A workflow for processing global datasets: application to intercropping

Rémi Mahmoud[1], Pierre Casadebaig[1]*, Nadine Hilgert[2], Noémie Gaudio[1]

(1) AGIR, Univ. Toulouse, INRAE, Castanet-Tolosan, France
(2) MISTEA, Univ. Montpellier, INRAE, Institut Agro, Montpellier, France

(*) Corresponding author (pierre.casadebaig@inrae.fr)

ORCID iDs

Rémi Mahmoud - https://orcid.org/0000-0003-0853-0834
Pierre Casadebaig - https://orcid.org/0000-0001-7225-936X
Noémie Gaudio - https://orcid.org/0000-0002-4528-9851

Abstract

Field experiments are a key source of data and knowledge in agricultural research. An emerging practice is to compile the measurements and results of these experiments (rather than the results of publications, as in meta-analysis) into global datasets. Our aim in the present study was to provide several methodological paths related to the design of global datasets. We considered 37 field experiments as the use case for designing a global dataset and illustrated how tidying and disseminating the data are the first steps towards open science practices. We developed a method to identify complete factorial designs within global datasets using tools from graph theory. We discuss the position of global datasets in the continuum between data and knowledge, compared to other approaches such as meta-analysis. We advocate using global datasets more widely in agricultural research.

Introduction

Field experiments, whether conducted on farms or at experimental research stations, have traditionally been the primary approach for acquiring knowledge in crop sciences (Maat, 2011). Yet, extrapolating applicable principles from localized experiments remains a challenging task (Makowski et al., 2014). To derive general rules about agroecosystem functioning, meta-analysis, ~~i.e.~~ *i.e.* a “statistical analysis of a large collection of analysis results from individual studies to integrate the findings” (Glass, 1976), is typically employed. Alternatively, global datasets, corresponding to the aggregation of observations from numerous experiments, can serve as another valuable tool for analyzing agronomic data. While the use of meta-analysis to report results is growing in crop science, it is not a mainstream analysis method compared to reports based on a repeated (years) set of one or two field trials. Distinguishing themselves from meta-analyses, global datasets compile raw experimental results on a detailed scale, such as repeated measurements on individuals or multiple state variables on the canopy. In contrast, meta-analysis is typically restricted to published results with a limited set of variables.

Although examples of comprehensive agronomic datasets exist (Kattge et al., 2011; Newman and Furbank, 2021), only a few studies have been based on global datasets (Licker et al., 2010; Lobell et al., 2020; Newman and Furbank, 2021) with even less focus on methods for this type of datasets in crop science (Senft et al., 2022). One significant advantage of agronomic global datasets relies on the fact that they include diverse phenotypic observations from varying soils and climates, enabling more reliable generalization of local findings (Tardieu, 2020). These datasets reduce the risk of spurious correlations (Krajewski et al., 2015; Tardieu, 2020) and maximize the utility of experimental data yet to be used in scientific publications (Zamir, 2013).

However, global datasets come with their own challenges. Assembling these datasets requires extensive data collection, standardization, and homogenization across diverse experiments conducted by different research teams (White and Van Evert, 2008; Makowski et al., 2014). This tedious curation step is an undervalued task, whose duration could be reduced from the adoption of good practices upstream. Recent efforts and international initiatives aimed at opening and standardizing data are emerging, highlighting that data standardization is crucial for improving the interpretation of experimental results and the generalization of knowledge acquisition. It also facilitates statistical meta-analysis and data publication (Krajewski et al., 2015). However, datasets for plant and crop measurements in controlled field trials are still scarce in public databases. The different field experiments gathered often have diverse objectives, leading to unbalanced and incomplete designs. Confounding factors,

~~i.e.~~ *i.e.* the unintended mixing of two or more effects making them indistinguishable, can also be challenging (Casler, 2015). Consequently, using and analyzing global datasets require a thorough understanding of the dataset, judicious interpretation of the results, identification of balanced data subsets for specific research questions, and acceptance that the effects of some factors may remain indistinguishable. Therefore, the application of statistical learning techniques on global datasets is only feasible after extensive data pre-processing.

Despite these challenges, crop science would greatly benefit from the study of global datasets combining multiple experiments (White and Van Evert, 2008; Zamir, 2013; Cruz and Nascimento, 2019). This approach is particularly relevant considering the current agricultural landscape, where crop diversification is crucial for sustainable farming (Duru et al., 2015). This diversification mandates extensive experimentation, requiring robust data-federation efforts. The joint analysis of global datasets makes it possible to understand the context-dependent nature of diverse experiments and enhances comprehension of the interaction between crop diversity and agroecosystem functioning.

To achieve this, we recommend adopting practices for designing and analyzing global datasets that align with tidy data (Wickham, 2014; Broman and Woo, 2018) and FAIR principles (Findable, Accessible, Interoperable, and Reusable) (Wilkinson et al., 2016). As a use case, we illustrate the design of a global dataset for intercropping systems, in which at least two crop species are grown in the same field for a significant part of their growth cycle. We describe the main steps involved in designing a global dataset gathering 37 intercropping experiments across Europe. We also describe and apply an original method ~~for identifying factorial designs, which is~~ to identify complete factorial design subsets of interest. This methodological development was aimed at helping the potential collaborators to explore and get an overview of the dataset as a function of their factor of interest, a key step in assisting further modeling and analysis steps.

Our global aim was to describe our workflow in a realistic manner, hoping to promote these practices and to encourage the scientific community to move towards a more open approach to conducting experimental science in agronomy, making it more reproducible and shared.

Design steps of global datasets

This section presents the generic steps involved in designing a global dataset. As the gathering, cleaning, and formatting of the spare source datasets is time-consuming, we followed tidy data specifications (Wickham, 2014) and a global data science workflow as presented by Wickham and Grolemund (2016) (Figure 1).

