The International Society of Precision Agriculture presents the
15th International Conference on
Precision Agriculture
26–29 JUNE 2022
Minneapolis Marriott City Center | Minneapolis, Minnesota USA

From fragmented data to unified insights: leveraging data standardization tools for better collaboration and agronomic big data analysis

Shai Sela
Chief Scientist, Agmatix (sela.shai@agmatix.com)
Sapir tower, Tuval 40, Ramat Gan, Israel

A paper from the Proceedings of the
15th International Conference on Precision Agriculture
June 26-29, 2022
Minneapolis, Minnesota, United States

Abstract.

The quantity and scope of agronomic data available for researchers in both industry and academia is increasing rapidly. Data sources include a myriad of different streams, such as field experiments, sensors, climatic data, socioeconomic data or remote sensing. The lack of standards and workflows frequently leads agronomic data to be fragmented and siloed, hampering collaboration efforts. Implementing data standardization schemes can enable efficient collaboration, and leveraging the collective power of the research community to address critical agronomic knowledge gaps.

This presentation will provide an overview of available research data standardization tools and explain the underlying FAIR and other data management principles. Using Agmatix's standardization platform as an example, we will demonstrate how data from multiple sources can be standardized and used for insightful modeling. We have used 3774 experimental data points from different sources in the United States – universities, commercial associations and farm-management systems - to construct a corn prediction model. Production environment descriptors such as nutrients inputs, soil texture, soil organic matter and planting dates were augmented with relevant climatic data retrieved for specific growth stages periods in each experiment. The model, built as an ensemble of decision trees, was able to achieve good accuracy in yield prediction across the different production environments (R2 = 0.9. RMSE = 1.0 Mg/Ha, mean absolute prediction error of 7%). Standard feature permutation procedures found the key factors affecting the model: nutrients inputs, soil organic matter, soil texture, previous crop, and climatic conditions during two critical crop growth periods. Analyzing the data for the effect of these factors found informative yield trends, which can now be further explored.

While demonstrated here on corn yield, many different agronomic research domains can benefit from standardization of data. We call on the agronomic research community to adopt standardization tools, and share their data through public repositories, community-managed or private-public data platforms. This will allow better transparency, enable collaborative efforts, and an increased potential for tackling the current global agronomic challenges.

Keywords.
Data standardization; Maize, Modeling; Machine learning; Agmatix

Sela, S. (2022). From fragmented data to unified insights: leveraging data standardization tools for better collaboration and agronomic big data analysis. Proceedings of the 15th International Conference on Precision Agriculture (unpaginated, online). Monticello, IL: International Society of Precision Agriculture.

1. Introduction

Agronomic production is affected by many factors, such as soil type, weather, pests, nutrients availability and many other factors. Understanding the interactions between these factors, to enable system optimization, requires the coverage of large parameter spaces. This in turn requires large quantities of data, leading the agronomic research sphere to be data driven. Agronomic research has become an ever increasing collaborative effort – between stakeholders in academia, government and non-governmental organizations. Typical agronomic study will involve utilizing data from a myriad of sources, such as field and lab experiments, remote sensing, soil sensors, machinery data, weather data, or data from previous research efforts.

The lack of data standards, and in many cases the collection of data in a manual way, creates large variability in file types and parameter naming. This can occur even within the same research group, and more often between research groups. Researchers tend to name data in ways understandable to them, and frequently adequate metadata is missing. Research personnel change, and recovering lost metadata involves tremendous amount of effort. Often, the end result is data fragmentation and dispersion, reduced usability of past research efforts, limited ability to collaborate with other researchers, and loss of potential new insights. Manually unifying data from different sources into a single database is a time consuming task for researchers. The way forward would be to adopt data standardization solutions.

There are several open data networks available for agronomy researchers to explore and share research data, and in some cases, offer tools to aid in standardization of data according to the FAIR principles. These include (among others) CGIAR GARDIAN: (https://gardian.bigdata.cgiar.org/#/), Global Open Data for Agriculture and Nutrition (GODAN, https://www.godan.info/), DataONE (https://www.dataone.org/), API-Agro (https://api-agro.eu/en/), or Agmatix (www.agmatix.com).

