



HAL
open science

A workflow for processing global datasets: application to intercropping

Rémi Mahmoud, Pierre Casadebaig, Nadine Hilgert, Noémie Gaudio

► To cite this version:

Rémi Mahmoud, Pierre Casadebaig, Nadine Hilgert, Noémie Gaudio. A workflow for processing global datasets: application to intercropping. 2023. hal-04145269v1

HAL Id: hal-04145269

<https://hal.inrae.fr/hal-04145269v1>

Preprint submitted on 29 Jun 2023 (v1), last revised 27 Mar 2024 (v4)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial - ShareAlike 4.0 International License

1 **A workflow for processing global datasets: application to** 2 **intercropping**

3 Rémi Mahmoud¹, Pierre Casadebaig^{1*}, Nadine Hilgert², Noémie Gaudio¹

4 (1) AGIR, Univ. Toulouse, INRAE, Castanet-Tolosan, France

5 (2) MISTEA, Univ. Montpellier, INRAE, Institut Agro, Montpellier, France

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

7 **Abstract**

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

18 Introduction

19 Field experiments, whether conducted on farms or at experimental research stations, have
20 traditionally been the primary approach for acquiring knowledge in crop sciences (Maat,
21 2011). Yet, extrapolating applicable principles from localized experiments remains a chal-
22 lenging task (Makowski et al., 2014). To derive general rules about agroecosystem function-
23 ing, meta-analysis, i.e. a “statistical analysis of a large collection of analysis results from
24 individual studies to integrate the findings” (Glass, 1976), is typically employed. Alter-
25 natively, global datasets, corresponding to the aggregation of observations from numerous
26 experiments, can serve as another valuable tool for analyzing agronomic data. Distinguish-
27 ing themselves from meta-analyses, global datasets compile raw experimental results on a
28 detailed scale, such as repeated measurements on individuals or multiple state variables on
29 the canopy. In contrast, meta-analysis is typically restricted to published results with a
30 limited set of variables.

31 Although examples of comprehensive agronomic datasets exist (Kattge et al., 2011; New-
32 man and Furbank, 2021), only a few studies have been based on global datasets (Licker et
33 al., 2010; Lobell et al., 2020; Newman and Furbank, 2021) with even less focus on methods
34 for this type of datasets in crop science (Senft et al., 2022). One significant advantage of
35 agronomic global datasets relies on the fact that they include diverse phenotypic observa-
36 tions from varying soils and climates, enabling more reliable generalization of local findings
37 (Tardieu, 2020). These datasets reduce the risk of spurious correlations (Tardieu, 2020) and
38 maximize the utility of experimental data yet to be used in scientific publications (Zamir,
39 2013).

40 However, global datasets come with their own challenges. Assembling these datasets requires
41 extensive data collection, standardization, and homogenization across diverse experiments
42 conducted by different research teams (White and Van Evert, 2008; Makowski et al., 2014).
43 The different field experiments often have diverse objectives, leading to unbalanced and
44 incomplete designs. Confounding factors, i.e. the unintended mixing of two or more effects
45 making them indistinguishable, can also be challenging (Casler, 2015). Consequently, using
46 and analyzing global datasets require a thorough understanding of the dataset, judicious in-
47 terpretation of results, identification of balanced data subsets for specific research questions,
48 and acceptance that the effects of some factors may remain indistinguishable. Therefore,
49 the application of statistical learning techniques on global datasets is only feasible after
50 extensive data pre-processing.

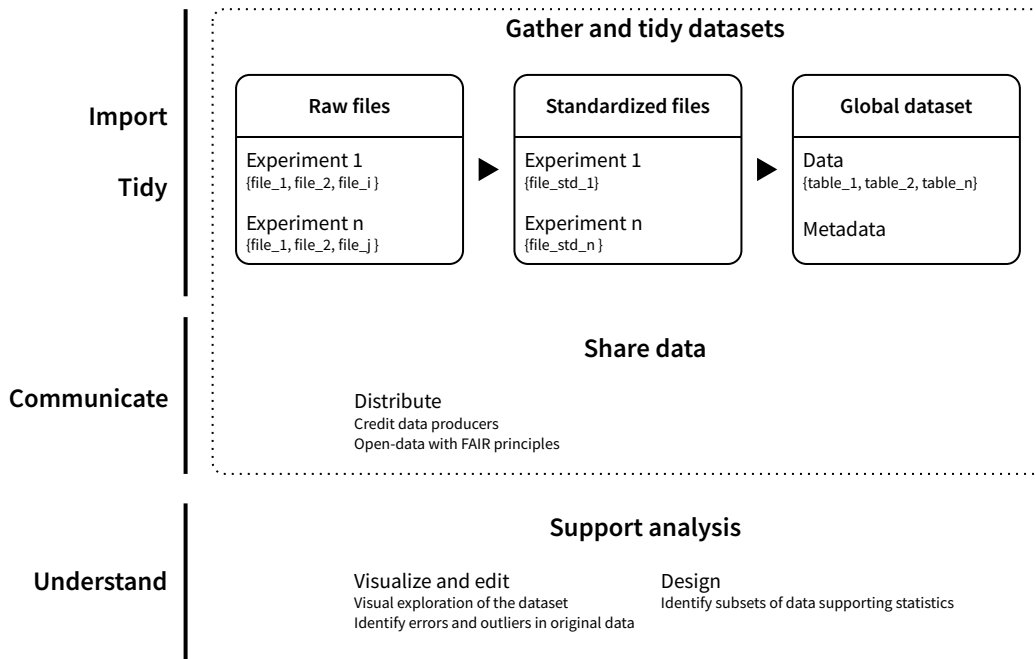
51 Despite these challenges, crop science would greatly benefit from the study of global datasets
52 combining multiple experiments (White and Van Evert, 2008; Zamir, 2013; Cruz and Nasci-

