

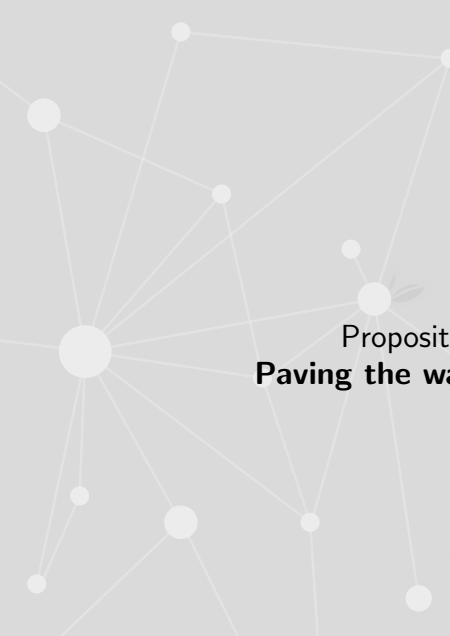
# Paving the way for FAIR data in plant phenotyping

Evangelia A. Papoutsoglou



# Propositions

1. As opposed to metadata, there is no reason to supply RDF distributions of data files.  
(this thesis)
2. Minimum information standards are not only domain-dependent.  
(this thesis)
3. Scientific journals should only publish articles when the accompanying datasets adhere to the latest metadata standards in their respective field.
4. Scientific communication with the public should be left to professional communicators rather than researchers.
5. A proposition database would aid aspiring PhDs.
6. Impostor syndrome is made worse by the knowledge than individuals perceived as more competent suffer from it as well.



Propositions belonging to the thesis entitled  
**Paving the way for FAIR data in plant phenotyping**  
Evangelia A. Papoutsoglou  
Wageningen, 16 June 2021

# **Paving the way for FAIR data in plant phenotyping**

Evangelia A. Papoutsoglou

## **Thesis committee**

### **Promotor**

Prof. Dr R. G. F. Visser  
Professor of Plant Breeding  
Wageningen University & Research

### **Co-promotors**

Dr R. Finkers  
Senior scientist, Plant Breeding  
Wageningen University & Research

Prof. Dr I. N. Athanasiadis  
Professor of Artificial Intelligence & Data Science  
Wageningen University & Research

### **Other members**

Prof. C. J. Lawrence-Dill, Iowa State University, USA  
Prof. Dr D. de Ridder, Wageningen University & Research  
Prof. M. D. Wilkinson, Universidad Politécnica de Madrid, Spain  
Dr W. J. J. Knibbe, Wageningen University & Research

This research was conducted under the auspices of the Graduate School of Experimental Plant Sciences (EPS).



# **Paving the way for FAIR data in plant phenotyping**

Evangelia A. Papoutsoglou

## **Thesis**

submitted in fulfilment of the requirements for the degree of doctor  
at Wageningen University  
by the authority of the Rector Magnificus,  
Prof. Dr A. P. J. Mol,  
in the presence of the  
Thesis Committee appointed by the Academic Board  
to be defended in public  
on Wednesday 16 June 2021  
at 11 a.m. in the Aula.

Evangelia A. Papoutsoglou  
Paving the way for FAIR data in plant phenotyping, 210 pages.

PhD thesis, Wageningen University, Wageningen, the Netherlands (2021)  
With references, with summary in English