Data standardization schemes should follow the FAIR guiding principles. In essence, the data should be (Wilkinson et al. 2016):

i) Findable: data are described with sufficient meta-data, which is assigned a unique identifier. The meta-data file should be indexed within a searchable resource.

ii) Accessible: meta-data can be retrieved using a standardized communication protocol. Meta-data should be accessible even if the data is no longer available.

iii) Interoperable: meta-data should use a formal and broadly applicable language for knowledge representation.

iv) Reusable: Meta-data are sufficiently described, with relevant attributes. Meta data should have a clear data usage license, and meet the domain community standards.

The number of datasets that follow the FAIR principals is increasing. Efforts are made to quantify and report the FAIRness of data sets within data management platforms (i.e. Jones et al. 2019). While the findability and accessibility aspects of the data are easier to achieved, the interoperability and reusability aspects of data remain a challenge. An extensive recent review on FAIR implementation in agriculture and food data found difficulties in FAIR implementation due to (among others), lack in shared vocabularies and data sharing handling practices (Top et al. 2022).

The objective of this paper is to demonstrate the usefulness of data standardization platforms, and show how they can be used to unify data from multiple sources and generate new insights. This is exemplified here using Agmatix's standardization platform. Unified data will subsequently be used as input to a machine learning model, to predict corn yield across different production environments in the US.

2. Methods

2.1 Agmatix platform, and the GUARDS protocol

Agmatix is a private agronomic data standardization platform (www.agmatix.com). The platform enables unification of any data source (tabular, word documents, pdf's etc.) into a secure database, where collaborators can query the data and perform statistical analyses.

The core of the platform is an extensive library of agronomic ontologies developed by Agmatix, called GUARDS – Growing Universal Agronomic Research Data Standard. The GUARDS library borrows and extends upon ontologies from several publically available ontologies libraries, such as Agro (Aubert et al. 2017), ENVO (Buttigieg et al. 2013), and PPO (Stucky et al. 2018). The GUARDS library covers many domains related to experimental setup and the agriculture production environment, including research metadata, soil, water (irrigation), chemical and biological inputs, sampling protocols, climate and more. The ontologies are generated in a bottom-up approach, allowing flexibility in accommodation of any data type. The relation between the ontology entities is mapped and recorded, and the GUARDS library currently consist of more than 4000 entities and 200 relation types. In the ingestion process, the data is curated and outliers are identified.

2.2 Data sources

Corn crop data used for the analysis came from different sources (Figure 1):

i) Supplemental data of a large meta-analysis study of agriculture Nitrous oxide emissions, published by Eagle et al. (2020). Individual references within the meta-data study are reported in the legend of Figure 1.

ii) Experimental corn data provided courtesy of Achim Dobermann of the International Fertilizer Association (IFA).

iii) Data on long term corn experiments, published by Puntel et al. (2016).

iv) Data from multiple corn N response trials provided courtesy of Michael Castellano (ISU).

v) Data from farm management systems.

All data files were ingested and standardized to a single database by the Agmatix platform. Each dataset on its own has many potential explanatory parameters, some of which are very detailed. We were limited however to the common parameters available across all datasets: total N applied, previous crop, soil texture, soil organic matter, and irrigation data (if available). This basic set of explanatory parameters was enriched with weather data: precipitation and temperature (source: Awhere inc, 4X4 km resolution) for specific corn growth stages. A corn-growing model based on Growing Degree Days (Nielsen 2019) was used to delineate the growing season of each experiment into 7 windows of interest:

i) Planting to VE

ii) VE to V3

iii) V3 to V10

iv) V10 to tasseling

v) Tasseling to 10 days past tasseling

vi) 10-20 days past tasseling

vii) 20-30 days past tasseling

As planting dates differed between experiments, these windows of interest occurred in different dates in each experiment. For each window of interest, the mean GDD and the total water availability (precipitation + irrigation) was calculated for all experimental data. Altogether, 3774 yield observations and 19 inputs were available for model development. The dataset spanned a wide range of corn yield values (Figure 2).

2.3 Modeling approach

A XGBoost algorithm implemented in R (Chen & Guestrin 2016) was used to construct a yield prediction model. To ensure good coverage of the yield parameter space, a stratified sampling approach was applied to split the data to training (80%) and testing (20%) datasets. After the model was calibrated, it was validated using the independent test data set. Model efficiency was quantified using five indexes: i) R2; ii) Mean Absolute Prediction Error (MAPE, [%]); iii) Root Mean Square Error (RMSE, [ppm]); iv) Normalized RMSE (RMSE divided by the range of observations [%]); and v) Mean Absolute Deviation (MAD).

