows are observations. In the case of an LTE, the observation is a plot year. The data are also provided in an Excel representation as this remains a popular file format for data exchange.

Semantic Annotation

The Frictionless specification includes an rdfType property which supports annotation of fields with an RDF Class. This property allows us to enrich the CSV data by adding a meaningful definition to each field using ontology concepts. Ontologies and controlled vocabularies used include Agronomy Ontology, Agrovoc, Plant Experimental Conditions Ontology (PECO), Trait Ontology, Environment Ontology and ChEBI. For example, in a dataset where nitrogen application is a treatment factor, the field defining the factor levels can be annotated with the PECO term nitrogen fertilizer exposure (http://purl.obolibrary.org/obo/PECO_0007102). If the factor levels are different forms of nitrogen, for example ammonium nitrate vs sodium nitrate, then these categorical values are further mapped to the ChEBI terms http://purl.obolibrary.org/obo/CHEBI_63038 and http://purl.obolibrary.org/obo/CHEBI_63005.

Providing this level of semantic annotation on the data improves the potential interoperability of the data and is a useful step for moving to a linked data format.

Publishing with DataCite DOIs

All datasets are published with DataCite DOIs and make maximum use of the DataCite schema. Publishing with a DOI improves the findability of the datasets and means they can be formally cited, and citations measured for impact reporting to funders. The DataCite Schema's RelatedIdentifier property is used extensively to link a dataset to related outputs including publications, other datasets and supporting documents such as experiment plans. This allows us to publish datasets in context with a plurality of relevant documentation.

A further advantage of DOIs and using the RelatedIdentifier property is it will allow us to generate a PID Graph (Aryani, 2019) by describing and uncovering relationships between datasets, experiments, supporting materials and publications (Fig. 1). However, the success of DOIs for measuring impact and in the PIDGraph depends on their adoption by researchers, dataset citation by authors and enforcement by journal editors.

Curry, in this volume, has discussed duplication issues facing gene banks which in part arise from data management practices of the time, and similar issues face LTE data from previous practices for sharing unidentified and unversioned data. We know there are older and duplicate versions of LTE datasets in circulation, so while adopting DOIs cannot remove these, going forward, their use provides a centralises the discovery of datasets and the relationships between datasets (Fig. 2).

Reducing Barriers to Access

Access to the new LTE Standard datasets requires user registration and this allows us to continue collecting data use metrics for reporting to funders. Including a

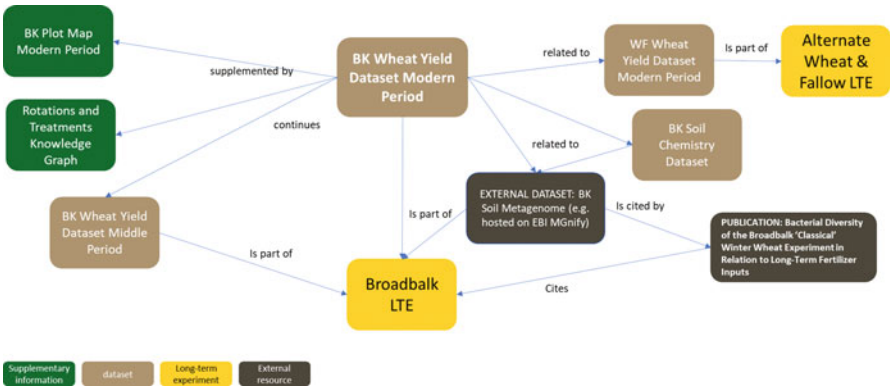


Fig. 2 PID graph showing relationships between experiments, datasets and supplementary material

registration wall is a recognised barrier to reuse (Sébastien Martin et al., 2013), and for e-RA, while there is anecdotal evidence that it has deterred requests and is viewed as archaic it is not possible to quantify this. To mitigate this potential risk the registration process has been simplified to provide a more streamlined experience compared to e-RA Data access. Users submit a one-time registration form which mandates entering a valid email address and optionally asks users to provide additional information on the intended use of the data. Users are sent a link to confirm their email and can then download the dataset.

To avoid deterring users with a lengthy form, most questions requesting information are optional. Completion of optional questions is encouraged by reminding potential users that continued funding and access to LTE data relies on our being able to demonstrate impact to funders and completing optional questions supports this.