ISBN: 978-94-6395-797-7  
DOI: 10.18174/546089

# Table of contents

|  |     |
|--|-----|
| <b>List of abbreviations</b> . . . . .   | 6   |
| <b>Chapter 1:</b> General introduction . . . . .   | 9   |
| <b>Chapter 2:</b> Enabling reusability of plant phenomic datasets with<br>MIAPPE 1.1 . . . . .   | 29  |
| <b>Chapter 3:</b> BrAPI — An application programming interface for<br>plant breeding applications . . . . .  | 51  |
| <b>Chapter 4:</b> Using the MIAPPE standard to improve reusability of<br>plant phenotyping data:<br>Lessons learned from reusing multi-location potato<br>field trial data . . . . . | 71  |
| <b>Chapter 5:</b> Extracting knowledge networks from plant scientific<br>literature:<br>Potato tuber flesh color as an exemplary trait . . . . .                                     | 95  |
| <b>Chapter 6:</b> General discussion . . . . .   | 115 |
| <b>Supplementary materials</b> . . . . .   | 137 |
| <b>References</b> . . . . .  | 169 |
| <b>Summary</b> . . . . .   | 198 |
| <b>Acknowledgements</b> . . . . .  | 201 |
| <b>About the author</b> . . . . .  | 205 |
| <b>Education statement</b> . . . . .   | 207 |

# List of abbreviations

| Abbreviation | Meaning  |
|--------------|--|
| AFLP         | Amplified Fragment Length Polymorphism         |
| API          | Application Programming Interface              |
| BCH          | Beta-Carotene Hydroxylase                      |
| BrAPI        | Breeding API                                   |
| CO           | Crop Ontology                                  |
| CSV          | Comma-Separated Values                         |
| DOI          | Digital Object Identifier                      |
| ENA          | European Nucleotide Archive                    |
| eQTL         | expression QTL                                 |
| EVA          | European Variation Archive                     |
| FAIR         | Findable, Accessible, Interoperable, Reusable  |
| FAO          | Food and Agriculture Organization              |
| FDP          | FAIR Data Point                                |
| FDT          | Farm Data Train                                |
| FN           | False Negatives                                |
| FP           | False Positives                                |
| GO           | Gene Ontology                                  |
| GxE          | Genotype by Environment                        |
| HTTP         | Hypertext Transfer Protocol                    |
| IBM          | International Business Machines Corporation    |
| ISA          | Investigation - Study - Assay                  |
| ISA-Tab      | ISA-Tabular                                    |
| ISO          | International Organization for Standardization |
| JSON         | JavaScript Object Notation                     |
| JSON-LD      | JSON-Linked Data                               |
| KN           | Knowledge Network                              |
| <i>LeZEP</i> | <i>Lycopersicon esculentum</i> ZEP             |
| MAGE-Tab     | MicroArray Gene Expression-Tabular             |

| Abbreviation | Meaning  |
|--------------|--|
| MCPD         | Multi-Crop Passport Descriptor                           |
| MHC          | Major Histocompatibility Complex                         |
| MIAME        | Minimum Information About a MicroArray Experiment        |
| MIAPE        | Minimum Information About a Proteomics Experiment        |
| MIAPPE       | Minimum Information About a Plant Phenotyping Experiment |
| MIC          | Minimal Inhibitory Concentration                         |
| MIC          | MHC class I Chain-related                                |
| NCBI         | National Center for Biotechnology Information            |
| NER          | Named Entity Recognition                                 |
| NLP          | Natural Language Processing                              |
| ORCID        | Open Researcher and Contributor ID                       |
| OWL          | Web Ontology Language                                    |
| PBTT         | Photo-Beta Thermal Time                                  |
| PHT          | Personal Health Train                                    |
| PPEO         | Plant Phenotype Experiment Ontology                      |
| QTL          | Quantitative Trait Locus                                 |
| RDF          | Resource Description Framework                           |
| REST         | Representational State Transfer                          |
| RFLP         | Restriction Fragment Length Polymorphism                 |
| RIL          | Recombinant Inbred Line                                  |
| SNP          | Single-Nucleotide Polymorphism                           |
| SPARQL       | SPARQL Protocol and RDF Query Language                   |
| StAN1        | <i>Solanum tuberosum</i> Anthocyanin 1                   |
| TP           | True Positives   |
| TTL          | Terse RDF Triple Language                                |
| URI          | Uniform Resource Identifier                              |
| URL          | Uniform Resource Locator                                 |
| WEx          | Watson Explorer  |
| WKS          | Watson Knowledge Studio                                  |
| XML          | eXtensible Markup Language                               |
| ZEP          | Zeaxanthin Epoxidase                                     |

