



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. 2024. hal-04145269v2

HAL Id: hal-04145269

<https://hal.inrae.fr/hal-04145269v2>

Preprint submitted on 8 Feb 2024 (v2), 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 - 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 **ORCID iDs**

8 Rémi Mahmoud - <https://orcid.org/0000-0003-0853-0834>

9 Pierre Casadebaig - <https://orcid.org/0000-0001-7225-936X>

10 Noémie Gaudio - <https://orcid.org/0000-0002-4528-9851>

11 **Abstract**

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

22 Introduction

23 Field experiments, whether conducted on farms or at experimental research stations, have
24 traditionally been the primary approach for acquiring knowledge in crop sciences (Maat,
25 2011). Yet, extrapolating applicable principles from localized experiments remains a chal-
26 lenging task (Makowski et al., 2014). To derive general rules about agroecosystem function-
27 ing, meta-analysis, *i.e.* a “statistical analysis of a large collection of analysis results from
28 individual studies to integrate the findings” (Glass, 1976), is typically employed. Alter-
29 natively, global datasets, corresponding to the aggregation of observations from numerous
30 experiments, can serve as another valuable tool for analyzing agronomic data. While the
31 use of meta-analysis to report results is growing in crop science, it is not a mainstream
32 analysis method compared to reports based on a repeated (years) set of one or two field
33 trials. Distinguishing themselves from meta-analyses, global datasets compile raw experi-
34 mental results on a detailed scale, such as repeated measurements on individuals or multiple
35 state variables on the canopy. In contrast, meta-analysis is typically restricted to published
36 results with a limited set of variables.

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

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

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

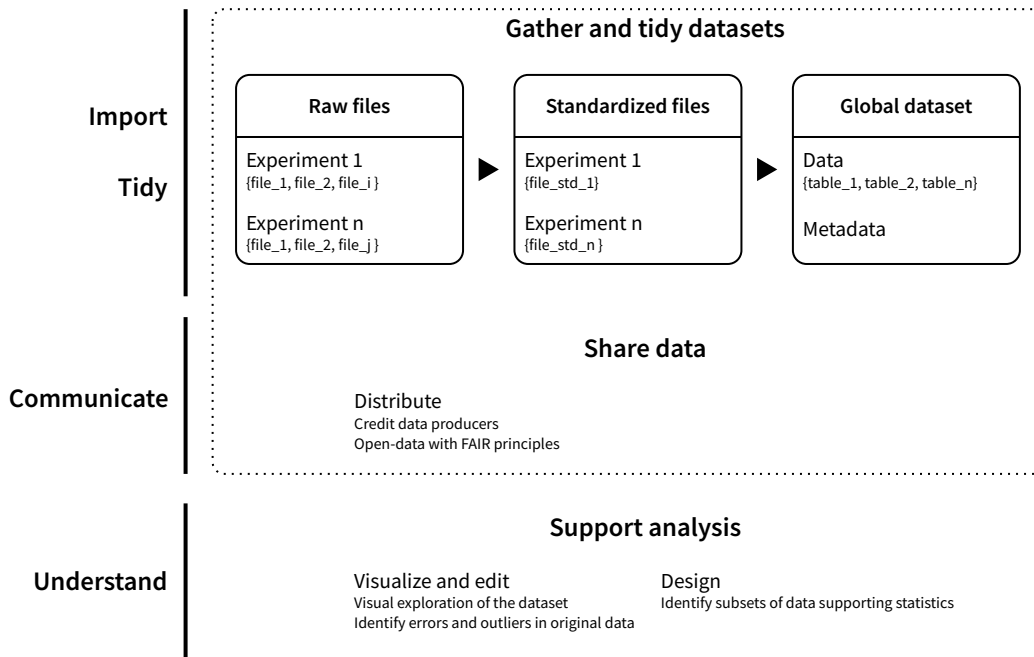
63 Despite these challenges, crop science would greatly benefit from the study of global datasets
64 combining multiple experiments (White and Van Evert, 2008; Zamir, 2013; Cruz and Nasci-
65 mento, 2019). This approach is particularly relevant considering the current agricultural
66 landscape, where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