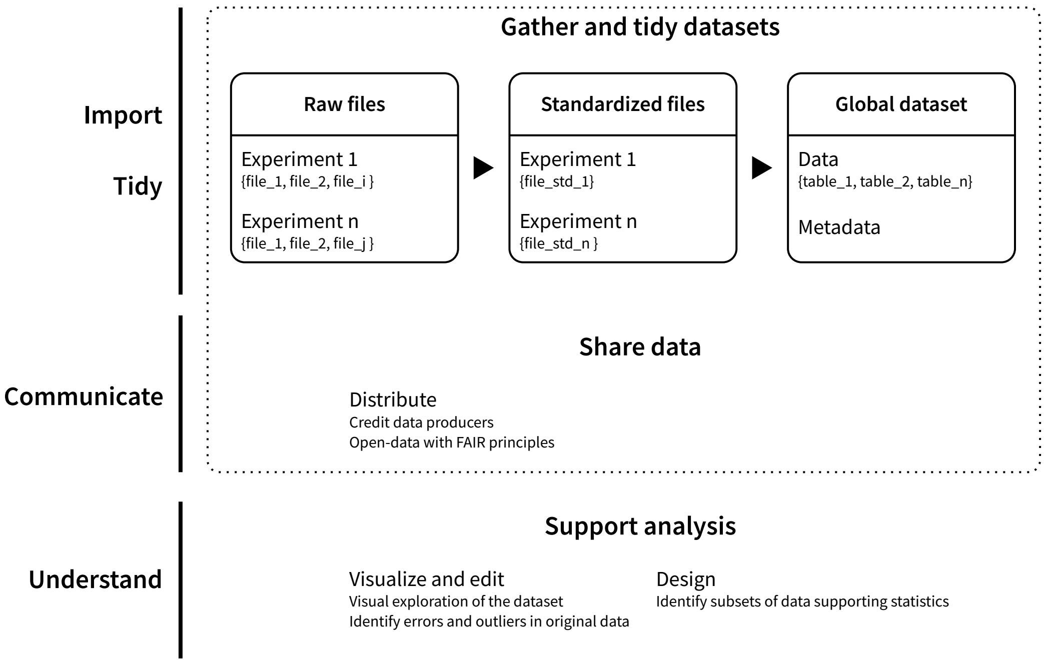

Figure 1. Main steps for designing global datasets. The left column corresponds to a classical data science workflow. We adapted these steps for global dataset design specificities, to illustrate the importance of data gathering, tidying, and sharing (dotted frame). While some actions supporting subsequent data analysis are generic (visualization, editing), most depend on the chosen analysis strategy.

1. Gather and tidy source datasets

1.1. Conceptual framework

Overall, the aim of this gathering and tidying step is to transform a highly heterogeneous set of tables, scattered in various files according to the logic of each practitioner, into a structured and documented set of rectangular files.

In a first step, the research groups that conducted the experiments whose features are interesting for a global dataset shall be identified and contacted. While the data processing step is often known to be very time-consuming in the overall data science workflow (Wickham, 2014), this contact and convincing step is also very long, with potential disappointing responses (Popkin, 2019).

Then, a basic database model for the global dataset has to be developed. This step involves defining the structure of a database, including the number of tables needed and the relationships between them. It also involves describing the metadata, such as the variables measured or collected, their definitions, and units.

Using this database model, the raw experimental files are standardized, from various spreadsheet formats into a single and coherent dataset. In crop science, operating by field experiment makes the whole process easier, by focusing standardization efforts on a set of files sharing common properties (illustrated by moving from *raw* to *standardized* files in Figure 1). These standardized files are then combined and documented to make the data "analysis-friendly" (Wilson et al., 2017), which enables detection of errors and data exploration, validation and analysis. A good practice is to work with "tidy" data which is a standard way of mapping the meaning of a dataset to its structure (Wickham, 2014). A dataset is messy or tidy depending on how rows, columns and tables are matched up with observations, variables and types. In tidy data, every column is a variable, every row is an observation, and every cell is a single value. Messy data is any other arrangement of the data (Wickham and Grolemund, 2016; Broman and Woo, 2018).

1.2. Case study

~~While there are relatively few incentives to share agronomical (Senft et al., 2022) or ecological (Jenkins et al., 2023) datasets, requirements and practices need to evolve. The ability to easily disseminates data is thus a key feature in designing a dataset, since it determines how other researchers will be able to interact with the data, and potentially increase its reuse. Open data should be designed in accordance with the FAIR data principles ().~~

~~When discussing with the involved research groups, one recurrent constraint to open their data was the perception that their contribution could not be credited unless sharing authorship in research articles. If applied consistently, open-data FAIR requirements will allow contributors to be specifically acknowledged for their work, through citation of the dataset they contributed to (Jenkins et al., 2023).~~

~~Once the data are in a tractable format, visual exploration allows for a comprehensive overview of data patterns, aiding in the identification of anomalies such as errors and outliers that may not be immediately apparent through numerical analysis alone.~~

~~Later, additional processes are required to render the dataset operational for analytical and modeling studies, such as data imputation, dimension reduction, or data normalization.~~

~~Because these steps depend largely on the chosen analytical workflow, they are not directly included in the communicated open datasets, but rather tailored by the subsequent analytical team (dotted frame in Figure 1).~~

~~Nonetheless, sharing methods can support the future reuse of the dataset. In our case in crop ecology, we illustrated this step with the development of an original method aiming at identifying subsets in the overall dataset corresponding to complete factorial designs. This method is presented in the following section.~~

~~We briefly describe the features of the available field experiments to highlight their richness and heterogeneity (see Gaudio et al. (2021) and Mahmoud et al. (2022) for full details and experimental protocols; see Gaudio et al. (2023) for the global dataset online).~~

Although combining results from a few experiments (usually two years, often sequential) is common in the intercropping literature (and more generally in crop science), no study includes joint analysis of dozens of experiments to infer more generic results about intercropping functioning. To this end, we designed, built and analyzed a global dataset gathering the results of 37 field experiments that involved cereal-legume intercrops and the corresponding sole crops. Globally, the aim of these field experiments was to compare the growth and grain yield (t.ha^{-1}) of multiple combinations of species grown in intercrop to their sole-crop reference. The field experiments were carried in 5 European countries (France, Denmark, Italy, Germany and England) from 2001 to ~~2017 (Figure 2).~~ 2017. The global dataset included 5 legume species (chickpea, faba bean, lentil, lupin and pea), 3 cereal species (barley, durum wheat and soft wheat) and 8 resulting intercrops, *i.e.* i) barley associated with faba bean, lupin or pea, ii) durum wheat associated with chickpea, faba bean or pea, and iii) soft wheat associated with lentil or pea.

~~**Figure 2.** Location of the 37 intercropping experiments gathered within the global dataset.~~

To gather the 37 experiments, six research teams were contacted. For each experiment, several ~~excel files~~ spreadsheet files (all in Excel format) were retrieved, ranging from 1 to 10 per experiment. These files differed by the number of ~~spreadsheets~~ sheets they contained, ranging from 1 to 67. We finally collected a total of 86 excel files ~~and~~ (412 ~~spreadsheets~~sheets). These raw data were highly heterogeneous at all levels, whether concerning the variables (~~e.g.~~ *e.g.* type, name, unit, measured scale) or the format of the file itself (~~e.g. one spreadsheet~~ *e.g.* one sheet per date or per variable, different tables on a same ~~spreadsheet~~sheet, calculations and graphs ~~within raw data files~~mixed with raw data cells, different languages and encoding format).