The registration process will be monitored alongside other tracking methods such as Google Analytics and DOI metrics. If these are demonstrated to provide suitable reporting data, the need for a registration wall may be removed, however it is worth noting there are benefits from registration. By logging user emails and the datasets being accessed, subject to consent we can notify when new dataset versions and corrections are published.

3.4 Measuring Impact

Publishing datasets with a DOI should offer an attractive route to measuring the impact of LTE dataset reuse, however, there are currently cultural and technological limitations to this.

Dataset citation is not yet normalised across academic publishing; not all authors are in the habitat of citing datasets in reference sections and editors do not always enforce journal guidelines for dataset publishing. For example, a paper (Shtiliyanova

et al., 2017) published in an Elsevier journal using Rothamsted Meteorological Station data did not cite the dataset despite it being published with a DOI and editorial guidelines encouraging authors to cite datasets in the reference list (https://www.elsevier.com/journals/computers-and-electronics-in-agriculture/0168-1699/guide-for-authors).

Nor is the infrastructure to support PID Graphs mature enough to provide reliable metrics. For example, the Broadbalk mean long-term winter wheat yields dataset (2017) has been formally cited in publications at least three times, but these citations are not reflected in results from querying the DataCite Commons (https://commons.datacite.org/) and DataCite GraphQL API, giving 1 and 0 results respectively.

For reporting, a further limitation of dataset citation is it only reports examples of data reuse in the public domain, it does not reveal unpublished works. Impact from unpublished work can only be reported to funders if users of the data volunteer this information.

3.5 Preventing New Data Loss/Supporting Best Practices for Externally Generated Data

As described previously, LTEs can be used as living laboratories by external researchers. These activities can result in new externally generated datasets, but there is no current mechanism to govern how these data are managed, it relies on the researchers generating the data to follow best practices. This is a serious data stewardship issue for the long-term experiments with consequence if datasets are not well stewarded:

- The National Capability is unaware of the full extent of data pertaining to an experiment
- The e-RA Data Curators are unaware of additional datasets relevant to a data request.
- Opportunities to create new time series from repeated, but irregular samplings or observations by different groups are missed
- Opportunities for collaboration or coordination of data collection activities may be missed.
- Opportunities to increase impact through re-use for both the National Capability and the external researchers generating the data are missed
- Costly data collection activities are repeated because the fate of data generated by previous researchers is unknown.

The Rothamsted Long-term experiments are maintained as a publicly funded resource for the benefit of all and this principle should be extended to externally generated datasets, regardless of their funding and as a condition of LTE access. In future, requests for access to use the Rothamsted LTEs and sample archive should be assessed not only on their scientific merit, but also on their sampling and analytical

methods and data management plan. The LTE National Capability must have agency to prescribe actions to ensure externally generated data well managed, accessible and maintains continuity with other related LTE data. This means reviewing and agreeing DMPs to ensure there is an appropriate curation and publication pipeline, with a minimum expectation that metadata is published; recommending methodologies where these will facilitate comparability of new data with previously generated data, provided this does not conflict with the scientific question being addressed.

Enforcing these behaviours is clearly difficult, but a role of the e-RA Curators should be to liaise with the researcher to ensure the data curation terms agreed to are being followed. In cases where a researcher fails to follow the agreed DMP, the National Capability should reserve the right to deny future access to LTE resources.

In exchange for these stricter terms, the National Capability must be resourced to provide data management support and data hosting for external researchers or advise on appropriate alternative data hosting.

Adopting this as a best practice has the potential to create a virtuous circle for data linkage in which FAIRer datasets can be more easily identified and link, although it carries the risk of alienating researchers unwilling to comply with new access terms and burdening research with legal agreements.

3.6 Addressing Legacy Data

On the e-RA 2013 website the well-known Classical Experiments were extensively described with links to supporting publications and documents. For example the Broadbalk Experiment https://web.archive.org/web/20170210114304/http://www.era.rothamsted.ac.uk/Broadbalk, however, other experiments were more briefly described with only limited information, for example the Woburn Ley Arable Experiment https://web.archive.org/web/20170210235005/http://www.era.rothamsted.ac.uk/Other#SEC9.