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

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

85 **Design steps of global datasets**

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



90

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

96 1. Gather and tidy source datasets

97 1.1. Conceptual framework

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

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

106 Then, a basic database model for the global dataset has to be developed. This step in-
 107 volves defining the structure of a database, including the number of tables needed and the

108 relationships between them. It also involves describing the metadata, such as the variables
109 measured or collected, their definitions, and units.

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

122 **1.2. Case study**

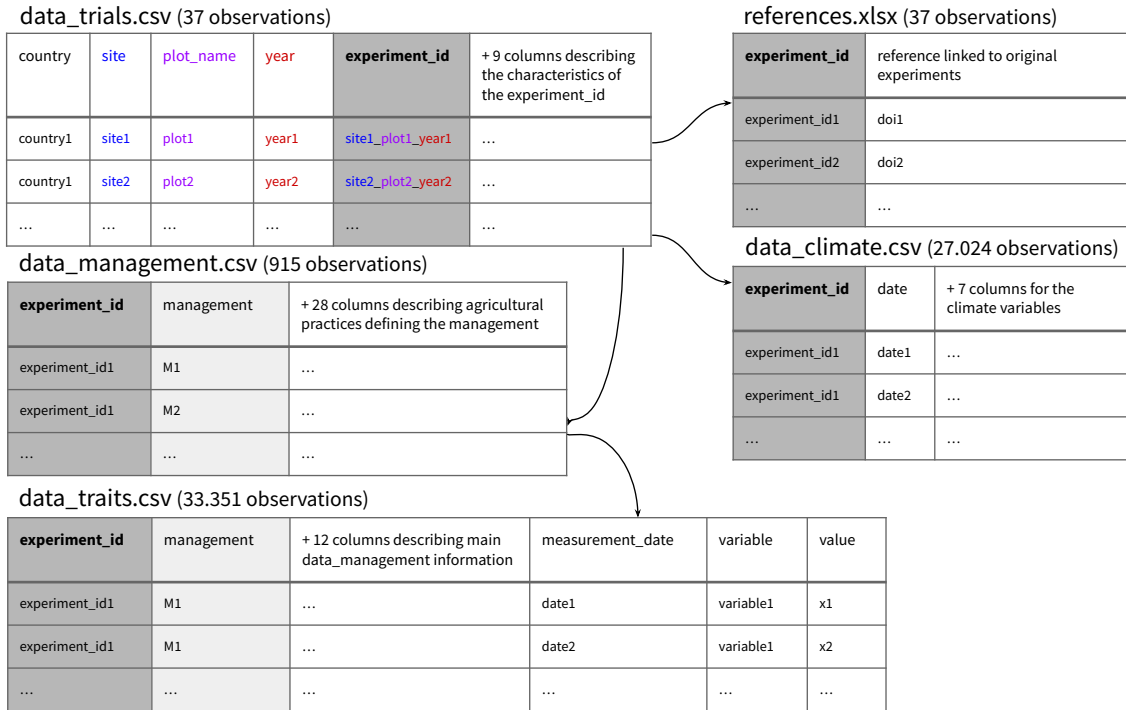
123 Although combining results from a few experiments (usually two years, often sequential)
124 is common in the intercropping literature (and more generally in crop science), no study
125 includes joint analysis of dozens of experiments to infer more generic results about intercrop-
126 ping functioning.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

136 To gather the 37 experiments, six research teams were contacted. For each experiment,
137 several spreadsheet files (all in Excel format) were retrieved, ranging from 1 to 10 per
138 experiment. These files differed by the number of sheets they contained, ranging from 1
139 to 67. We finally collected a total of 86 excel files (412 sheets). These raw data were
140 highly heterogeneous at all levels, whether concerning the variables (*e.g.* type, name, unit,
141 measured scale) or the format of the file itself (*e.g.* one sheet per date or per variable,

142 different tables on a same sheet, calculations and graphs mixed with raw data cells, different
143 languages and encoding format).