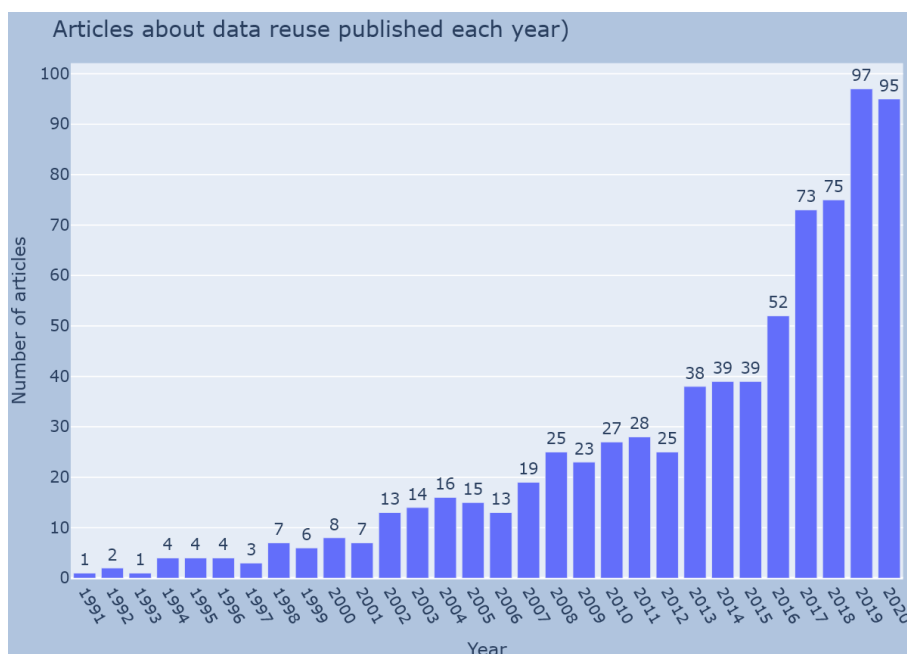


# Chapter 1

## General introduction



The notion that scientific datasets are not treated as they should be is not new. As early as 1991, Neches *et al.*, 1991 discussed the limited scope of “knowledge bases”, data fragmentation and the heterogeneity of knowledge representation formats, all of which limit data reuse. They identified the need for data integration and interoperability, as well as the potential of composite, multi-faceted datasets, which can enable more powerful analyses, without resources being spent on generating new data. Since then, this need has become even more imperative: the numbers of scientific datasets are exploding (doubling every 12 months (Szalay and J. Gray, 2006)), but scientific output simply cannot keep up (estimated to be doubling only every 9 years (Bornmann and Mutz, 2015)). Similarly, interest in data reuse is also growing, as reflected in the number of yearly publications around this topic (**Figure 1.1**).



**Figure 1.1:** A bar graph reflecting the increasing number of scientific articles published each year with data reuse as one of the core topics discussed. It has been generated based on metrics from the Web of Science, using the query  $TS=$ (“data reus\*”), i.e. the “data reus\*” pattern is searched in the “topic” fields, which include the title, abstract, author keywords and keywords generated based on the titles of cited articles.

Researchers are also generally open to the possibility of data reuse. In a recent study inquiring about “Data sharing, management, use, and reuse”, over 85% of the respondents indicated that they were willing to share data or reuse data collected by others, provided it was easily accessible (Tenopir *et al.*, 2011). Accessibility is a key factor undermining not only sharing but also reproducibility according to 90% of scientists in a 2016 survey (Baker, 2016). Indeed, datasets accompanying publications have been found to become inaccessible at a rate of as much as 17% per year (Vines *et al.*, 2014); a different study shows that 80% of data is lost after 20 years (Gibney



and Van Noorden, 2013). In practice, challenges arise because datasets are not readily discoverable and researchers have issues understanding them and judging their suitability for a research goal (due to insufficient documentation), as well as verifying their quality (Curty *et al.*, 2017).