53 mento, 2019). This approach is particularly relevant considering the current agricultural
54 landscape, where crop diversification is crucial for sustainable farming (Duru et al., 2015).
55 This diversification mandates extensive experimentation, requiring robust data-federation
56 efforts. The joint analysis of global datasets makes it possible to understand the context-
57 dependent nature of diverse experiments and enhances comprehension of the interaction
58 between crop diversity and agroecosystem functioning.

59 To achieve this, we recommend adopting practices for designing and analyzing global
60 datasets that align with tidy data (Wickham, 2014; Broman and Woo, 2018) and FAIR
61 principles (Findable, Accessible, Interoperable, and Reusable) (Wilkinson et al., 2016). As
62 a use case, we illustrate the design of a global dataset for intercropping systems, in which
63 at least two crop species are grown in the same field for a significant part of their growth
64 cycle. We describe the main steps involved in designing a global dataset gathering 37
65 intercropping experiments across Europe. We also describe and apply an original method
66 for identifying factorial designs, which is a key step in assisting modeling and analysis
67 steps.

68 **Designing global datasets**

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



73

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

79 1. Gathering and tidying datasets

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

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

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

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

104 2. Distributing datasets

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

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

116 **3. Supporting analysis**

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

120 Later, additional processes are required to render the dataset operational for analytical
121 and modeling studies, such as data imputation, dimension reduction, or data normaliza-
122 tion. Because these steps depend largely on the chosen analytical workflow, they are not
123 directly included in the communicated open datasets, but rather tailored by the subsequent
124 analytical team (dotted frame in Figure 1).

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

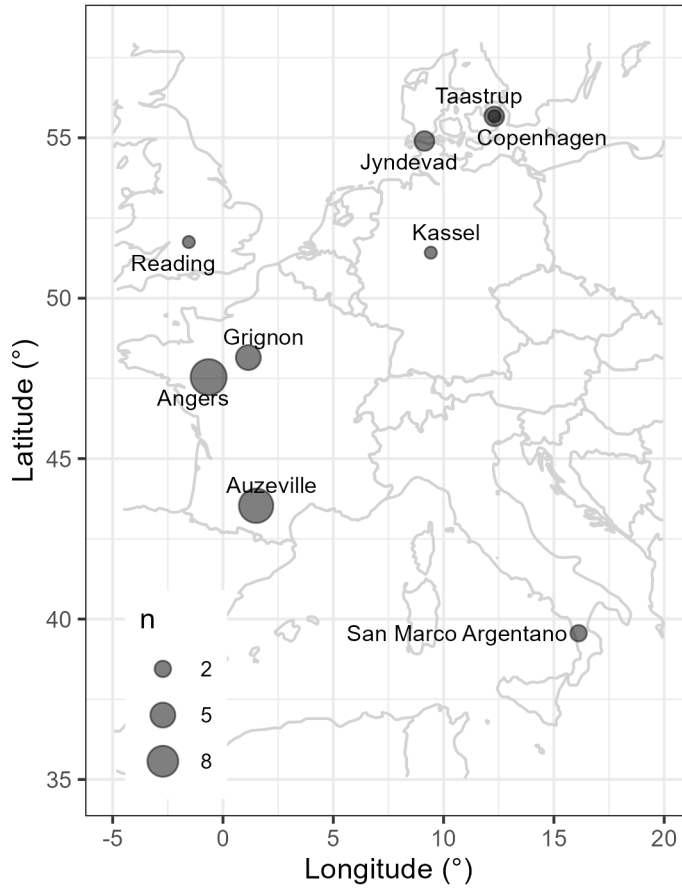
129 **Case study**

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

133 **1. Intercropping context and experimental data**

134 Although combining results from a few experiments (usually two years, often sequential)
135 is common in the intercropping literature (and more generally in crop science), no study
136 includes joint analysis of dozens of experiments to infer more generic results about intercrop-
137 ping functioning. To this end, we designed, built and analyzed a global dataset gathering the
138 results of 37 field experiments that involved cereal-legume intercrops and the corresponding
139 sole crops. Globally, the aim of these field experiments was to compare the growth and
140 grain yield ($\text{t}\cdot\text{ha}^{-1}$) of multiple combinations of species grown in intercrop to their sole-crop