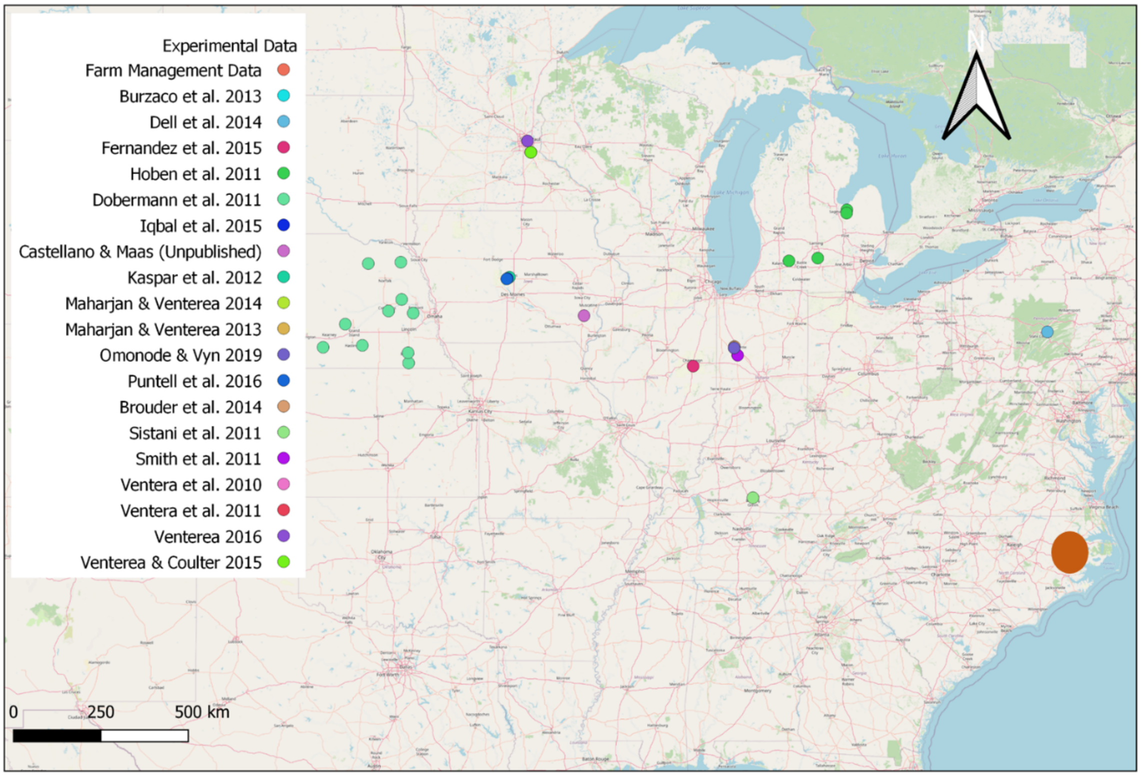

Figure 1. Locations and sources of the experimental data used for the analysis.