Data reusability can be improved through better documentation and metadata practices. A case study concluded that metadata is usually not readily available because the original data producers could not foresee who might be able and willing to reuse their datasets and therefore did not invest in submitting them to relevant repositories (Wallis *et al.*, 2013). On the other end, scientists have noted that they would have increased confidence in data generated by others if it were accompanied by a) documentation about its collection details and quality assurance methods, b) explicitly mentioned metadata standards, and c) provided provenance information (Tenopir *et al.*, 2020). This is also a manifestation of the chicken-and-egg conundrum, as researchers need to share data for others to be able to reuse it, but reuse requires the data to be already shared (and a supporting infrastructure) in the first place.

## The data landscape in plant phenotyping

The agricultural sciences are at the forefront of research crucial to the continued survival of our species, tackling the challenge of providing for the nutritional needs of an ever increasing population in shifting climates (Cakmak, 2002). Plant breeders have the task of producing cultivars that can perform well and maintain their yield stability in the face of environmental challenges (e.g. extreme climates, pathogens) by crossing and hybridization, using the latest innovations in breeding techniques to speed up the process and increase the efficiency of developing new cultivars.

The presented challenges with data sharing and reuse feature prominently in plant research. A driver for reuse is the unique insight attainable only through analyses exploiting multiple types of -omics data, for which there are multiple examples within the agricultural sciences (Sielemann *et al.*, 2020). Some -omics domains even have established standards (such as MIAME for microarray data (Brazma *et al.*, 2001), MIxS for sequence data (Yilmaz *et al.*, 2011), and MIAPE for proteomics data (C. F. Taylor *et al.*, 2007)) and community repositories (such as GenBank for genomics (Benson *et al.*, 2000), ArrayExpress for functional genomics (Brazma *et al.*, 2003), MetaboLights for metabolomics (Haug *et al.*, 2013), and the NCBI Taxonomy (Federhen, 2012)), which already facilitate discoverability and integration procedures.

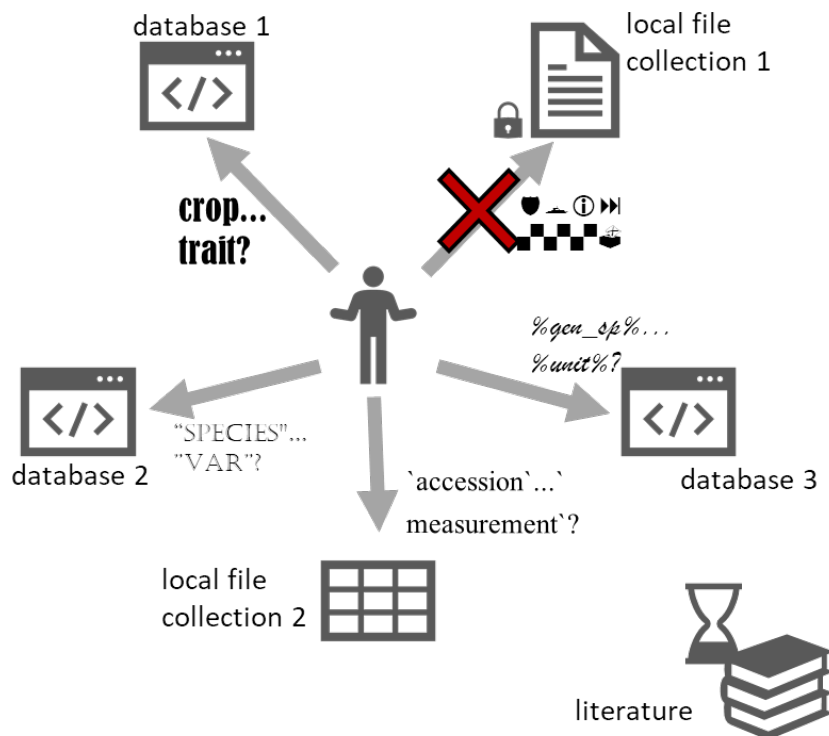
In the domain of plant phenotyping (or phenomics), the path to establishing a straightforward approach to data reuse is less clear. Plant phenotyping data is heterogeneous to the same degree that the experiments that generate it have different goals. Depending on the original goals, the experimental setups, management practices, methods, and studied plant/crop species, the traits measured can vary wildly. The same is true for data collection practices among researchers, who assemble datasets using no particular specifications with respect to file format, syntax, shape or content, and have no central community repository to deposit their datasets. Today, the landscape is composed of many institutional databases that are hard (if not impossible) to meaningfully integrate, and scattered files in local storage media, all contributing to data inaccessibility. The situation for data reuse is also rather grim as these disorganized