141 reference. The field experiments were carried in 5 European countries (France, Denmark,
142 Italy, Germany and England) from 2001 to 2017 (Figure 2). The global dataset included 5
143 legume species (chickpea, faba bean, lentil, lupin and pea), 3 cereal species (barley, durum
144 wheat and soft wheat) and 8 resulting intercrops.



145

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

147 **2. Gathering, tidying and distributing**

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

155 After the step of gathering, the files were transformed into standardized rectangular data
156 tables, following the tidy format and good practices (Wickham, 2014; Broman and Woo,
157 2018), resulting in the creation of one given file per experiment. Each file includes 6 spread-
158 sheets, in which the variables and values were placed as a function of the information they
159 provided (e.g. plant functioning, climate, agricultural practices). This step resulted in the
160 creation of 37 excel files (vs. 86) and 222 spreadsheets (vs. 412).

161 Finally, all the files were pooled together using R software, with a final table per type of
162 variable, i.e. four tables related respectively to climate, crop measurements, agricultural
163 practices and global information describing the site. Overall, the global dataset contained
164 308 and 299 statistical individuals (i.e. a unique combination of site * year * management)
165 in intercrop and sole crop, respectively (Table 1). The number of plant characteristics was
166 much larger (33351 observations, among which 12896 were measured in sole crops and 20455
167 in intercrops), since several variables were measured at the crop scale, sometimes several
168 times during the crop cycle.

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

172 Additional details on experimental designs and management practices are reported in the
173 reference publications for 26 of the 37 experiments (Knudsen et al., 2004; Corre-Hellou et
174 al., 2006; Hauggaard-Nielsen et al., 2008; Hauggaard-Nielsen et al., 2009a; b; Launay et al.,
175 2009; Bedoussac and Justes, 2010a; b; Naudin et al., 2010, 2014; Barillot et al., 2014; Pelzer
176 et al., 2016; Tang et al., 2016; Viguier et al., 2018; Kammoun et al., 2021).

177 **Table 1.** Diversity of the treatments in the global dataset by factor (columns) and experiment
 178 (rows). Within each column, each colored rectangle is a level of the factor considered. A
 179 rectangle in a given row and column indicates that the corresponding experiment contains at
 180 least one statistical individual with the corresponding factor level.

Experiment	No. of statistical individuals	No. of variables	Sowing spatial arrangement	Species mixture	Nitrogen fertilization
Taastrup_taastrup_2003	6	9	[Red]	[Green]	[Cyan]
SanMarco_sanMarco_2004	4	10	[Red]	[Green]	[Cyan]
SanMarco_sanMarco_2003	4	10	[Red]	[Green]	[Cyan]
Reading_reading_2003	6	10	[Red]	[Green]	[Cyan]
Kassel_kassel_2004	6	10	[Red]	[Green]	[Cyan]
Jynde vad_jyn_2003	24	11	[Red]	[Red][Green]	[Cyan]
Jynde vad_jyn_2002	24	12	[Red]	[Red][Green]	[Cyan]
Jynde vad_jyn_2001	24	12	[Red]	[Red][Green]	[Cyan]
Grignon_inrae_2017	19	6	[Red]	[Cyan]	[Cyan]
Grignon_inrae_2010	16	8	[Red]	[Cyan]	[Red][Cyan]
Grignon_inrae_2009	15	8	[Red]	[Cyan]	[Red][Cyan]
Grignon_inrae_2008	27	5	[Red]	[Cyan]	[Red][Cyan]
Grignon_inrae_2007	30	7	[Red]	[Cyan]	[Red][Cyan]
Copenhagen_hbg_2003	24	10	[Red]	[Red][Green]	[Cyan]
Copenhagen_hbg_2002	24	11	[Red]	[Red][Green]	[Cyan]
Copenhagen_hbg_2001	24	12	[Red]	[Red][Green]	[Cyan]
Auz_ZN_2012	58	24	[Red]	[Cyan]	[Cyan]
Auz_TO_2016	86	18	[Red]	[Green]	[Cyan]
Auz_TO_2013	93	24	[Red]	[Cyan]	[Red][Cyan]
Auz_TE_2006	13	20	[Red]	[Cyan]	[Red][Cyan]
Auz_SGs_2007	66	23	[Red]	[Cyan]	[Red][Cyan]
Auz_PP_2011	20	20	[Red]	[Cyan]	[Red][Cyan]
Auz_pk_2011	18	18	[Red]	[Cyan]	[Red][Cyan]
Auz_marinette_2_2015	85	13	[Red]	[Green]	[Cyan]
Auz_marinette_1_2015	22	13	[Red]	[Cyan]	[Cyan]
Auz_cochard_2010	60	21	[Red]	[Cyan]	[Red][Cyan]
Angers_thorigne_2009	11	12	[Red]	[Cyan]	[Red][Cyan]
Angers_thorigne_2008	15	14	[Red]	[Cyan]	[Red][Cyan]
Angers_thorigne_2007	11	12	[Red]	[Cyan]	[Red][Cyan]
Angers_thorigne_2006	6	8	[Red]	[Cyan]	[Red][Cyan]
Angers_thorigne_2004	6	10	[Red]	[Green]	[Cyan]
Angers_thorigne_2003	6	10	[Red]	[Green]	[Cyan]
Angers_jailliere_2008	22	16	[Red]	[Cyan]	[Red][Cyan]
Angers_jailliere_2007	14	16	[Red]	[Cyan]	[Red][Cyan]
Angers_fnams_2003	12	10	[Red]	[Green]	[Red][Cyan]
Angers_fnams_2002	4	8	[Red]	[Green]	[Red][Cyan]
Angers_brainsurlauthion_2011	10	5	[Red]	[Cyan]	[Cyan]