~~After the step of gathering , the files were~~ Aiming at improving machine and human readability (Wilson et al., 2017), variable names were chosen to be as explicit as possible. We settled for composite names separated by underscore and containing: as few abbreviations as possible, a reference to the organizational levels (organs: leaf, shoot; individuals: plants; population: crop), and a reference to the variable itself (biomass, number, length). After gathering step, the information of the files was transformed into standardized rectangular data tables, following ~~the tidy format and good practices~~ a *tidy* format (Wickham, 2014~~;~~) and recommended practices of data organization in spreadsheets (Broman and Woo, 2018), resulting in the creation of one given file per experiment. ~~Each file includes~~ The measured values were not normalized (for *e.g.* spatial field or experimenter effects) as the information on experimental design type and structure was only accessible in very few trials. Each file included 6 ~~spreadsheets, in which the variables and values were placed~~ sheets with one table per sheet, defined as a function of the ~~information~~ category of data they provided (~~e.g.~~ *e.g.* plant functioning, climate, agricultural practices). This step resulted in the creation of 37 excel files (vs. 86) and 222 ~~spreadsheets~~ sheets (vs. 412).

Finally, all the files were pooled together using R software, ~~with a final table per type of variable, i.e.~~ to create one global table per data category, *i.e.* four tables related respectively to climate, crop measurements, agricultural practices and global information describing the site (Figure 2). Overall, the global dataset contained 308 and 299 statistical individuals (~~i.e.~~ defined as a unique combination of {site * year * management}) in intercrop and sole crop, respectively (Table 1). The number of plant characteristics was much larger (33351 observations, among which 12896 were measured in sole crops and 20455 in intercrops), since several variables were measured at the crop scale, sometimes several times during the crop cycle.

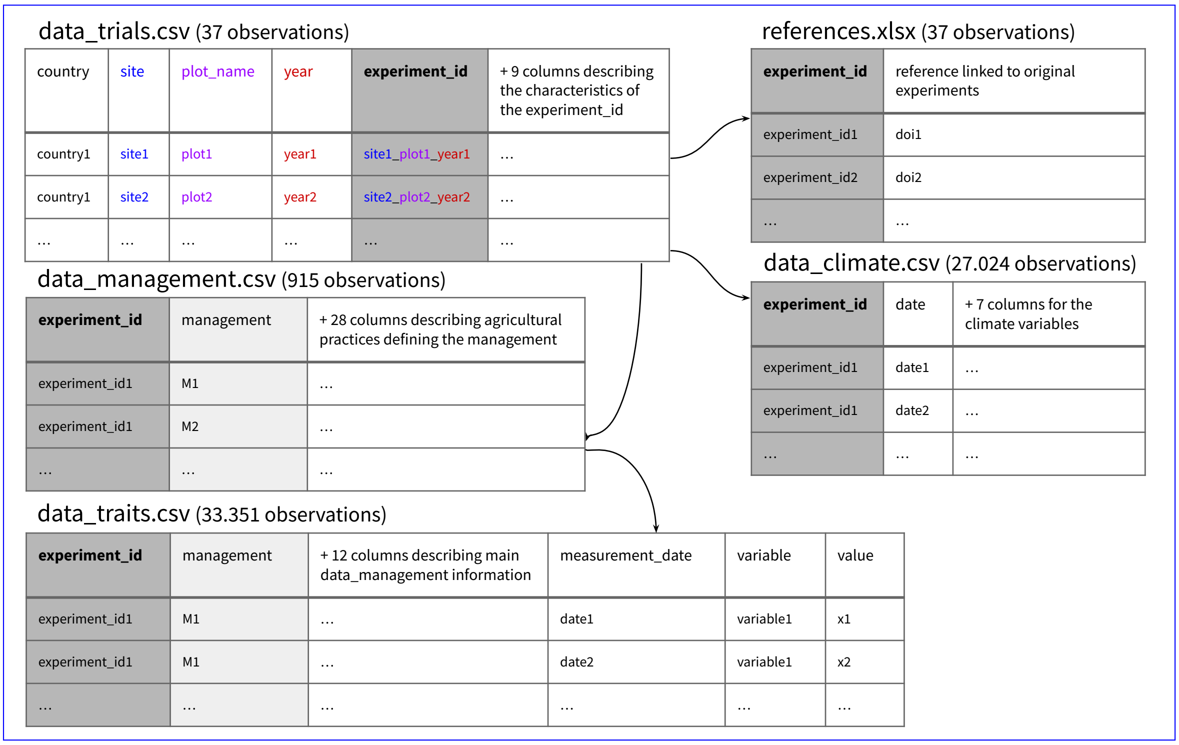

Figure 2. Representation of the relationships between tables identified in the global dataset. Five tables were defined to organize data, all sharing a common identifier (experiment_id, which is the concatenation of the site_plot_year of each experiment). The table *data_trials.csv* provides the main characteristics (*e.g.* latitude/longitude, soil texture) of each site, with one line per experiment (37 observations). The table *data_climate.csv* provides the climate time series during the growing season for each experiment (27.024 observations), retrieved using a gridded API (NASA POWER API, Sparks (2018)). The table *data_management.csv* describes the different agricultural practices used in each experimentation (*e.g.* species grown in sole- or intercrop, genotype, fertilization). The table *data_traits.csv* provides all the plant variables and their value as a function of time (measurement) per management and experiment (33.351 observations). Finally, the table *references.xlsx* provides the initial experimental references linked to each experiment (when existing).

Table 1. Overview of the diversity of the treatments in the global dataset by factors (columns) and experiments (rows). Within each column, each colored rectangle is a level of the factor considered. For instance, the two colors for the *Mixing pattern* indicate that the two species intercropped were sown in alternate rows or within the row; the two colors for the *Nitrogen (N) fertilization* indicate that the experiment included at least two N-treatments (no fertilization and N-fertilization, the latter of which may include several amounts of N); regarding *Species mixture*, the number of colors indicates the number of different species mixtures included in a given experiment. A rectangle in a given row and column indicates that the corresponding experiment contains at least one statistical individual with the corresponding factor level.