data stores are rarely accessible to and interpretable by third parties — or anyone other than the original collectors, due to lack of metadata. Consequently, scientists spend resources on generating new datasets instead of making full use of the existing ones.

For this reason, it is necessary to take steps toward improving the landscape of data reuse within plant phenotyping and the domains connected to it. One aspect of this revolves around the fact that plants have to be studied in the context of their environment, which has motivated multi-location trials aimed at disentangling the effects of genotype and environment on plant development (Millet *et al.*, 2019). Another factor exerting great pressure on the phenotypic data domain is the advent of automated high-throughput, high-resolution technologies which drives the production of more datasets at an even greater rate. To process the higher volumes of data we often rely on prediction models that require training, which can greatly benefit from a higher degree of dataset integration — for example with genomic selection (Spindel and S. R. McCouch, 2016). Improving data management practices for these datasets will pave the way for them to be reused in more large-scale meta-analyses and continue to contribute to our understanding (Coppens *et al.*, 2017). The broadly acknowledged aspects of these data management practices revolve around metadata standardization and the establishment/promotion of community repositories to host this data (D. Brown *et al.*, 2020; Fahlgren *et al.*, 2015; Furbank and Tester, 2011; Pauli *et al.*, 2016; Shakoore *et al.*, 2017; W. Yang *et al.*, 2020).

Plant phenotypes are the products of their genes and the environments in which they developed. This implies three types of data: phenotypic, genomic/genotypic and environmental. Among the three, genomic and genotypic data are the most structured and centralized thanks to repositories such as the European Nucleotide Archive (Leinonen *et al.*, 2010), the European Variation Archive (EBI, 2020) and Ensembl-Plants (Bolser *et al.*, 2016). Environmental data (e.g. weather conditions) is neither centralized, nor does it have widespread standards. However, environmental characterization is only a subset of the data collection process undertaken for some phenotyping experiments. The state of the domain is shown on **Figure 1.2**.

A representative scenario without loss of generality is the following: A meta-analysis investigates the response of developmental and agronomic traits of potato across a multitude of locations, by analyzing a composite dataset (i.e. obtained from the integration of multiple datasets). This yields better results than each of its component datasets alone (Hurtado-Lopez, 2012). Hurtado-Lopez created a collection of data for her doctoral thesis to garner insight into the identification of quantitative trait loci (QTLs) and their trait associations in the CxE potato population (a diploid backcross population) (Jacobs *et al.*, 1995). This population has frequently been studied in the WUR Plant Breeding department (Acharjee, 2013; Anithakumari, 2011; Carreño-Quintero, 2013; B. C. Celis-Gamboa, 2002; Eck, 1995; Getahun, 2017; Hurtado-Lopez, 2012; Jongedijk, 1991; Kloosterman, 2006; Park, 2005; Tessema, 2017; Werij, 2011; Willemsen, 2018). She performed multi-environment analyses studying QTL by environment interactions and observed QTLs that were stable across these environments for multiple plant traits. To do this, she incorporated weather data into her dataset, which enabled analyses of developmental, morphological and agronomic traits across different environments. **Figure 1.3** shows the field trials that were considered in Hurtado-Lopez's work, with phenotypic and environmental data provided by partner institutes from older experiments



**Figure 1.2:** A representation of the state of the landscape within the plant phenotypic data domain. Data is found in institutional databases, online and local file collections. There is no uniform nomenclature, structure or means of searching and obtaining data.