181

182 3. Supporting analysis

183 The brief description of the global dataset revealed the diversity of agronomic situations
 184 considered (Table 1). While the experimental designs had many similarities (e.g. species
 185 cultivated, agricultural management), the resulting overall design did not allow an immedi-
 186 ate statistical analysis of the global dataset. We thus developed a method to *a posteriori*
 187 identify subsets in the global dataset corresponding to complete factorial designs. This ap-
 188 proach can quickly assess whether the dataset is suited to answer a set of scientific questions,
 189 as long as the factors of interest are sufficiently represented in the global dataset.

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

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

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

195 Given a set of categorical variables X_1, \dots, X_k , each having values in a discrete set (i.e. $\forall i =$
 196 $1, \dots, k$ $X_i \in \mathcal{A}_i := \{x_{i,1}, \dots, x_{i,j_i}\}$, ($j_i \in \mathbb{N}^*$ denoting the number of levels of variable X_i)),
 197 a k -partite graph can be derived by setting $V = \bigcup_{i=1}^k \mathcal{A}_i$, (i.e. each level of each factor is a
 198 vertex), and $E = \{(x, y) \mid \text{levels } x \text{ and } y \text{ observed together}\}$.

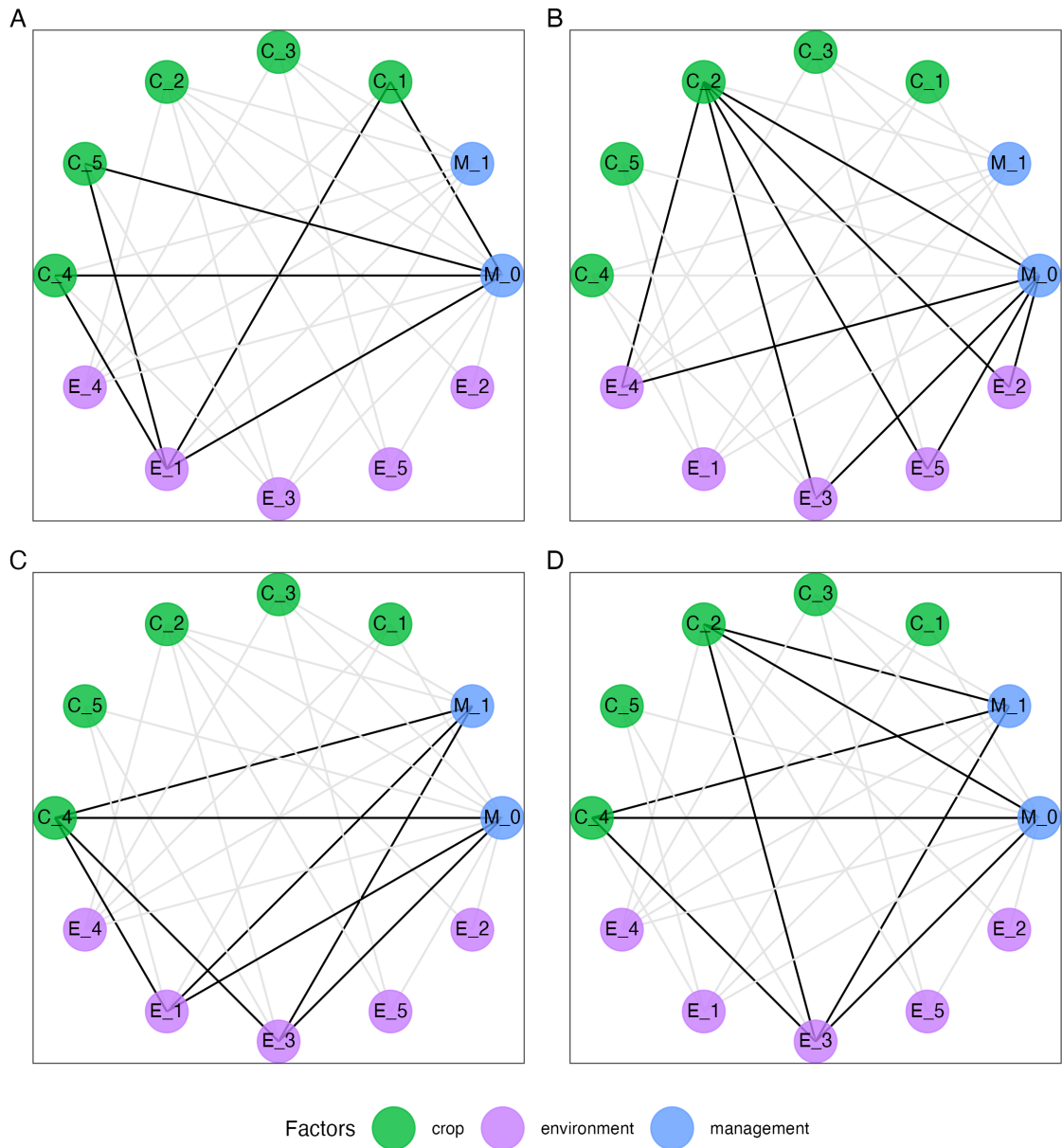
199 A factorial design is complete if, and only if, all possible combinations of the factor levels are
 200 present. For a graph $G = (V, E)$, this is equivalent to identifying a subgraph with an edge
 201 between each pair of vertices from independent sets (i.e. a k -clique). Thus, the challenge of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