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

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



165

166

167

168

169

170

171

172

173

174

175

176

177

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).

178
179
180
181
182
183
184
185
186
187

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	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	Green, Yellow, Cyan	Cyan
Jynde vad_jyn_2002	24	12	Red	Green, Yellow, Cyan	Cyan
Jynde vad_jyn_2001	24	12	Red	Green, Yellow, Cyan	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	Green, Yellow, Cyan	Cyan
Copenhagen_hbg_2002	24	11	Red	Green, Yellow, Cyan	Cyan
Copenhagen_hbg_2001	24	12	Red	Green, Yellow, Cyan	Cyan
Auz_ZN_2012	58	24	Red	Cyan, Purple, Pink	Cyan
Auz_TO_2016	86	18	Red	Green	Cyan
Auz_TO_2013	93	24	Red	Cyan, Purple, Pink	Cyan
Auz_TE_2006	13	20	Red	Cyan, Purple, Pink	Red, Cyan
Auz_SGs_2007	66	23	Red	Cyan, Purple, Pink	Red, Cyan
Auz_PP_2011	20	20	Red	Cyan, Purple, Pink	Red, Cyan
Auz_pk_2011	18	18	Red	Cyan, Purple, Pink	Red, Cyan
Auz_marinette_2_2015	85	13	Red	Green	Cyan
Auz_marinette_1_2015	22	13	Red	Cyan, Purple, Pink	Cyan
Auz_cochard_2010	60	21	Red	Cyan, Blue, Purple, Pink	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	Red, Cyan
Angers_thorigne_2003	6	10	Red	Green	Red, 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

188

189 **2. Share organized data**

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

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

201 This global dataset, as well as the metadata associated, are available on a data repository
202 in a FAIR way (Gaudio et al., 2023). Out of the 37 experiments gathered, 11 have never
203 been valued before. Additional details on experimental designs and management practices
204 are reported in the reference publications for 26 of the 37 experiments (Knudsen et al., 2004;
205 Corre-Hellou et al., 2006; Hauggaard-Nielsen et al., 2008; Hauggaard-Nielsen et al., 2009a;
206 b; Launay et al., 2009; Bedoussac and Justes, 2010a; b; Naudin et al., 2010, 2014; Barillot
207 et al., 2014; Pelzer et al., 2016; Tang et al., 2016; Viguier et al., 2018; Kammoun et al.,
208 2021).

209 **3. Support new analysis**

210 **3.1. Conceptual framework**

211 Once the data are in a tractable format, visual exploration allows for a comprehensive
212 overview of data patterns, aiding in the identification of anomalies such as errors and outliers
213 that may not be immediately apparent through numerical analysis alone. Later, additional
214 processes are required to render the dataset operational for analytical and modeling studies,
215 such as data imputation, dimension reduction, or data normalization. Because these steps
216 depend largely on the chosen analytical workflow, they are not directly included in the com-
217 municated open datasets, but rather tailored by the subsequent analytical team (Figure 1).
218 Nonetheless, sharing methods can support the future reuse of the dataset. In our case in
219 crop ecology, we illustrated this step with the development of an original method aiming at
220 identifying subsets in the overall dataset corresponding to complete factorial designs.