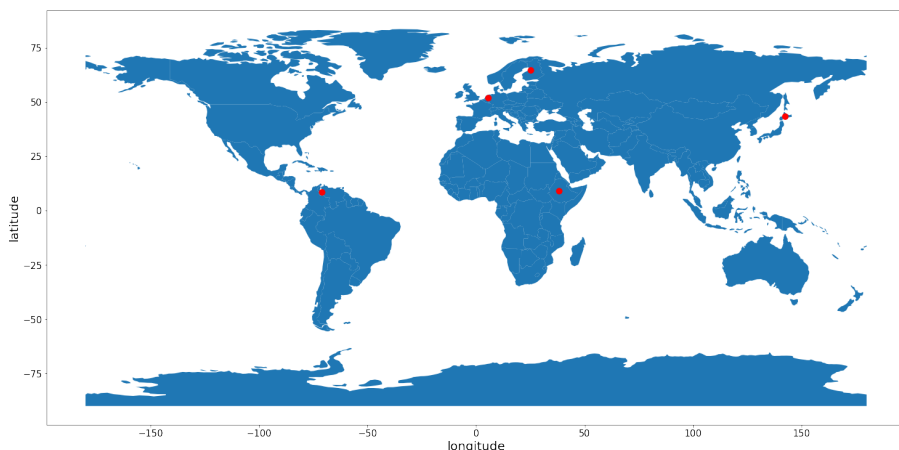
that had been conducted semi-independently, using the recommended guidelines of Wageningen University & Research (WUR). The data for the QTL analysis were acquired separately. Using this data, Hurtado-Lopez was able to establish a correlation between potato traits and the photoperiod they were exposed to, which varied according to latitude. For all of these analyses to be possible, it was necessary to ensure that the datasets from the different field trials were indeed compatible (with respect to e.g. experimental designs, management practices, plant trait observations). Part of this process revolved around maintaining traceability for the different CxE genotypes across the experiments, so they could be compared at later stages.

Over the years, researchers in WUR Plant Breeding have produced high volumes and different types of data (e.g. phenotypic, genotypic, molecular) for the CxE potato population (**Figure 1.4**). Experiments have taken place in different locations, on the field and in vitro, and new genetic maps have been calculated as marker technologies improved. The original (core) population has also been extended with seeds from the original cross, and the number of genotypes has recently been expanded to over 1600