204 Phillips et al. (2019) developed the Maximum Multipartite Clique Enumeration (MMCE)
 205 algorithm to count the number of maximal multipartite cliques within a k -partite graph.
 206 MMCE starts from the observation that if G is k -partite, and if another graph G' is built
 207 from G by adding all intrapartite edges, then C is a maximal k -partite clique in G if C
 208 is a maximal clique in G' with at least one vertex in each partite set. Thus, the initial
 209 question is a matter of a modified problem of maximal clique enumeration, which is a NP -
 210 hard problem (Lawler et al., 1980). To address this issue, the MMCE algorithm uses a
 211 graph inflation approach, by adding all possible intrapartite edges to G . It then identifies

212 maximal cliques in the inflated graph using a procedure of Bron and Kerbosch (1973) and
213 checks whether the cliques identified cover all of the partite sets. We coded MMCE in
214 the R programming language (<https://github.com/RemiMahmoud/kclique>). Although the
215 problem of identifying maximal k -partite cliques with the maximum number of vertices has
216 also been shown to be *NP*-hard for any $k \geq 3$ (Phillips et al., 2019), the relatively few
217 vertices ($|V| < 300$) in the global dataset allowed solutions to be found quickly.

218 Here, we illustrate this method with a fictive global dataset made up from an unbalanced
219 design of five environments (site*year), five crops, and two management levels (Figure 3).
220 When applied on this unbalanced design, this method identified 11 maximal 3-partite cliques,
221 with four examples illustrated in Figure 3. While each of these examples maximized the
222 representativeness of a factor of interest (crop, environment, or management), no factorial
223 design was found with two levels per factor in this fictive dataset.

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



231

232 **Figure 3. Four maximal 3-cliques that represent distinct complete factorial designs**
 233 **within an unbalanced design with three factors.** Black edges represent the edges of the 3-
 234 cliques and gray edges represent the factor combinations appearing in the initial design. Despite
 235 the potential richness of the global dataset, there was no case where two levels of each factor
 236 were combined in a factorial design: network *A* focused on crops, network *B* on environments,
 237 network *C* on management, and network *D* on crop and management together.

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 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. yield, nitrogen fertilization)	All available observations (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. The main advantage of global datasets in biology is that they consist of

259 phenotypic observations, which means that the studied processes are potentially observed
260 at lower levels than in meta-analysis. In this sense, global datasets could enable further in-
261 vestigation of potential causalities based on correlations in the data (Garside and Bell, 2011;
262 Gunawardena, 2014). Additionally, since agronomic global datasets contain plant-related
263 variables measured at multiple organizational levels (e.g. organ, plant, crop), they can target
264 a wide audience for data reuse. For instance, researchers developing functional–structural
265 plant models (Louarn et al., 2020) may be interested in variables measured at the plant
266 scale (e.g. number of tillers, inter-node length, plant height), while those who develop crop
267 models to predict yields (Berghuijs et al., 2021) may be interested in variables measured at
268 the crop scale (e.g. crop biomass, crop height).