221 **3.2. Case study**

222 **Method**

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

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

236 **Table 2. Definitions in graph theory used in the present study.**

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>i.e.</i> $\tilde{V} \subseteq V$) and whose edge set (\tilde{E}) is included in the edge set of G (<i>i.e.</i> $\tilde{E} \subseteq E$)
<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 non-empty, 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

237 Given a set of categorical variables X_1, \dots, X_k , each having values in a discrete set (*i.e.*
238 $\forall i = 1, \dots, k X_i \in \mathcal{A}_i := \{x_{i,1}, \dots, x_{i,j_i}\}$, ($j_i \in \mathbb{N}^*$ denoting the number of levels of variable
239 X_i)), a k -partite graph can be derived by setting $V = \bigcup_{i=1}^k \mathcal{A}_i$ (*i.e.* each level of each factor
240 is a vertex) and $E = \{(x, y) \mid \text{levels } x \text{ and } y \text{ observed together}\}$.

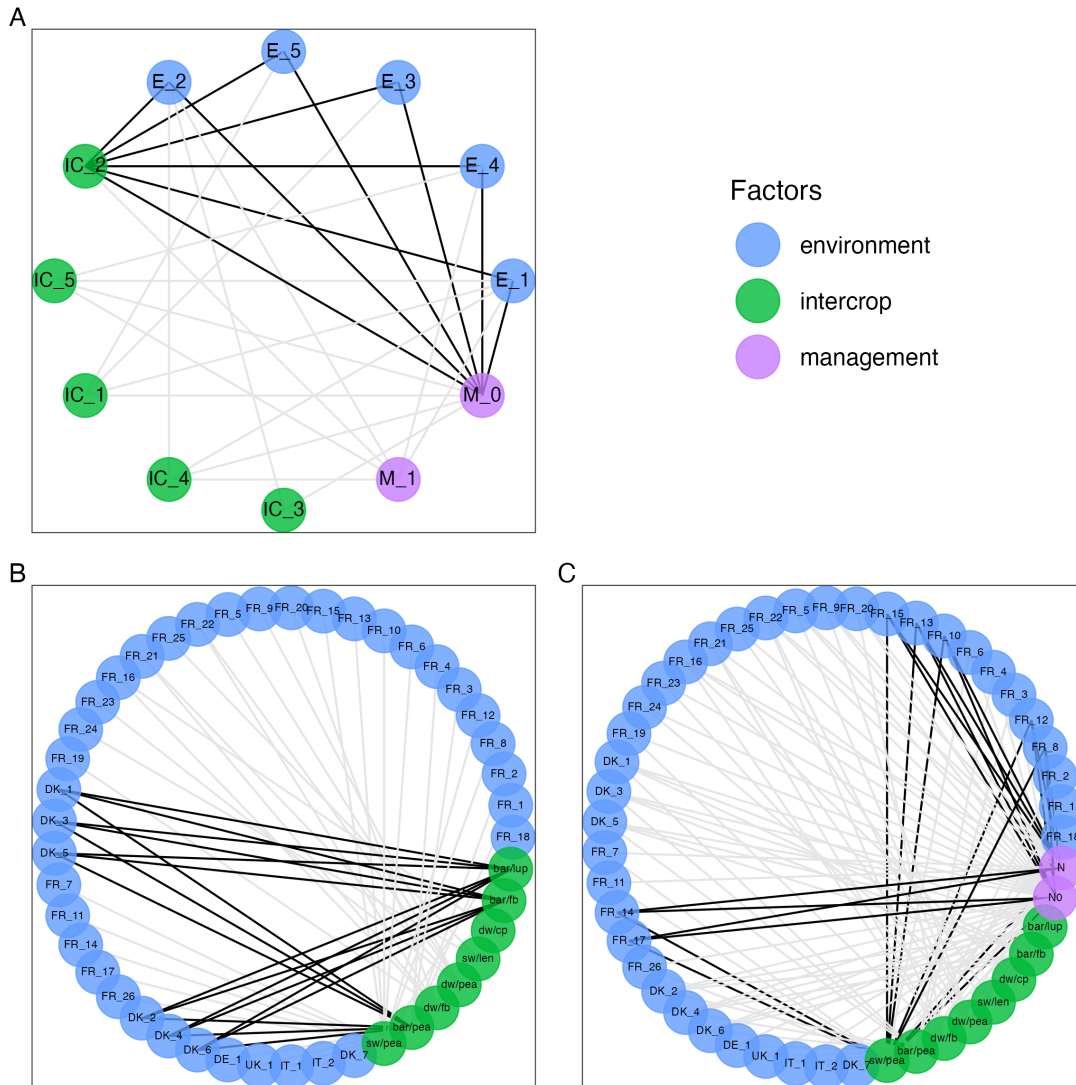
241 A factorial design is complete if, and only if, all possible combinations of the factor levels are
242 present. For a graph $G = (V, E)$, this is equivalent to identifying a subgraph with an edge
243 between each pair of vertices from independent sets (*i.e.* a k -clique). Thus, the challenge
244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