Experiment	No. of statistical individuals	No. of variables	Mixing pattern	Species mixture	Nitrogen fertilization
Taastrup_taastrup_2003	6	9			
SanMarco_sanMarco_2004	4	10			
SanMarco_sanMarco_2003	4	10			
Reading_reading_2003	6	10			
Kassel_kassel_2004	6	10			
Jyndevad_jyn_2003	24	11			
Jyndevad_jyn_2002	24	12			
Jyndevad_jyn_2001	24	12			
Grignon_inrae_2017	19	6			
Grignon_inrae_2010	16	8			
Grignon_inrae_2009	15	8			
Grignon_inrae_2008	27	5			
Grignon_inrae_2007	30	7			
Copenhagen_hbg_2003	24	10			
Copenhagen_hbg_2002	24	11			
Copenhagen_hbg_2001	24	12			
Auz_ZN_2012	58	24			
Auz_TO_2016	86	18			
Auz_TO_2013	93	24			
Auz_TE_2006	13	20			
Auz_SGs_2007	66	23			
Auz_PP_2011	20	20			
Auz_pk_2011	18	18			
Auz_marinette_2_2015	85	13			
Auz_marinette_1_2015	22	13			
Auz_cochard_2010	60	21			
Angers_thorigne_2009	11	12			
Angers_thorigne_2008	15	14			
Angers_thorigne_2007	11	12			
Angers_thorigne_2006	6	8			
Angers_thorigne_2004	6	10			
Angers_thorigne_2003	6	10			
Angers_jailliere_2008	22	16			
Angers_jailliere_2007	14	16			
Angers_fnams_2003	12	10			
Angers_fnams_2002	4	8			
Angers_brainsurlauthion_2011	10	5			

2. Share organized data

While there are relatively few incentives to share agronomical (Senft et al., 2022) or ecological (Jenkins et al., 2023) datasets, requirements and practices need to evolve (Krajewski et al., 2015). The ability to easily disseminates data is thus a key feature in designing a dataset, since it determines how other researchers will be able to interact with the data, and potentially increase its reuse. Open data should be designed in accordance with the FAIR data principles (https://force11.org/info/the-fair-data-principles/).

When discussing with the involved research groups, one recurrent constraint to open their data was the perception that their contribution could not be credited unless sharing authorship in research articles. If applied consistently, open-data FAIR requirements will allow contributors to be specifically acknowledged for their work, through citation of the dataset they contributed to (Jenkins et al., 2023).

This global dataset, as well as the metadata associated, are available on a data repository in a FAIR way (Gaudio et al., 2023). Out of the 37 experiments gathered, 11 have never been valued before.

Additional details on experimental designs and management practices are reported in the reference publications for 26 of the 37 experiments (Knudsen et al., 2004; Corre-Hellou et al., 2006; Hauggaard-Nielsen et al., 2008; Hauggaard-Nielsen et al., 2009a; b; Launay et al., 2009; Bedoussac and Justes, 2010a; b; Naudin et al., 2010, 2014; Barillot et al., 2014; Pelzer et al., 2016; Tang et al., 2016; Viguier et al., 2018; Kammoun et al., 2021).

3. Support new analysis

3.1. Conceptual framework

~~**Table 1.** Diversity of the treatments in the global dataset by factor (columns)and experiment (rows). Within each column, each colored rectangle is a level of the factor considered. A rectangle in a given row and column indicates that the corresponding experiment contains at least one statistical individual with the corresponding factor level.~~
Once the data are in a tractable format, visual exploration allows for a comprehensive overview of data patterns, aiding in the identification of anomalies such as errors and outliers that may not be immediately apparent through numerical analysis alone. Later, additional processes are required to render the dataset operational for analytical and modeling studies, such as data imputation, dimension reduction, or data normalization. Because these steps depend largely on the chosen analytical workflow, they are not directly

included in the communicated open datasets, but rather tailored by the subsequent analytical team (Figure 1). Nonetheless, sharing methods can support the future reuse of the dataset. In our case in crop ecology, we illustrated this step with the development of an original method aiming at identifying subsets in the overall dataset corresponding to complete factorial designs.

3.2. Case study

Method

The brief description of the global dataset revealed the diversity of agronomic situations considered (Table 1). While the experimental designs ~~had~~ share many similarities (~~e.g.~~ *e.g.* species cultivated, agricultural ~~management~~practices), the resulting overall design ~~did not allow an immediate statistical analysis of the global dataset~~is unbalanced. We thus developed a method to *a posteriori* identify subsets in the global dataset corresponding to complete factorial designs. This approach can quickly assess whether the dataset is suited to answer a set of scientific questions, as long as the factors of interest are sufficiently represented in the global dataset. The role of this method was not to identify potential confounding factors, which is left for the interpretation of the results of further statistical analysis

To identify the largest data subsets associated with complete factorial designs in the global dataset, we used tools from graph theory (Phillips et al., 2019). In graph theory, a graph G is a pair $G = (V, E)$ where V is a set of vertices, and E is a set of edges that connect some of the vertices (Table 2).

Table ~~2~~2. Definitions in graph theory used in the present study. ~~. Definitions in graph theory used in the present study (Phillips et al., 2019)~~

Term	Definition
subgraph $\widetilde{G} = (\widetilde{V}, \widetilde{E})$ of a graph $G = (V, E)$	A graph whose vertex set $(\widetilde{V})$ is included in the vertex set of G (~~i.e~~ *i.e.* $\widetilde{V} \subseteq V$) and whose edge set $(\widetilde{E})$ is included in the edge set of G (i.e $\widetilde{E} \subseteq E$)
complete graph	A graph whose vertices are all connected
clique of a graph G	A complete subgraph of G
maximal clique of a graph G	A clique that cannot be extended by including one more adjacent vertex
k-partite graph	A graph that can be partitioned into k ~~nonempty~~non-empty, vertex-disjoint, edgeless subgraphs
k-partite clique or *k-clique*	A set of vertices that induces a complete k-partite subgraph

Term	**Definition**
maximal k-partite clique	A k-clique that cannot be extended by including one more adjacent vertex

Given a set of categorical variables $X_1, ..., X_k$, each having values in a discrete set (~~i.e.~~ *i.e.* $\forall i = 1, ..., k \; X_i \in \mathcal{A}_i := \{x_{i,1}, ..., x_{i,j_i}\}$, ($j_i \in \mathbb{N}^*$ denoting the number of levels of variable X_i)), a k-partite graph can be derived by setting $V = \bigcup_{i=1}^{k} \mathcal{A}_i$ ~~, (i.e.~~ (*i.e.* each level of each factor is a vertex) ~~,~~ and $E = \{(x, y)|$ levels x and y observed together$\}$.

A factorial design is complete if, and only if, all possible combinations of the factor levels are present. For a graph $G = (V, E)$, this is equivalent to identifying a subgraph with an edge between each pair of vertices from independent sets (~~i.e.~~ *i.e.* a k-clique). Thus, the challenge of identifying the largest complete factorial designs within a global dataset can be reduced to counting the number of maximal k-cliques in the graph.

Phillips et al. (2019) developed the Maximum Multipartite Clique Enumeration (MMCE) algorithm to count the number of maximal multipartite cliques within a k-partite graph. MMCE starts from the observation that if G is k-partite, and if another graph G' is built from G by adding all intrapartite edges, then C is a maximal k-partite clique in G if C is a maximal clique in G' with at least one vertex in each partite set. Thus, the initial question is a matter of a modified problem of maximal clique enumeration, which is a NP-hard problem (Lawler et al., 1980). To address this issue, the MMCE algorithm uses a graph inflation approach, by adding all possible intrapartite edges to G. It then identifies maximal cliques in the inflated graph using a procedure of Bron and Kerbosch (1973) and checks whether the cliques identified cover all of the partite sets. We coded MMCE in the R programming language (https://github.com/RemiMahmoud/kclique). Although the problem of identifying maximal k-partite cliques with the maximum number of vertices has also been shown to be NP-hard for any $k \geq 3$ (Phillips et al., 2019), the relatively few vertices ($|V| < 300$) in the global dataset allowed solutions to be found quickly.