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

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

295 **Acknowledgements**

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

303 **Competing Interests**

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

306 **Author Contributions**

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

309 **Data Availability**

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

311 **References**

- 312 Ackoff, R.L. 1989. From data to wisdom. *Journal of applied systems analysis* 16(1): 3–9.
- 313 Allen, I.E., and I. Olkin. 1999. Estimating Time to Conduct a Meta-analysis From Number
314 of Citations Retrieved. *JAMA* 282(7): 634–635. doi: [10.1001/JAMA.282.7.634](https://doi.org/10.1001/JAMA.282.7.634).
- 315 Barillot, R., D. Combes, S. Pineau, P. Huynh, and A.J. Escobar-Gutierrez. 2014. Com-
316 parison of the morphogenesis of three genotypes of pea (*Pisum sativum*) grown in pure
317 stands and wheat-based intercrops. *Aob Plants* 6: plu006. doi: [https://doi.org/10.1093/](https://doi.org/10.1093/aobpla/plu006)
318 [aobpla/plu006](https://doi.org/10.1093/aobpla/plu006).

- 319 Bedoussac, L., and E. Justes. 2010a. Dynamic analysis of competition and complementarity
320 for light and N use to understand the yield and the protein content of a durum wheat–
321 winter pea intercrop. *Plant and Soil* 330(1-2): 37–54. doi: [https://doi.org/10.1007/
322 s11104-010-0303-8](https://doi.org/10.1007/s11104-010-0303-8).
- 323 Bedoussac, L., and E. Justes. 2010b. The efficiency of a durum wheat-winter pea intercrop
324 to improve yield and wheat grain protein concentration depends on N availability during
325 early growth. *Plant and Soil* 330(1-2): 19–35. doi: [https://doi.org/10.1007/s11104-009-
326 0082-2](https://doi.org/10.1007/s11104-009-0082-2).
- 327 Berghuijs, H.N.C., M. Weih, W. Van Der Werf, A.J. Karley, E. Adam, et al. 2021. Calibrat-
328 ing and testing APSIM for wheat-faba bean pure cultures and intercrops across Europe.
329 *Field Crops Research* 264: 108088. doi: [10.1016/j.fcr.2021.108088](https://doi.org/10.1016/j.fcr.2021.108088).
- 330 Broman, K.W., and K.H. Woo. 2018. Data organization in spreadsheets. *The American*
331 *Statistician* 72(1): 2–10. doi: [10.1080/00031305.2017.1375989](https://doi.org/10.1080/00031305.2017.1375989).
- 332 Bron, C., and J. Kerbosch. 1973. Algorithm 457: Finding All Cliques of an Undirected
333 Graph [H]. *Communications of the ACM* 16(9): 575–577. doi: [10.1145/362342.362367](https://doi.org/10.1145/362342.362367).
- 334 Casler, M.D. 2015. Fundamentals of experimental design: Guidelines for designing suc-
335 cessful experiments. *Agronomy Journal* 107(2): 692–705. doi: [https://doi.org/10.2134/
336 agronj2013.0114](https://doi.org/10.2134/agronj2013.0114).
- 337 Corre-Hellou, G., J. Fustec, and Y. Crozat. 2006. Interspecific competition for soil N and
338 its interaction with N-2 fixation, leaf expansion and crop growth in pea-barley intercrops.
339 *Plant and Soil* 282(1-2): 195–208. doi: <https://doi.org/10.1007/s11104-005-5777-4>.
- 340 Cruz, S.M.S. da, and J.A.P. do Nascimento. 2019. Towards integration of data-driven
341 agronomic experiments with data provenance. *Computers and Electronics in Agriculture*
342 161(September 2018): 14–28. doi: [10.1016/j.compag.2019.01.044](https://doi.org/10.1016/j.compag.2019.01.044).
- 343 Duru, M., O. Therond, G. Martin, R. Martin-Clouaire, M.-A. Magne, et al. 2015. How
344 to implement biodiversity-based agriculture to enhance ecosystem services: A review.
345 *Agronomy for Sustainable Development* 35(4): 1259–1281. doi: [10.1007/s13593-015-
346 0306-1](https://doi.org/10.1007/s13593-015-0306-1).
- 347 Garside, A.L., and M.J. Bell. 2011. Growth and yield responses to amendments to the
348 sugarcane monoculture: Towards identifying the reasons behind the response to breaks.
349 *Crop and Pasture Science* 62(9): 776–789. doi: [10.1071/CP11055](https://doi.org/10.1071/CP11055).
- 350 Gaudio, N., R. Mahmoud, L. Bedoussac, E. Justes, E.-P. Journet, et al. 2023. A global
351 dataset gathering 37 field experiments involving cereal-legume intercrops and their cor-
352 responding sole crops. doi: [10.5281/zenodo.8081577](https://doi.org/10.5281/zenodo.8081577).
- 353 Gaudio, N., C. Violle, X. Gendre, F. Fort, R. Mahmoud, et al. 2021. Interspecific inter-
354 actions regulate plant reproductive allometry in cereal–legume intercropping systems.
355 *Journal of Applied Ecology* 58(11): 2579–2589. doi: [https://doi.org/10.1111/1365-
356 2664.13979](https://doi.org/10.1111/1365-2664.13979).