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

260 **Application**

261 Here, we illustrate this method with two datasets : (1) a theoretical one, where we generated
262 an unbalanced design of five environments, five intercrops, and two management levels
263 (Figure 3A); and (2) a practical one, corresponding to the global dataset presented in this
264 study (Figure 3B and 3C).

265 When applied on the theoretical unbalanced design (Figure 3A), this method identified
266 8 maximal 3-partite cliques, each of these designs having different number of modalities
267 in considered factors (environment, intercrops or management). There is only one design
268 maximizing the number of environments, and no factorial design was found with two levels
269 per factor.



270

271

272

273

274

275

276

277

278

279

280

281

Figure 3. Three maximal k -cliques that represent distinct complete factorial designs within theoretical (A) and experimental (B-C) unbalanced designs. Black edges represent the edges of the cliques and gray edges represent the factor combinations appearing in the initial design. In the case A, 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 and 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 *NO* for no fertilization, and *N* for fertilization.

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

291 Discussion

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

301 **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 (<i>e.g.</i> yield, nitrogen fertilization)	All available observations (<i>e.g.</i> 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

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

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

323 Alternatively, global datasets might have a role in increasing the discovery and use of non-
324 published experimental data. In our case, almost 30% of the experimental data gathered
325 have not been published through a research article. Bringing them together with other
326 experiments valued the time and energy required to conduct those field experiments. It
327 was also a friction point, since researchers may be reluctant to share unpublished data. For
328 instance, in our use case, 11 of the 37 experiments were not included in published articles
329 or database before this initiative, while each is now described within the global dataset
330 (Gaudio et al., 2023) and linked back groups leading field experiments in 1-4 scientific
331 publications (Gaudio et al., 2021; Louarn et al., 2021; Mahmoud et al., 2022; Meunier et al.,
332 2022). Based on the global dataset developed in this study, Gaudio et al. (2021) extracted
333 a subset of 28 experiments to assess the influence of intercropping on the relation between
334 plant biomass and grain yield; Louarn et al. (2021) extracted a subset of 15 experiments to
335 validate the adaptation of Nitrogen Nutrition Index (NNI) to intercropping; Mahmoud et al.
336 (2022) extracted a subset of 11 experiments to assess the influence of nitrogen fertilization
337 on plant-plant interactions in intercrops; and Meunier et al. (2022) extracted a subset of 31

338 experiments to calibrate a statistical model used in a modeling chain to predict ecosystem
339 services as a function of the species associated in cereal-legume intercrops.

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

349 **Acknowledgements**

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

355 **Funding**

356 This research was supported by the French National Research Agency under the Investments
357 for the Future Program (referred to as ANR-16-CONV-0004 and ANR-20-PCPA-0006) and
358 by the European Research Council under the European Union's Horizon Europe research
359 and innovation program in the framework of the IntercropValuES (Developing Intercropping
360 for agrifood Value chains and Ecosystem Services delivery in Europe and Southern countries,
361 <https://intercropvalues.eu/>) starting from November 2022 [grant number 101081973].