Application

Here, we illustrate this method with ~~a fictive global dataset made up from~~ two datasets : (1) a theoretical one, where we generated an unbalanced design of five environments~~(site*year), five crops~~, five intercrops, and two management levels (Figure ~~3).~~ 3A); and (2) a practical one, corresponding to the global dataset presented in this study (Figure 3B and 3C).

When applied on ~~this unbalanced design~~ the theoretical unbalanced design (Figure 3A), this method identified ~~11~~ 8 maximal 3-partite cliques, ~~with four examples illustrated in Figure 3. While~~ each of these ~~examples maximized the representativeness of a factor of interest (crop, environment,~~ designs having different number of modalities in considered factors (environment, intercrops or management)~~,~~. There is only one design maximizing the number of environments, and no factorial design was found with two levels per factor~~in this fictive dataset~~.

~~We also applied this method to address a specific issue (Mahmoud et al., 2022), in which we analyzed how nitrogen (N) fertilization influenced plant-plant interactions within intercrops. To this end, we looked for experiments that included both N-fertilized and unfertilized treatments by looking for a maximal 2-clique in a graph composed of two sets of vertices: i) field experiments and ii) N fertilization (i.e. unfertilized and N fertilized levels). The targeted maximal 2-clique needed to contain the two levels of the sets of N-fertilization vertices.~~

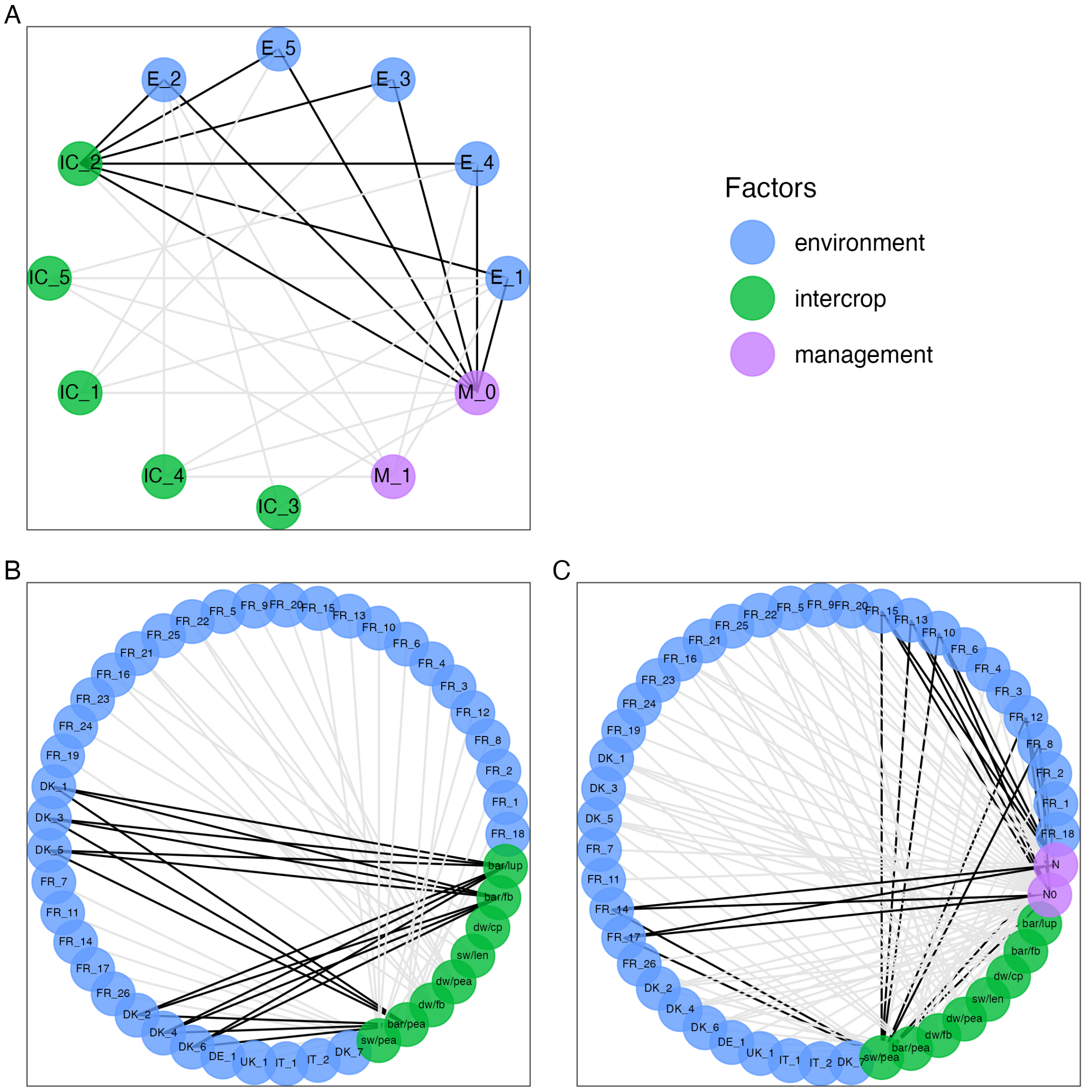

Figure 3. ~~Four~~ Three maximal ~~3-cliques~~ k-cliques that represent distinct complete factorial designs within ~~an~~ theoretical (A) and experimental (B-C) unbalanced ~~design with three factors~~designs. Black edges represent the edges of the ~~3-cliques~~ cliques and gray edges represent the factor combinations appearing in the initial design. ~~Despite the potential richness of the global dataset, there was no case where two levels of each factor were combined in a factorial design: network~~ In the case *A*~~focused on crops, network~~, we generated a random unbalanced design for three factors and illustrated the 3-clique maximizing the number of environments. The experimental design in the cases *B* ~~on environments, network *C* on management,~~ and ~~network *D* on crop and management together.~~ *C* corresponds to the aggregation of the 37 experimentations (blue nodes). In case *B*, we searched for any intercrop

observed at least in two environments. In case *C*, there was an additional constraint on two levels of nitrogen (N) fertilization. Countries were abbreviated with their ISO 3166 codes; species were abbreviated as barley (*bar*), chickpea (*cp*), durum wheat (*dw*), faba bean (*fb*), lentil (*len*), lupin (*lup*), soft wheat (*sw*); nitrogen fertilization was abbreviated as *N0* for no fertilization, and *N* for fertilization.