A primary challenge for these less well-known experiments is finding information about them. Previously we stated in the 5 years following e-RA 2013's launch there were over 400 requests for data, but these headline figures hide significant variation between experiments. Over 85% of requests were for Broadbalk, Park Grass and Hoosfield, with Broadbalk accounting for nearly half of all requests. The remaining requests were spread over 27 other experiments listed on e-RA. This does not mean data from these experiments has less scientific value, rather it highlights their lower profile and lack of documentation and accessible data. For e-RA, experiment fame appears to be the determinant of the extensiveness of available data and documentation and drives a positive loop for maintaining the likelihood of re-use. Therefore, it is reasonable to assert less well-known experiments are caught in an opposite loop where a lack of data and documentation disincentivises requests and a lack of requests lowers the perceived research value of these experiments and priority for mobilising them and so they remain under-used. As can be seen in the case study (Box 3), prior lack of use is not an indicator of the research value of a legacy dataset.

To raise the profile of these experiments in e-RA 2021 we are describing them using the GLTEN Schema and publishing with the same visibility as the more famous experiments. This reduces the presentation bias between different LTEs and allows researchers to find and make a more informed decision on the potential usefulness of an experiment. Key metadata to characterise LTEs includes cropping system, climate and soil classifications, treatment factors, design and, available data. The site now indicates when data may be available but requires curation and at this point a request for data can be made. If the data are judged to be both useable (i.e. sufficient documentation exists to describe the data) and useful, the e-RA data curators will prepare a dataset for publication. Given the volume of legacy data available, this reactive strategy makes efficient use of staff time to support users and provide new, scientifically valuable datasets.

Box 3: The Long-Term Liming Experiment Case Study

In July 2016 the e-RA Data Curators received a request from the James Hutton Institute for long-term data on liming. After internal discussions the Rothamsted and Woburn Long-term Liming Experiments were suggested as potential datasets.

These experiments ran from 1962 to 1996 and e-RA provided limited information which was both difficult to find and insufficient for an external researcher to make an informed decision their usefulness. The potential value of the long-term liming experiments was only identified by a long serving staff member with deep knowledge of the long-term experiments.

The data was in poor state with the first 12 years data on paper and the remainder as early Genstat formats. Significant work was required to transcribe from and update data formats. Some data was organised by plot numbers while other data was organised by treatment, fortunately a paper key for the mapping between plots and treatments was found, otherwise a portion of the data would have been unusable.

Datasets for soil chemistry and yields were compiled and two papers investigating the effects of liming on yields and economic returns (Holland & Behrendt, 2020; Holland et al., 2019) published. The second paper on the economics of liming in arable crop rotations has since been reported on by Farmers Weekly (Clarke, 2021), demonstrating the applied agricultural interest in an experiment ended 25 years previously.

4 Conclusion

The new version of e-RA has made significant progress to make the Rothamsted Long-term Experiments a FAIRer data resource. This has required the e-RA team to look critically at how data has been managed and provided, understand where there are weaknesses and how they can be addressed. As with any change process this can

be a difficult when existing conventions that work and, anxieties about giving researchers greater freedom to use the data are challenged.

Arguments against more open sharing of data are the fears of misrepresentation or misinterpretation. Earlier we gave an example of a paper not using a data citation (Shtiliyanova et al., 2017), in fact this paper also made a false assumption about the data available. Referencing the data would have highlighted this mistake. Data citation provides a degree of confidence in that data used to assert conclusions is available for verification. The challenge for the LTE Data Curators remains to ensure LTE data are presented with sufficient metadata to support independent re-use by researchers and internally consistent to avoid accidental misinterpretations.

The experience of e-RA demonstrates long-term data stewardship needs specialist data skills and continuing investment to maintain and develop both skills and the infrastructures to support the data. Importantly this support should be extended as a service to external researchers generating data and providing it should be viewed as a matter of self-interest as additional well curated accessible and interoperable data only enhances the overall value of the experiments. Neglecting data stewardship does a disservice to the work of every technician, field worker and lab assistant who created it and future generations who can benefit from it, and as can be seen from the example of the Long-term Liming Experiments, restoring neglected data into an accessible and useable product can require significant cost and effort.