362 **Conflict of interest disclosure**

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

365 **Author Contributions**

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

368 **Data Availability**

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

370 **References**

- 371 Ackoff, R.L. 1989. From data to wisdom. *Journal of applied systems analysis* 16(1): 3–9.
- 372 Allen, I.E., and I. Olkin. 1999. Estimating Time to Conduct a Meta-analysis From Number
373 of Citations Retrieved. *JAMA* 282(7): 634–635. doi: [10.1001/JAMA.282.7.634](https://doi.org/10.1001/JAMA.282.7.634).
- 374 Barillot, R., D. Combes, S. Pineau, P. Huynh, and A.J. Escobar-Gutierrez. 2014. Com-
375 parison of the morphogenesis of three genotypes of pea (*Pisum sativum*) grown in pure
376 stands and wheat-based intercrops. *Aob Plants* 6: plu006. doi: [https://doi.org/10.1093/
377 aobpla/plu006](https://doi.org/10.1093/aobpla/plu006).
- 378 Bedoussac, L., and E. Justes. 2010a. Dynamic analysis of competition and complementarity
379 for light and N use to understand the yield and the protein content of a durum wheat–
380 winter pea intercrop. *Plant and Soil* 330(1-2): 37–54. doi: [https://doi.org/10.1007/
381 s11104-010-0303-8](https://doi.org/10.1007/s11104-010-0303-8).
- 382 Bedoussac, L., and E. Justes. 2010b. The efficiency of a durum wheat-winter pea intercrop
383 to improve yield and wheat grain protein concentration depends on N availability during
384 early growth. *Plant and Soil* 330(1-2): 19–35. doi: [https://doi.org/10.1007/s11104-009-
385 0082-2](https://doi.org/10.1007/s11104-009-0082-2).
- 386 Berghuijs, H.N.C., M. Weih, W. Van Der Werf, A.J. Karley, E. Adam, et al. 2021. Calibrat-
387 ing and testing APSIM for wheat-faba bean pure cultures and intercrops across Europe.
388 *Field Crops Research* 264: 108088. doi: [10.1016/j.fcr.2021.108088](https://doi.org/10.1016/j.fcr.2021.108088).
- 389 Broman, K.W., and K.H. Woo. 2018. Data organization in spreadsheets. *The American*
390 *Statistician* 72(1): 2–10. doi: [10.1080/00031305.2017.1375989](https://doi.org/10.1080/00031305.2017.1375989).
- 391 Bron, C., and J. Kerbosch. 1973. Algorithm 457: Finding All Cliques of an Undirected
392 Graph [H]. *Communications of the ACM* 16(9): 575–577. doi: [10.1145/362342.362367](https://doi.org/10.1145/362342.362367).
- 393 Casler, M.D. 2015. Fundamentals of experimental design: Guidelines for designing suc-
394 cessful experiments. *Agronomy Journal* 107(2): 692–705. doi: [https://doi.org/10.2134/
395 agronj2013.0114](https://doi.org/10.2134/agronj2013.0114).