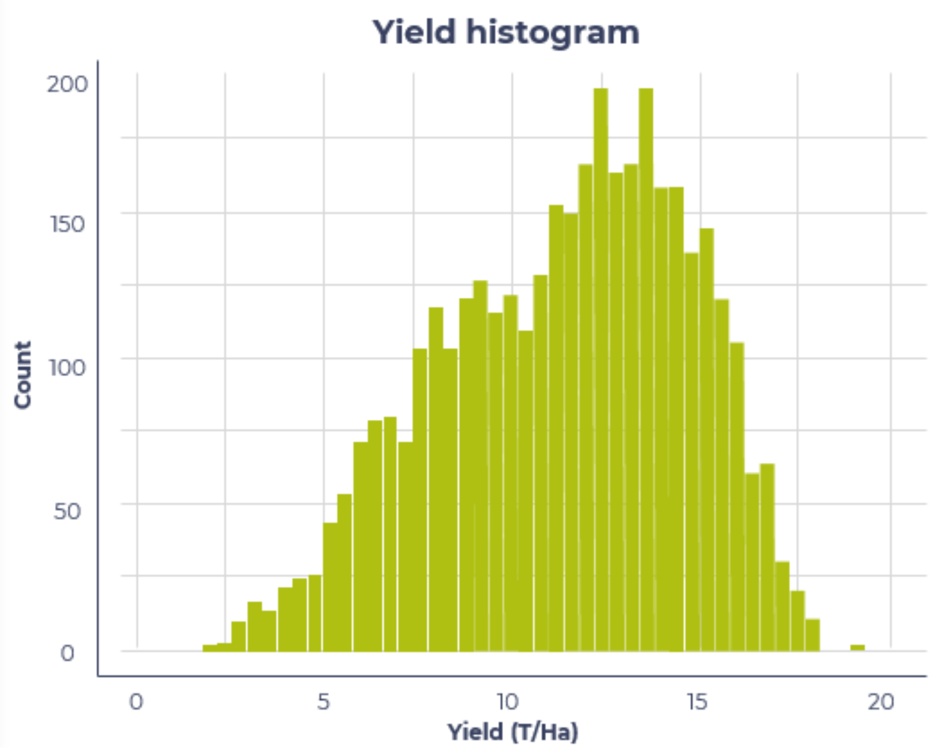

Figure 2. Histogram of corn yield data used for the analysis.

3. Results and discussion

Figure 3 presents the results for the validation set. The model was able to predict the test data with a RMSE of 1.00 Mg/Ha, MAPE of 6.7% and R^2 of 0.90. The model was able to predict adequately both high and low values of yield. The unified dataset covers a large yield parameter space, occurring under different production conditions. It could be of interest to explore the factors effecting the yield to identify trends. Table 1 present a feature permutation analysis conducted on the model. In line with previous studies (Sela et al. 2016), the analysis found total N, SOM, previous crop, and soil texture to be factors that govern corn yield. Interestingly, the 5th and 6th most effecting parameters are climatic – water availability between V10 and tasseling, and early season temperature stress. A subset of the data was created where total N was between 150 and 200 Kg/Ha, SOM% was between 2-3%, and the soil texture was medium (either Silt Loam or Loam). For this subset of data, Figure 4 presents a box plot of yield versus three groups of water availability (irrigation + precipitation) during V10 to VT. In line with previous studies, the results suggest a reduction in yield with lower availability of water during this critical time of corn growth (Cakir 2004). Since all data are standardized and unified, one can now easily explore the conditions where this yield trend occurs. Similarly, all other parameters in Table 1 can be explored for potential new insights.