To date, most of the effort on re-thinking e-RA has focused on providing a better experience for the users. The present state of the art for e-RA are adoption of the GLTEN schema to describe long-term experiments and the move to LTE Standard Datasets which provide published datasets following FAIR principles. This is a significant advancement towards linked data; DOI metadata can identify relationships between datasets, and semantic annotation supports data linkage between equivalently described datasets. With Frictionless data there are opportunities to develop schema profiles that can better meet the needs of different user communities. However, the fundamental unit for publishing and sharing data is the dataset and LTE Standard datasets are really a convenience for grouping related observations and variables as a coherent set. The next logical step for e-RA is to provide better linkage at the observational scale, however, this will bring a new set of challenges, notably how to cite and measure impact for dynamically accessed and linked data and how to capture the experiment narrative as a set of rules to prevent invalid combinations of observations.

Compared to many research institutes managing long-term agricultural experiments, Rothamsted is now relatively advanced in adopting the FAIR Data Principles and the data stewardship approaches being used can be a template for best practices within the long-term experiments community. From the experiences of colleagues working across the GLTEN to analyse data across multiple LTEs, there are remain significant blocks to integrating and re-using data from LTEs managed by different institutes. From the initial identification of appropriate experiments and available data, understanding the methods and experiment design, and wrangling data into a usable and interoperable form, the 80/20 rule, that 80% of time is spent finding, cleansing and organising data, still applies to LTE data.

Adopting the approaches outlined here by the wider LTE community, namely adopting the GLTEN schema for robust LTE descriptions and characterisation and implementing FAIR Data Principles using the Frictionless approach may provide more opportunities for impact and in turn demonstrate evidence for continued investment to maintain these unique resources. But for now, at least having access to data stewards with intimate knowledge of an LTE remains essential for successful re-use.

References

1862. *Memoranda of the plan and results of the Rothamsted field experiments*.

1928. *Rothamsted Experimental Station report for 1927–28 with the supplement to the guide to the experimental plots*. Rothamsted Experimental Station.

2017. Broadbalk mean long-term winter wheat grain yields. In R. Research (Ed.) (1 ed.). Electronic Rothamsted Archive.

2019. *Guide to the classical and other long-term experiments, datasets and sample archive*. Rothamsted Research.

Addy, J. W. G., Ellis, R. H., Macdonald, A. J., Semenov, M. A., & Mead, A. (2020). Investigating the effects of inter-annual weather variation (1968–2016) on the functional response of cereal grain yield to applied nitrogen, using data from the Rothamsted Long-Term Experiments. *Agricultural and Forest Meteorology, 284*, 107898.

Aryani, M. F. A. A. (2019). *Introducing the PID graph* [Online]. Available: https://blog.datacite.org/introducing-the-pid-graph/. Accessed 07 Feb 2021.

Brandon Whitehead, C. C., & Aubin, S. (2019). *39 hints to facilitate the use of semantics for data on agriculture and nutrition*. https://rd-alliance.org/

Clarke, A. (2021). *Why regular liming can raise profits by up to £436/ha/year* [Online]. Available: https://www.fwi.co.uk/arable/land-preparation/soils/why-regular-liming-can-raise-profits-by-up-to-436-ha-year. Accessed 20/02/2021.

Dyke, G. V. (1974). Chapter Seven – Long-term experiments. In G. V. Dyke (Ed.), *Comparative experiments with field crops*. Butterworth-Heinemann.

Edwards, C. A., & Lofty, J. R. (1982). Nitrogenous fertilizers and earthworm populations in agricultural soils. *Soil Biology and Biochemistry, 14*, 515–521.

Fan, M.-S., Zhao, F.-J., Poulton, P. R., & Mcgrath, S. P. (2008). Historical changes in the concentrations of selenium in soil and wheat grain from the Broadbalk experiment over the last 160 years. *Science of the Total Environment, 389*, 532–538.

Glendining, M. J., & Poulton, P. R. (1996). Interpretation difficulties with long-term experiments. In D. S. Powlson, P. Smith, & J. U. Smith (Eds.), *Evaluation of soil organic matter models, 1996* (pp. 99–109). Springer.

Group, D. M. W. (2019). *DataCite metadata schema documentation for the publication and citation of research data version 4.3*.

Hawkins, N. J., Cools, H. J., Sierotzki, H., Shaw, M. W., Knogge, W., Kelly, S. L., Kelly, D. E., & Fraaije, B. A. (2014). Paralog re-emergence: A novel, historically contingent mechanism in the evolution of antimicrobial resistance. *Molecular Biology and Evolution, 31*, 1793–1802.