- 396 Corre-Hellou, G., J. Fustec, and Y. Crozat. 2006. Interspecific competition for soil N and
397 its interaction with N-2 fixation, leaf expansion and crop growth in pea-barley intercrops.
398 *Plant and Soil* 282(1-2): 195–208. doi: <https://doi.org/10.1007/s11104-005-5777-4>.
- 399 Cruz, S.M.S. da, and J.A.P. do Nascimento. 2019. Towards integration of data-driven
400 agronomic experiments with data provenance. *Computers and Electronics in Agriculture*
401 161(September 2018): 14–28. doi: [10.1016/j.compag.2019.01.044](https://doi.org/10.1016/j.compag.2019.01.044).
- 402 Duru, M., O. Therond, G. Martin, R. Martin-Clouaire, M.-A. Magne, et al. 2015. How
403 to implement biodiversity-based agriculture to enhance ecosystem services: A review.
404 *Agronomy for Sustainable Development* 35(4): 1259–1281. doi: [10.1007/s13593-015-0306-1](https://doi.org/10.1007/s13593-015-0306-1).
- 406 Garside, A.L., and M.J. Bell. 2011. Growth and yield responses to amendments to the
407 sugarcane monoculture: Towards identifying the reasons behind the response to breaks.
408 *Crop and Pasture Science* 62(9): 776–789. doi: [10.1071/CP11055](https://doi.org/10.1071/CP11055).
- 409 Gaudio, N., R. Mahmoud, L. Bedoussac, E. Justes, E.-P. Journet, et al. 2023. A global
410 dataset gathering 37 field experiments involving cereal-legume intercrops and their cor-
411 responding sole crops. doi: [10.5281/zenodo.8081577](https://doi.org/10.5281/zenodo.8081577).
- 412 Gaudio, N., C. Violle, X. Gendre, F. Fort, R. Mahmoud, et al. 2021. Interspecific inter-
413 actions regulate plant reproductive allometry in cereal–legume intercropping systems.
414 *Journal of Applied Ecology* 58(11): 2579–2589. doi: <https://doi.org/10.1111/1365-2664.13979>.
- 416 Glass, G.V. 1976. Primary, secondary, and meta-analysis of research. *Educational re-
417 searcher* 5(10): 3–8.
- 418 Gunawardena, J. 2014. Models in biology: 'Accurate descriptions of our pathetic thinking'.
419 *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).
- 420 Hauggaard-Nielsen, H., M. Gooding, P. Ambus, G. Corre-Hellou, Y. Crozat, et al. 2009a.
421 Pea-barley intercropping for efficient symbiotic N-2-fixation, soil N acquisition and use
422 of other nutrients in European organic cropping systems. *Field Crops Research* 113(1):
423 64–71. doi: <https://doi.org/10.1016/j.fcr.2009.04.009>.
- 424 Hauggaard-Nielsen, H., M. Gooding, P. Ambus, G. Corre-Hellou, Y. Crozat, et al. 2009b.
425 Pea-barley intercropping and short-term subsequent crop effects across European organic
426 cropping conditions. *Nutrient Cycling in Agroecosystems* 85(2): 141–155. doi: <https://doi.org/10.1007/s10705-009-9254-y>.
- 428 Hauggaard-Nielsen, H., B. Jørnsgaard, J. Kinane, and E.S. Jensen. 2008. Grain legume–
429 cereal intercropping: The practical application of diversity, competition and facilitation
430 in arable and organic cropping systems. *Renewable Agriculture and Food Systems* 23(1):
431 3–12. doi: <https://doi.org/10.1017/S1742170507002025>.
- 432 Jenkins, G.B., A.P. Beckerman, C. Bellard, A. Benítez-López, A.M. Ellison, et al. 2023. Re-
433 producibility in ecology and evolution: Minimum standards for data and code. *Ecology*

434 and Evolution 13(5). doi: [10.1002/ece3.9961](https://doi.org/10.1002/ece3.9961).

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

438

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

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

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

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

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

455 Licker, R., M. Johnston, J.A. Foley, C. Barford, C.J. Kucharik, et al. 2010. Mind the gap:
456 How do climate and agricultural management explain the 'yield gap' of croplands around
457 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)
458 [8238.2010.00563.x](https://doi.org/10.1111/j.1466-8238.2010.00563.x).

459 Lobell, D.B., J.M. Deines, and S.D. Tommaso. 2020. Changes in the drought sensitivity of
460 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).

461 Louarn, G., R. Barillot, Di. Combes, and A. Escobar-Gutiérrez. 2020. Towards intercrop
462 ideotypes: Non-random trait assembly can promote overyielding and stability of species
463 proportion in simulated legume-based mixtures. *Annals of Botany* 126(4): 671–685. doi:
464 [10.1093/aob/mcaa014](https://doi.org/10.1093/aob/mcaa014).