We considered two examples for the application on the agronomic global dataset. In the first one, we searched for any number of intercrops observed at least in two environments. Two designs were identified: the one with the most environmental modalities is illustrated in Figure 3B; the alternative design was, crossing {environments} x {intercrops}, {FR_22, FR_21} x {dw/pea, dw/fb}. The second example was the same request with an additional constraint on two levels of nitrogen (N) fertilization. In this case, three designs were identified, the largest one being illustrated in Figure 3C. The alternative designs were, crossing {environments} x {intercrops} x {N-fertilization}, {FR_9, FR_5, FR_22} x {dw/pea} x {N0, N} and {FR_22, FR_20, FR_16} x {dw/fb} x {N0, N}.

Discussion

One key reason to use agricultural data is to improve knowledge in crop science, as in other scientific fields. This can be generalized with the Data, Information, Knowledge and Wisdom pyramid (Ackoff, 1989), which describes the continuum between data and the knowledge it provides. Thus, the issue is to use appropriate methods based on the available data to provide insights and understanding of a studied system's functioning. Depending on whether data come from experimental data or from scientific publications, methods related to global datasets or meta-analysis, respectively, will be used (Makowski et al., 2014)~~, and both~~. Both are useful for studying global issues in agronomy (Table 3). Two important issues arise from this observation: data availability and the knowledge that one wants to provide.

Table 3. Overview of a comparison between meta-analysis and global datasets.

Criterion	Meta-analysis	Global datasets
Scope	All practices studied in multiple scientific publications	All practices tested in multiple experiments
Time required to collect and tidy the data	Long to very long (dozen to hundreds of hours)	Very long
Variables used	Often standard variables (~~e.g.~~ *e.g.* yield, nitrogen fertilization)	All available observations (~~e.g.~~ *e.g.* agronomic practices, phenotypic measurements, climate)
Number of observations	Moderate to large (dozens to hundreds)	Large (hundreds to thousands)
Reuse	Possible, but limited to the present variables	Possible once the data are formatted
Data sources	Scientific publications	Experimental files

In meta-analysis, data are available because they are already published, even if it takes a long time to retrieve them. Conducting a meta-analysis is thus time-consuming, especially the pre-analysis search and development of the database, which represent around 60% of the working time (Allen and Olkin, 1999). Meta-analysis requires identifying and extracting the values of interest from scientific publications, while being cautious to avoid potential bias.

In contrast, building global datasets requires interacting with the research teams that conducted the experiments and adapting their raw experimental files to a standard format (Figure 1). This step itself is very likely to necessitate more time than meta-analysis data processing step~~.~~ , and would greatly benefit from improved upstream data standardization practices (Krajewski et al., 2015). The main advantage of global datasets in biology is that they consist of phenotypic observations, which means that the studied processes are potentially observed at lower levels than in meta-analysis. In this sense, global datasets could enable further investigation of potential causalities based on correlations in the data (Garside and Bell, 2011; Gunawardena, 2014). Additionally, since agronomic global datasets contain plant-related variables measured at multiple organizational levels (~~e.g.~~ *e.g.* organ, plant, crop), they can target a wide audience for data reuse. For instance, researchers developing functional–structural plant models (Louarn et al., 2020) may be interested in variables measured at the plant scale (~~e.g.~~ *e.g.* number of tillers, inter-node length, plant height), while those who develop crop models to predict ~~yields~~ yield (Berghuijs et al., 2021) may be interested in variables measured at the crop scale (~~e.g.~~ *e.g.* crop biomass, crop height).

Alternatively, global datasets might have a role in increasing the discovery and use of non-published experimental data. In our case, almost 30% of the experimental data gathered have not been published through a research article. Bringing them together with other experiments valued the time and energy required to conduct those field experiments. It was also a friction point, since researchers may be reluctant to share unpublished data. For instance, in our use case, 11 of the 37 experiments were not included in published articles or database before this initiative, while each is now described within the global dataset (Gaudio et al., 2023) and linked back groups leading field experiments in 1-4 scientific publications (Gaudio et al., 2021; Louarn et al., 2021; Mahmoud et al., 2022; Meunier et al., 2022). Based on the global dataset developed in this study, Gaudio et al. (2021) extracted a subset of 28 experiments to assess the influence of intercropping on the relation between plant biomass and grain yield; Louarn et al. (2021) extracted a subset of 15 experiments to validate the adaptation of Nitrogen Nutrition Index (NNI) to intercropping; Mahmoud et al. (2022) extracted a subset of 11 experiments to assess the influence of ~~N~~ nitrogen fertilization on plant-plant interactions in intercrops; and Meunier et al. (2022) extracted a subset of 31 experiments to calibrate a statistical model used in a modeling chain to predict ecosystem services as a function of the species associated in cereal-legume intercrops.

We argue that crop science can benefit from global datasets because they decrease the cost of data (reuse) and increase the reproducibility of studies along with open data science tools (Lowndes et al., 2017). Ultimately, global datasets contribute to new findings through joint analysis of multiple experiments - a key consideration given the pressing need for consolidating results in the context of an increasingly variable and changing climate. Despite these needs for advancements, the challenges associated with the data standardization and proprietary rights present significant obstacles to the ~~utilization~~ building of these global datasets in crop science. A tighter integration between experimental and modeling research communities is the first step in a way forward.

Acknowledgements

We thank the entire technical staff of the different research teams who shared their data, for all the huge work they have done, without which this paper and the associated dataset would not exist. We thank Michael and Michelle Corson for their helpful comments and English revision, and the three reviewers for their valuable comments and corrections which highly contribute to improve the manuscript.

Funding

This research was supported by the French National Research Agency under the Investments for the Future Program (referred to as ANR-16-CONV-0004 and ANR-20-PCPA-0006) and by the European Research Council under the European Union's Horizon Europe research and innovation program in the framework of the IntercropValuES (Developing Intercropping for agrifood Value chains and Ecosystem Services delivery in Europe and Southern countries, https://intercropvalues.eu/) ~~project~~ starting from November 2022 [grant number 101081973]. ~~We thank Michael and Michelle Corson for their helpful comments and English revision.~~

Conflict of interest disclosure

The authors have no relevant financial or non-financial interests to disclose. On behalf of all authors, the corresponding author states that there is no conflict of interest.

Author Contributions

All authors contributed to funding acquisition, data collection and formatting, writing and editing the manuscript.

Data Availability

The global dataset is available on Zenodo open data repository (Gaudio et al., 2023).

References

Ackoff, R.L. 1989. From data to wisdom. Journal of applied systems analysis 16(1): 3–9.

Allen, I.E., and I. Olkin. 1999. Estimating Time to Conduct a Meta-analysis From Number of Citations Retrieved. JAMA 282(7): 634–635. doi: 10.1001/JAMA.282.7.634.