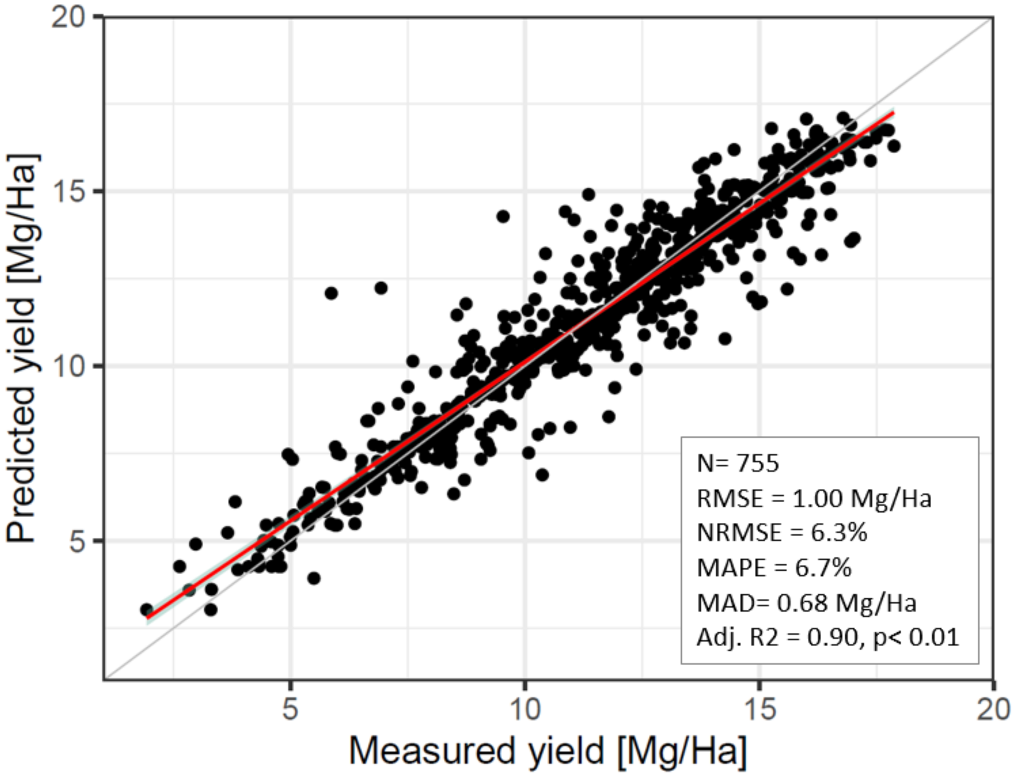

Figure 3. Predicted versus measured corn yield.

Table 1. Feature permutation analysis (mean RMSE value, and 5th-95th percentiles).

Rank	Parameter	RMSE (T/ha) (5%-95%)	Error increase compared to base RMSE (1.00 T/ha)
1	Total N applied	2.70 (2.58-2.81)	170%
2	SOM%	1.87 (1.80-1.94)	87%
3	Previous crop	1.42 (1.37-1.47)	42%
4	Soil texture	1.26 (1.21-1.30)	26%
5	Water availability V10 - VT	1.24 (1.21-1.28)	24%
6	GDD VE-V3	1.19 (1.17-1.22)	19%
7	Water availability V3 – V10	1.17 (1.14-1.20)	17%
8	Water availability PL – VE	1.14 (1.11-1.16)	14%
9	GDD V3-V10	1.13 (1.11-1.16)	13%
10	GDD VT to VT+10	1.13 (1.11-1.16)	13%
11	GDD VT+10 to VT+20	1.10 (1.07-1.12)	10%
12	Water availability PP - PL	1.08 (1.06-1.10)	8%
13	GDD VT+20 to VT+30	1.08 (1.06-1.09)	8%
14	GDD PL - VE	1.07 (1.06-1.09)	7%
15	Water availability VE-V3	1.07 (1.04-1.09)	7%
16	Water availability VT+20 to VT+30	1.06 (1.05-1.07)	6%
17	Water availability VT to VT+10	1.05 (1.04-1.06)	5%
18	Water availability VT+10 to VT+20	1.04 (1.03-1.06)	4%
19	GDD V10-VT	1.03 (1.02-1.04)	3%

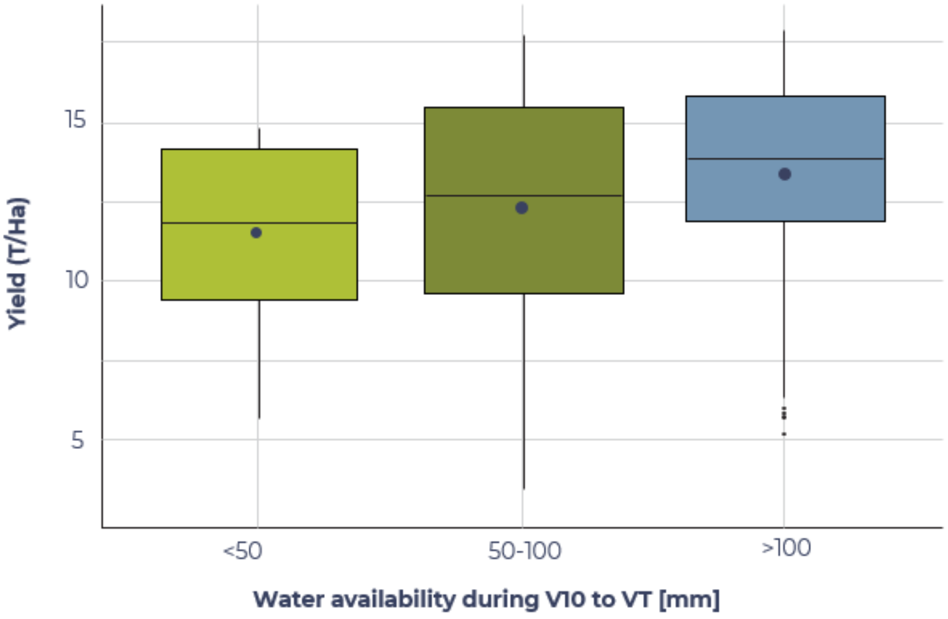

Figure 4. Effect of water availability during V10 to VT on corn yield.