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

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

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

472 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).

473 Mahmoud, R., P. Casadebaig, N. Hilgert, L. Alletto, G.T. Freschet, et al. 2022. Species
474 choice and n fertilization influence yield gains through complementarity and selection
475 effects in cereal-legume intercrops. *Agronomy for sustainable development*. doi:
476 [10.1007/s13593-022-00754-y](https://doi.org/10.1007/s13593-022-00754-y).

477 Makowski, D., T. Nesme, F. Papy, and T. Doré. 2014. Global agronomy, a new field
478 of research. A review. *Agronomy for Sustainable Development* 34(2): 293–307. doi:
479 [10.1007/s13593-013-0179-0](https://doi.org/10.1007/s13593-013-0179-0).

480 Meunier, C., L. Alletto, L. Bedoussac, J.E. Bergez, P. Casadebaig, et al. 2022. A modelling
481 chain combining soft and hard models to assess a bundle of ecosystem services provided
482 by a diversity of cereal-legume intercrops. *European Journal of Agronomy* 132(October
483 2021). doi: [10.1016/j.eja.2021.126412](https://doi.org/10.1016/j.eja.2021.126412).

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

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

491

492 Newman, S.J., and R.T. Furbank. 2021. A multiple species, continent-wide, million-
493 phenotype agronomic plant dataset. *Scientific Data* 8(1): 1–8. doi: [10.1038/s41597-
494 021-00898-8](https://doi.org/10.1038/s41597-021-00898-8).

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

498 Phillips, C.A., K. Wang, E.J. Baker, J.A. Bubier, E.J. Chesler, et al. 2019. On Finding and
499 enumerating maximal and maximum k-partite cliques in k-partite graphs. *Algorithms*
500 12(1). doi: [10.3390/a12010023](https://doi.org/10.3390/a12010023).

501 Popkin, G. 2019. Data sharing and how it can benefit your scientific career. *Nature*
502 569(7756): 445–447. doi: [10.1038/d41586-019-01506-x](https://doi.org/10.1038/d41586-019-01506-x).

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

506 Sparks, A.H. 2018. Nasapower: A NASA POWER Global Meteorology, Surface Solar En-
507 ergy and Climatology Data Client for R. *Journal of Open Source Software* 3(30): 1035.
508 doi: [10.21105/joss.01035](https://doi.org/10.21105/joss.01035).

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

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

513 Tardieu, F. 2020. Educated big data to study sensitivity to drought. *Nature Food* 1(11):
514 669–670. doi: [10.1038/s43016-020-00187-4](https://doi.org/10.1038/s43016-020-00187-4).

515 Viguier, L., L. Bedoussac, E.-P. Journet, and E. Justes. 2018. Yield gap analysis extended
516 to marketable grain reveals the profitability of organic lentil-spring wheat intercrops.
517 *Agronomy for Sustainable Development* 38(4): 39. doi: [https://doi.org/10.1007/s13593-
518 018-0515-5](https://doi.org/10.1007/s13593-018-0515-5).

519 White, J.W., and F.K. Van Evert. 2008. Publishing agronomic data. *Agronomy Journal*
520 100(5): 1396–1400. doi: [10.2134/agronj2008.0080F](https://doi.org/10.2134/agronj2008.0080F).

521 Wickham, H. 2014. Tidy data. *Journal of Statistical Software* 59(10). doi: [10.18637/jss.v059.i10](https://doi.org/10.18637/jss.v059.i10).

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

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

527 Wilson, G., J. Bryan, K. Cranston, J. Kitzes, L. Nederbragt, et al. 2017. Good enough
528 practices in scientific computing. *PLOS Computational Biology* 13(6): e1005510. doi:
529 [10.1371/JOURNAL.PCBI.1005510](https://doi.org/10.1371/JOURNAL.PCBI.1005510).

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