Barillot, R., D. Combes, S. Pineau, P. Huynh, and A.J. Escobar-Gutierrez. 2014. Comparison of the morphogenesis of three genotypes of pea (Pisum sativum) grown in pure stands and wheat-based intercrops. Aob Plants 6: plu006. doi: https://doi.org/10.1093/aobpla/plu006.

Bedoussac, L., and E. Justes. 2010a. Dynamic analysis of competition and complementarity for light and N use to understand the yield and the protein content of a durum wheat–winter pea intercrop. Plant and Soil 330(1-2): 37–54. doi: https://doi.org/10.1007/s11104-010-0303-8.

Bedoussac, L., and E. Justes. 2010b. The efficiency of a durum wheat-winter pea intercrop to improve yield and wheat grain protein concentration depends on N availability during early growth. Plant and Soil 330(1-2): 19–35. doi: https://doi.org/10.1007/s11104-009-0082-2.

Berghuijs, H.N.C., M. Weih, W. Van Der Werf, A.J. Karley, E. Adam, et al. 2021. Calibrating and testing APSIM for wheat-faba bean pure cultures and intercrops across Europe. Field Crops Research 264: 108088. doi: 10.1016/j.fcr.2021.108088.

Broman, K.W., and K.H. Woo. 2018. Data organization in spreadsheets. The American Statistician 72(1): 2–10. doi: 10.1080/00031305.2017.1375989.

Bron, C., and J. Kerbosch. 1973. Algorithm 457: Finding All Cliques of an Undirected Graph [H]. Communications of the ACM 16(9): 575–577. doi: 10.1145/362342.362367.

Casler, M.D. 2015. Fundamentals of experimental design: Guidelines for designing successful experiments. Agronomy Journal 107(2): 692–705. doi: https://doi.org/10.2134/agronj2013.0114.

Corre-Hellou, G., J. Fustec, and Y. Crozat. 2006. Interspecific competition for soil N and its interaction with N-2 fixation, leaf expansion and crop growth in pea-barley intercrops. Plant and Soil 282(1-2): 195–208. doi: https://doi.org/10.1007/s11104-005-5777-4.

Cruz, S.M.S. da, and J.A.P. do Nascimento. 2019. Towards integration of data-driven agronomic experiments with data provenance. Computers and Electronics in Agriculture 161(September 2018): 14–28. doi: 10.1016/j.compag.2019.01.044.

Duru, M., O. Therond, G. Martin, R. Martin-Clouaire, M.-A. Magne, et al. 2015. How to implement biodiversity-based agriculture to enhance ecosystem services: A review. Agronomy for Sustainable Development 35(4): 1259–1281. doi: 10.1007/s13593-015-0306-1.

Garside, A.L., and M.J. Bell. 2011. Growth and yield responses to amendments to the sugarcane monoculture: Towards identifying the reasons behind the response to breaks. Crop and Pasture Science 62(9): 776–789. doi: 10.1071/CP11055.

Gaudio, N., R. Mahmoud, L. Bedoussac, E. Justes, E.-P. Journet, et al. 2023. A global dataset gathering 37 field experiments involving cereal-legume intercrops and their corresponding sole crops. doi: 10.5281/zenodo.8081577.

Gaudio, N., C. Violle, X. Gendre, F. Fort, R. Mahmoud, et al. 2021. Interspecific interactions regulate plant reproductive allometry in cereal–legume intercropping systems. Journal of Applied Ecology 58(11): 2579–2589. doi: https://doi.org/10.1111/1365-2664.13979.

Glass, G.V. 1976. Primary, secondary, and meta-analysis of research. Educational researcher 5(10): 3–8.

Gunawardena, J. 2014. Models in biology: 'Accurate descriptions of our pathetic thinking'. BMC Biology 12(1): 1–11. doi: 10.1186/1741-7007-12-29/FIGURES/3.

Hauggaard-Nielsen, H., M. Gooding, P. Ambus, G. Corre-Hellou, Y. Crozat, et al. 2009a. Pea-barley intercropping for efficient symbiotic N-2-fixation, soil N acquisition and use of other nutrients in European organic cropping systems. Field Crops Research 113(1): 64–71. doi: https://doi.org/10.1016/j.fcr.2009.04.009.

Hauggaard-Nielsen, H., M. Gooding, P. Ambus, G. Corre-Hellou, Y. Crozat, et al. 2009b. Pea-barley intercropping and short-term subsequent crop effects across European organic cropping conditions. Nutrient Cycling in Agroecosystems 85(2): 141–155. doi: https://doi.org/10.1007/s10705-009-9254-y.

Hauggaard-Nielsen, H., B. Jørnsgaard, J. Kinane, and E.S. Jensen. 2008. Grain legume–cereal intercropping: The practical application of diversity, competition and facilitation in arable and organic cropping systems. Renewable Agriculture and Food Systems 23(1): 3–12. doi: https://doi.org/10.1017/S1742170507002025.

Jenkins, G.B., A.P. Beckerman, C. Bellard, A. Benıtez-López, A.M. Ellison, et al. 2023. Reproducibility in ecology and evolution: Minimum standards for data and code. Ecology and Evolution 13(5). doi: 10.1002/ece3.9961.

Kammoun, B., E.-P. Journet, E. Justes, and L. Bedoussac. 2021. Cultivar Grain Yield in Durum Wheat-Grain Legume Intercrops Could Be Estimated From Sole Crop Yields and Interspecific Interaction Index. Frontiers in Plant Science 12: 2191. doi: https://doi.org/10.3389/fpls.2021.733705.

Kattge, J., S. Diaz, S. Lavorel, I.C. Prentice, P. Leadley, et al. 2011. TRY–a global database of plant traits. Global change biology 17(9): 2905–2935.

Knudsen, M.T., H. Hauggaard-Nielsen, B. Jornsgard, and E.S. Jensen. 2004. Comparison of interspecific competition and N use in pea-barley, faba bean-barley and lupin-barley intercrops grown at two temperate locations. Journal of Agricultural Science 142: 617–627. doi: https://doi.org/10.1017/S0021859604004745.

Krajewski, P., D. Chen, H. Ćwiek, A.D.J. van Dijk, F. Fiorani, et al. 2015. Towards recommendations for metadata and data handling in plant phenotyping. Journal of Experimental Botany 66(18): 5417–5427. doi: https://doi.org/10.1093/jxb/erv271.

Launay, M., N. Brisson, S. Satger, H. Hauggaard-Nielsen, G. Corre-Hellou, et al. 2009. pre Exploring options for managing strategies for pea-barley intercropping using a modeling approach. European Journal of Agronomy 31(2): 85–98. doi: https://doi.org/10.1016/j.eja.2009.04.002.