Holland, J. E., & Behrendt, K. (2020). The economics of liming in arable crop rotations: Analysis of the 35-year Rothamsted and Woburn liming experiments. *Soil Use and Management*, n/a.

Holland, J. E., White, P. J., Glendining, M. J., Goulding, K. W. T., & Mcgrath, S. P. (2019). Yield responses of arable crops to liming – An evaluation of relationships between yields and soil pH from a long-term liming experiment. *European Journal of Agronomy, 105*, 176–188.

Johnston, A. E., & Poulton, P. R. (2018). The importance of long-term experiments in agriculture: Their management to ensure continued crop production and soil fertility; the Rothamsted experience. *European Journal of Soil Science, 69*, 113–125.

Lawes, J. B., & Gilbert, J. H. (1864). Report of experiments on the growth of wheat for twenty years in succession on the same land. *Journal of the Royal Agricultural Society of England, 25, 10*.

Macholdt, J., Piepho, H. P., Honermeier, B., Perryman, S., Macdonald, A., & Poulton, P. (2020). The effects of cropping sequence, fertilization and straw management on the yield stability of winter wheat (1986–2017) in the Broadbalk Wheat Experiment, Rothamsted, UK. *The Journal of Agricultural Science, 158*, 65–79.

Mariem, S. B., Gámez, A. L., Larraya, L., Fuertes-Mendizabal, T., Cañameras, N., Araus, J. L., Mcgrath, S. P., Hawkesford, M. J., Murua, C. G., Gaudeul, M., Medina, L., Paton, A., Cattivelli, L., Fangmeier, A., Bunce, J., Tausz-Posch, S., Macdonald, A. J., & Aranjuelo, I. (2020). Assessing the evolution of wheat grain traits during the last 166 years using archived samples. *Scientific Reports, 10*, 21828.

Morris, M. G. (1992). Responses of Auchenorhyncha (homoptera) to fertiliser and liming treatments at Park Grass, Rothamsted. *Agriculture, Ecosystems & Environment, 41*, 263–283.

Neal, A. L., Bacq-Labreuil, A., Zhang, X., Clark, I. M., Coleman, K., Mooney, S. J., Ritz, K., & Crawford, J. W. (2020). Soil as an extended composite phenotype of the microbial metagenome. *Scientific Reports, 10*, 10649.

Perryman, S. A. M., Castells-Brooke, N. I. D., Glendining, M. J., Goulding, K. W. T., Hawkesford, M. J., Macdonald, A. J., Ostler, R. J., Poulton, P. R., Rawlings, C. J., Scott, T., & Verrier, P. J. (2018). The electronic Rothamsted Archive (e-RA), an online resource for data from the Rothamsted long-term experiments. *Scientific Data, 5*, 180072.

Perryman, S. A. M., Scott, T., & Hall, C. (2020). *Annual mean air temperature at Rothamsted 1878–2019* (2nd ed.). Electronic Rothamsted Archive: Rothamsted Research.

Sansone, S.-A., Rocca-Serra, P., Field, D., Maguire, E., Taylor, C., Hofmann, O., Fang, H., Neumann, S., Tong, W., Amaral-Zettler, L., Begley, K., Booth, T., Bougueleret, L., Burns, G., Chapman, B., Clark, T., Coleman, L.-A., Copeland, J., Das, S., et al. (2012). Toward interoperable bioscience data. *Nature Genetics, 44*, 121–126.

Sébastien Martin, M. F., Turki, S., & Ihadjadene, M. (2013). Risk analysis to overcome barriers to open data. *European Journal of e-Government, 11*, 348–359.

Shtiliyanova, A., Bellocchi, G., Borras, D., Eza, U., Martin, R., & Carrère, P. (2017). Kriging-based approach to predict missing air temperature data. *Computers and Electronics in Agriculture, 142*, 440–449.

Stroud, J. L. (2018). Co-produced data: Open access tests trust. *Nature, 562*, 344.

White, J. W., Hunt, L. A., Boote, K. J., Jones, J. W., Koo, J., Kim, S., Porter, C. H., Wilkens, P. W., & Hoogenboom, G. (2013). Integrated description of agricultural field experiments and production: The ICASA Version 2.0 data standards. *Computers and Electronics in Agriculture, 96*, 1–12.

Wickham, H. (2014). Tidy data. *Journal of Statistical Software, 1*(10), 2014.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data, 3*, 160018.