- 357 Glass, G.V. 1976. Primary, secondary, and meta-analysis of research. Educational re-
358 searcher 5(10): 3–8.
- 359 Gunawardena, J. 2014. Models in biology: 'Accurate descriptions of our pathetic thinking'.
360 BMC Biology 12(1): 1–11. doi: [10.1186/1741-7007-12-29/FIGURES/3](https://doi.org/10.1186/1741-7007-12-29/FIGURES/3).
- 361 Hauggaard-Nielsen, H., M. Gooding, P. Ambus, G. Corre-Hellou, Y. Crozat, et al. 2009a.
362 Pea-barley intercropping for efficient symbiotic N₂-fixation, soil N acquisition and use
363 of other nutrients in European organic cropping systems. Field Crops Research 113(1):
364 64–71. doi: <https://doi.org/10.1016/j.fcr.2009.04.009>.
- 365 Hauggaard-Nielsen, H., M. Gooding, P. Ambus, G. Corre-Hellou, Y. Crozat, et al. 2009b.
366 Pea-barley intercropping and short-term subsequent crop effects across European organic
367 cropping conditions. Nutrient Cycling in Agroecosystems 85(2): 141–155. doi: <https://doi.org/10.1007/s10705-009-9254-y>.
- 369 Hauggaard-Nielsen, H., B. Jørnsgaard, J. Kinane, and E.S. Jensen. 2008. Grain legume-
370 cereal intercropping: The practical application of diversity, competition and facilitation
371 in arable and organic cropping systems. Renewable Agriculture and Food Systems 23(1):
372 3–12. doi: <https://doi.org/10.1017/S1742170507002025>.
- 373 Jenkins, G.B., A.P. Beckerman, C. Bellard, A. Benítez-López, A.M. Ellison, et al. 2023. Re-
374 producibility in ecology and evolution: Minimum standards for data and code. Ecology
375 and Evolution 13(5). doi: [10.1002/ece3.9961](https://doi.org/10.1002/ece3.9961).
- 376 Kammoun, B., E.-P. Journet, E. Justes, and L. Bedoussac. 2021. Cultivar Grain Yield
377 in Durum Wheat-Grain Legume Intercrops Could Be Estimated From Sole Crop Yields
378 and Interspecific Interaction Index. Frontiers in Plant Science 12: 2191. doi: <https://doi.org/10.3389/fpls.2021.733705>.
- 380 Kattge, J., S. Diaz, S. Lavorel, I.C. Prentice, P. Leadley, et al. 2011. TRY—a global database
381 of plant traits. Global change biology 17(9): 2905–2935.
- 382 Knudsen, M.T., H. Hauggaard-Nielsen, B. Jørnsgaard, and E.S. Jensen. 2004. Comparison
383 of interspecific competition and N use in pea-barley, faba bean-barley and lupin-barley
384 intercrops grown at two temperate locations. Journal of Agricultural Science 142: 617–
385 627. doi: <https://doi.org/10.1017/S0021859604004745>.
- 386 Launay, M., N. Brisson, S. Satger, H. Hauggaard-Nielsen, G. Corre-Hellou, et al. 2009.
387 Exploring options for managing strategies for pea-barley intercropping using a modeling
388 approach. European Journal of Agronomy 31(2): 85–98. doi: <https://doi.org/10.1016/j.eja.2009.04.002>.
- 390 Lawler, E.L., J.K. Lenstra, and A.H.G. Rinnooy Kan. 1980. Generating All Maximal
391 Independent Sets: NP-Hardness and Polynomial-Time Algorithms. SIAM Journal on
392 Computing 9(3): 558–565. doi: [10.1137/0209042](https://doi.org/10.1137/0209042).
- 393 Licker, R., M. Johnston, J.A. Foley, C. Barford, C.J. Kucharik, et al. 2010. Mind the gap:
394 How do climate and agricultural management explain the 'yield gap' of croplands around

395 the world? *Global Ecology and Biogeography* 19(6): 769–782. doi: [10.1111/j.1466-](https://doi.org/10.1111/j.1466-8238.2010.00563.x)
396 [8238.2010.00563.x](https://doi.org/10.1111/j.1466-8238.2010.00563.x).

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

399 Louarn, G., R. Barillot, Di. Combes, and A. Escobar-Gutiérrez. 2020. Towards intercrop
400 ideotypes: Non-random trait assembly can promote overyielding and stability of species
401 proportion in simulated legume-based mixtures. *Annals of Botany* 126(4): 671–685. doi:
402 [10.1093/aob/mcaa014](https://doi.org/10.1093/aob/mcaa014).