Lawler, E.L., J.K. Lenstra, and A.H.G. Rinnooy Kan. 1980. Generating All Maximal Independent Sets: NP-Hardness and Polynomial-Time Algorithms. SIAM Journal on Computing 9(3): 558–565. doi: 10.1137/0209042.

Licker, R., M. Johnston, J.A. Foley, C. Barford, C.J. Kucharik, et al. 2010. Mind the gap: How do climate and agricultural management explain the 'yield gap' of croplands around the world? Global Ecology and Biogeography 19(6): 769–782. doi: 10.1111/j.1466-8238.2010.00563.x.

Lobell, D.B., J.M. Deines, and S.D. Tommaso. 2020. Changes in the drought sensitivity of US maize yields. Nature Food 1(11): 729–735. doi: 10.1038/s43016-020-00165-w.

Louarn, G., R. Barillot, Di. Combes, and A. Escobar-Gutiérrez. 2020. Towards intercrop ideotypes: Non-random trait assembly can promote overyielding and stability of species proportion in simulated legume-based mixtures. Annals of Botany 126(4): 671–685. doi: 10.1093/aob/mcaa014.

Louarn, G., L. Bedoussac, N. Gaudio, E.P. Journet, D. Moreau, et al. 2021. Plant nitrogen nutrition status in intercrops– a review of concepts and methods. European Journal of Agronomy 124: 126229. doi: 10.1016/J.EJA.2021.126229.

Lowndes, J.S.S., B.D. Best, C. Scarborough, J.C. Afflerbach, M.R. Frazier, et al. 2017. Our path to better science in less time using open data science tools. Nature Ecology & Evolution 1(6): 1–7. doi: https://doi.org/10.1038/s41559-017-0160.

Maat, H. 2011. The history and future of agricultural experiments. NJAS - Wageningen

Journal of Life Sciences 57(3-4): 187–195. doi: 10.1016/j.njas.2010.11.001.

Mahmoud, R., P. Casadebaig, N. Hilgert, L. Alletto, G.T. Freschet, et al. 2022. Species choice and n fertilization influence yield gains through complementarity and selection effects in cereal-legume intercrops. Agronomy for sustainable development. doi: 10.1007/s13593-022-00754-y.

Makowski, D., T. Nesme, F. Papy, and T. Doré. 2014. Global agronomy, a new field of research. A review. Agronomy for Sustainable Development 34(2): 293–307. doi: 10.1007/s13593-013-0179-0.

Meunier, C., L. Alletto, L. Bedoussac, J.E. Bergez, P. Casadebaig, et al. 2022. A modelling chain combining soft and hard models to assess a bundle of ecosystem services provided by a diversity of cereal-legume intercrops. European Journal of Agronomy 132(October 2021). doi: 10.1016/j.eja.2021.126412.

Naudin, C., G. Corre-Hellou, S. Pineau, Y. Crozat, and M.-H. Jeuffroy. 2010. The effect of various dynamics of N availability on winter pea-wheat intercrops: Crop growth, N partitioning and symbiotic N-2 fixation. Field Crops Research 119(1): 2–11. doi: https://doi.org/10.1016/j.fcr.2010.06.002.

Naudin, C., H.M.G. van der Werf, M.-H. Jeuffroy, and G. Corre-Hellou. 2014. Life cycle assessment applied to pea-wheat intercrops: A new method for handling the impacts of co-products. Journal of Cleaner Production 73: 80–87. doi: https://doi.org/10.1016/j.jclepro.2013.12.029.

Newman, S.J., and R.T. Furbank. 2021. A multiple species, continent-wide, million-phenotype agronomic plant dataset. Scientific Data 8(1): 1–8. doi: 10.1038/s41597-021-00898-8.

Pelzer, E., M. Bazot, L. Guichard, and M.-H. Jeuffroy. 2016. Crop Management Affects the Performance of a Winter Pea–Wheat Intercrop. Agronomy Journal 108(3): 1089–1100. doi: https://doi.org/10.2134/agronj2015.0440.

Phillips, C.A., K. Wang, E.J. Baker, J.A. Bubier, E.J. Chesler, et al. 2019. On Finding and enumerating maximal and maximum k-partite cliques in k-partite graphs. Algorithms 12(1). doi: 10.3390/a12010023.

Popkin, G. 2019. Data sharing and how it can benefit your scientific career. Nature 569(7756): 445–447. doi: 10.1038/d41586-019-01506-x.

Senft, M., U. Stahl, and N. Svoboda. 2022. Research data management in agricultural sciences in germany: We are not yet where we want to be (C. Pulvento, editor). PLOS ONE 17(9): e0274677. doi: 10.1371/journal.pone.0274677.

Sparks, A.H. 2018. Nasapower: A NASA POWER Global Meteorology, Surface Solar Energy and Climatology Data Client for R. Journal of Open Source Software 3(30): 1035. doi: 10.21105/joss.01035.

Tang, X., S.A. Placella, F. Dayde, L. Bernard, A. Robin, et al. 2016. Phosphorus availability

pre
and microbial community in the rhizosphere of intercropped cereal and legume along a P-fertilizer gradient. Plant and Soil 407(1-2): 119–134. doi: https://doi.org/10.1007/s11104-016-2949-3.

Tardieu, F. 2020. Educated big data to study sensitivity to drought. Nature Food 1(11): 669–670. doi: 10.1038/s43016-020-00187-4.

Viguier, L., L. Bedoussac, E.-P. Journet, and E. Justes. 2018. Yield gap analysis extended to marketable grain reveals the profitability of organic lentil-spring wheat intercrops. Agronomy for Sustainable Development 38(4): 39. doi: https://doi.org/10.1007/s13593-018-0515-5.

White, J.W., and F.K. Van Evert. 2008. Publishing agronomic data. Agronomy Journal 100(5): 1396–1400. doi: 10.2134/agronj2008.0080F.

Wickham, H. 2014. Tidy data. Journal of Statistical Software 59(10). doi: 10.18637/jss.v059.i10.

Wickham, H., and G. Grolemund. 2016. R for data science: Import, tidy, transform, visualize, and model data. " O'Reilly Media, Inc.".

Wilkinson, M.D., M. Dumontier, Ij.J. Aalbersberg, G. Appleton, M. Axton, et al. 2016. Comment: The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3: 1–9. doi: 10.1038/sdata.2016.18.

Wilson, G., J. Bryan, K. Cranston, J. Kitzes, L. Nederbragt, et al. 2017. Good enough practices in scientific computing. PLOS Computational Biology 13(6): e1005510. doi: 10.1371/JOURNAL.PCBI.1005510.

Zamir, D. 2013. Where Have All the Crop Phenotypes Gone? PLoS Biology 11(6): 1–4. doi: 10.1371/journal.pbio.1001595.