4. Summary and conclusions

Agronomy researchers can utilize standardization platforms to overcome data fragmentation and disparity in file types and parameters naming. Standardization platforms substantially reduce the efforts devoted by researchers to unify, clean and ultimately understand shared data, streamlining collaboration efforts. Data standardization schemes can be applied both for single-source multi-year data, or for data originating from multiple sources. Once assembled, data can be used for modeling and analysis.

Exemplified here using the Agmatix standardization platform, a dataset of corn yield and potential affecting parameters was constructed from various data sources and types. The data was enriched with climatic data for specific growth stages. Using a machine learning model (XGBoost), it was shown that yield data can be adequately predicted across different production environments. A sensitivity analysis subsequently found candidate model parameters that can be further explored to better understand yield trends. While demonstrated here using corn yield data, the same approach can be applied to other agronomic parameter of interest.

We call on the agronomic research community to adopt standardization tools, and share their data through public repositories, community-managed or private-public data platforms. Tools are available for both legacy data and ongoing research data collection. Collaborative efforts and sharing of data can increase the potential of identifying new research directions, and ultimately help tackle current global agronomic challenges.

5. Acknowledgements

We thank Dr. Achim Dobermann (International Fertilizer Association), Prof. Sylvie Brouder (Purdue university), and Prof. Michael Castellano (Iowa State University) for making their data accessible, and for productive discussions on data standardization, data sharing and open science.

6. References

Aubert C., Buttigieg P.L., Laporte M.A., Devare M. & Arnaud E. (2017) CGIAR Agronomy Ontology, http://purl.obolibrary.org/obo/agro.owl

Buttigieg PL, Morrison N, Smith B, Mungall CJ, & Lewis SE (2013) The environment ontology: contextualising biological and biomedical entities. *Journal of Biomedical Semantics*, 4, 43.

Çakir, R. (2004). Effect of water stress at different development stages on vegetative and reproductive growth of corn, *Field Crops Research*, 89, 1-16 https://doi.org/10.1016/j.fcr.2004.01.005.

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). New York, NY, USA: ACM. https://doi.org/10.1145/2939672.2939785

Eagle, A. J., Brouder, S. M., Cambareri, G., Drury, C. F., Parkin, T. B., Smith, D. R. et al. (2020). Nitrous oxide emissions and field-level nitrogen balance of maize and other field crops: Data for meta-analysis. Purdue University Research Repository. doi:10.4231/DFB0-F030

Jones, M. B., Slaughter, P., Budden, A. E., & Habermann, T. (2019). Achieving FAIR: Improvement and Guidance using Quantitative Assessment of Datasets in the DataONE Federation. American Geophysical Union Fall Meeting abstract, 2019AGUFMIN14B..19J

Neilsen, R.L. (2019). Predict Leaf Stage Development in Corn Using Thermal Time, https://www.agry.purdue.edu/ext/corn/news/timeless/vstageprediction.html.

Puntel, L. A., Sawyer, J. E., Barker, D. W., Dietzel, R., Poffenbarger, H., Castellano, M.J. et al. (2016). Modeling Long-Term Corn Yield Response to Nitrogen Rate and Crop Rotation. *Frontiers in Plant Science*. 7, 1630. doi: 10.3389/fpls.2016.01630

Sela, S., van Es, H. M., Moebius-Clune, B. N., Marjerison, R., Melkonian, J., Moebius-Clune, D. et al. (2016). Adapt-N Outperforms Grower-Selected Nitrogen Rates in Northeast and Midwestern United States Strip Trials. *Agronomy Journal*, 108, 1726–1734https://doi.org/10.2134/agronj2015.0606

Stucky, B. J., Guralnick, R., Deck, J., Denny, E. G., Bolmgren, K. & Walls, R. (2018) The Plant Phenology Ontology: A New Informatics Resource for Large-Scale Integration of Plant Phenology Data. *Frontiers in Plant Science*. 9, 517. doi: 10.3389/fpls.2018.00517

Top, J., Janssen, S., Boogaard, H., Knapen.R., & Simsek-Senel, G. (2022) Cultivating FAIR principles for agri-food data, *Computers and Electronics in Agriculture, 196. https://doi.org/10.1016/j.compag.2022.106909*

Wilkinson, M., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data,* 3, 160018. https://doi.org/10.1038/sdata.2016.18