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

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

409 Maat, H. 2011. The history and future of agricultural experiments. *NJAS - Wageningen*
410 *Journal of Life Sciences* 57(3-4): 187–195. doi: [10.1016/j.njas.2010.11.001](https://doi.org/10.1016/j.njas.2010.11.001).

411 Mahmoud, R., P. Casadebaig, N. Hilgert, L. Alletto, G.T. Freschet, et al. 2022. Species
412 choice and n fertilization influence yield gains through complementarity and selection
413 effects in cereal-legume intercrops. *Agronomy for sustainable development*. doi:
414 [10.1007/s13593-022-00754-y](https://doi.org/10.1007/s13593-022-00754-y).

415 Makowski, D., T. Nesme, F. Papy, and T. Doré. 2014. Global agronomy, a new field
416 of research. A review. *Agronomy for Sustainable Development* 34(2): 293–307. doi:
417 [10.1007/s13593-013-0179-0](https://doi.org/10.1007/s13593-013-0179-0).

418 Meunier, C., L. Alletto, L. Bedoussac, J.E. Bergez, P. Casadebaig, et al. 2022. A modelling
419 chain combining soft and hard models to assess a bundle of ecosystem services provided
420 by a diversity of cereal-legume intercrops. *European Journal of Agronomy* 132(October
421 2021). doi: [10.1016/j.eja.2021.126412](https://doi.org/10.1016/j.eja.2021.126412).

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

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

430 Newman, S.J., and R.T. Furbank. 2021. A multiple species, continent-wide, million-
431 phenotype agronomic plant dataset. *Scientific Data* 8(1): 1–8. doi: [10.1038/s41597-](https://doi.org/10.1038/s41597-021-00898-8)
432 [021-00898-8](https://doi.org/10.1038/s41597-021-00898-8).

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

436 Phillips, C.A., K. Wang, E.J. Baker, J.A. Bubier, E.J. Chesler, et al. 2019. On Finding and
437 enumerating maximal and maximum k-partite cliques in k-partite graphs. *Algorithms*
438 12(1). doi: [10.3390/a12010023](https://doi.org/10.3390/a12010023).

439 Popkin, G. 2019. Data sharing and how it can benefit your scientific career. *Nature*
440 569(7756): 445–447. doi: [10.1038/d41586-019-01506-x](https://doi.org/10.1038/d41586-019-01506-x).

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

444 Tang, X., S.A. Placella, F. Dayde, L. Bernard, A. Robin, et al. 2016. Phosphorus availability
445 and microbial community in the rhizosphere of intercropped cereal and legume along a
446 P-fertilizer gradient. *Plant and Soil* 407(1-2): 119–134. doi: [https://doi.org/10.1007/
447 s11104-016-2949-3](https://doi.org/10.1007/s11104-016-2949-3).

448 Tardieu, F. 2020. Educated big data to study sensitivity to drought. *Nature Food* 1(11):
449 669–670. doi: [10.1038/s43016-020-00187-4](https://doi.org/10.1038/s43016-020-00187-4).

450 Viguier, L., L. Bedoussac, E.-P. Journet, and E. Justes. 2018. Yield gap analysis extended
451 to marketable grain reveals the profitability of organic lentil-spring wheat intercrops.
452 *Agronomy for Sustainable Development* 38(4): 39. doi: [https://doi.org/10.1007/s13593-
453 018-0515-5](https://doi.org/10.1007/s13593-018-0515-5).

454 White, J.W., and F.K. Van Evert. 2008. Publishing agronomic data. *Agronomy Journal*
455 100(5): 1396–1400. doi: [10.2134/agronj2008.0080F](https://doi.org/10.2134/agronj2008.0080F).

456 Wickham, H. 2014. Tidy data. *Journal of Statistical Software* 59(10).

457 Wickham, H., and G. Grolemund. 2016. *R for data science: Import, tidy, transform,
458 visualize, and model data.* ” O’Reilly Media, Inc.”.

459 Wilkinson, M.D., M. Dumontier, Ij.J. Aalbersberg, G. Appleton, M. Axton, et al. 2016.
460 Comment: The FAIR Guiding Principles for scientific data management and stewardship.
461 *Scientific Data* 3: 1–9. doi: [10.1038/sdata.2016.18](https://doi.org/10.1038/sdata.2016.18).

462 Wilson, G., J. Bryan, K. Cranston, J. Kitzes, L. Nederbragt, et al. 2017. Good enough
463 practices in scientific computing. *PLOS Computational Biology* 13(6): e1005510. doi:
464 [10.1371/JOURNAL.PCBI.1005510](https://doi.org/10.1371/JOURNAL.PCBI.1005510).

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