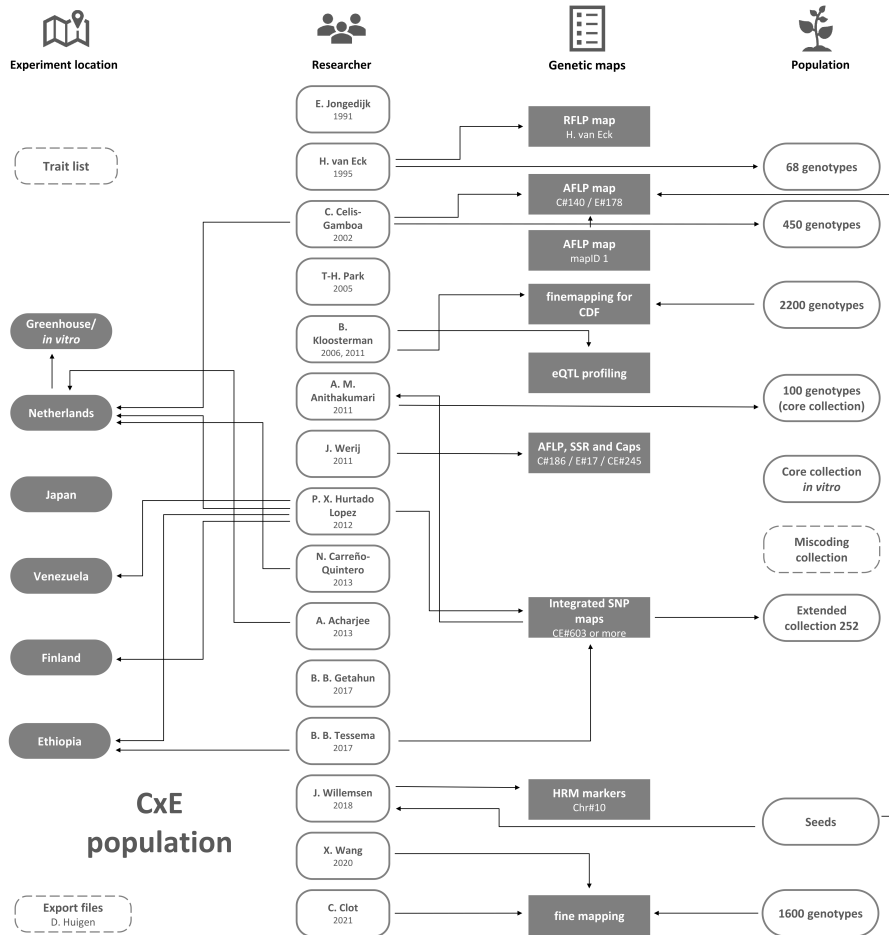
(for fine mapping) , but otherwise the propagation of the material over time has been clonal - i.e. the collection needs to run through a yearly maintenance cycle (planting, propagation, harvest, storage). Given the sheer amount of human labour involved, mix-ups related to the plant materials at various stages of that cycle were to be expected (i.e. tubers were accidentally assigned incorrect labels; tubes with leaf material for genotyping were switched in the lab). Resolving mix-ups was done by re-genotyping clones and comparing the pattern to the original data. Such analysis provides additional insight on the trustability of a given clone, which is relevant as an additional layer of metadata and should be included in meta-analyses of the population. As such errors cannot be prevented, it is crucial to use this metadata to maintain the traceability of the materials, preserve the results and the methods of analyses that uncovered the errors in the first place and ensure reproducibility and continuity.

## The FAIR data principles

The FAIR data principles state that data should be Findable, Accessible, Interoperable and Reusable, and are a set of domain-agnostic guidelines that data should abide by in order to be more suitable for reuse (Wilkinson *et al.*, 2016). In essence, FAIR data should be self-documenting and understandable to both humans and machines, enabling software to process it with minimal human intervention. The general sentiment behind the principles is not new, with a previous rendition of the same spirit having appeared as recently as 2014 in the shape of “Ten simple rules for the care and feeding of scientific data” (Goodman *et al.*, 2014). But this more concrete, equally human- and machine-centric manifestation has earned international praise and, unlike previous attempts, the FAIR principles have been gaining support and widespread popularity. Their main points are described as follows.



**Figure 1.3:** An overview of the locations where experiments have been conducted using the CxE population. The different latitudes mean that the photoperiods differ, affecting the plant traits (among other things).



**Figure 1.4:** A partial overview of experiments, people involved and marker maps using the CxE population, as found in WUR Plant Breeding (adapted from P. Hendrickx's illustration). Data from each field experiment has been analyzed by different people, with Hurtado-Lopez bringing multiple datasets together for a composite analysis. Various types of marker maps with different marker sets have been produced by different researchers over the years. The CxE population has 3 main sets. The traits evaluated for each field experiment varied as well. The theses mentioned in the illustration are: Acharjee, 2013; Anithakumari, 2011; Carreño-Quintero, 2013; B. C. Celis-Gamboa, 2002; Eck, 1995; Getahun, 2017; Hurtado-Lopez, 2012; Jongedijk, 1991; Kloosterman, 2006; Park, 2005; Tessema, 2017; Werij, 2011; Willemsen, 2018. Other relevant work includes Eck et al., 1995; Hurtado-Lopez et al., 2015; Kloosterman et al., 2013.

**Findability** is achieved with the attribution of unique, persistent identifiers to data. With the same being done to metadata, connections can be drawn between data points and their descriptors. These identifiers also enable unambiguous indexing in searchable data repositories, where users can be directed to best locate datasets relevant to their interests. Such repositories can then build federations.

**Accessibility** comes with the use of free, open protocols that allow (meta)data retrieval and authentication, and the persistence of metadata even when the data is no longer present. It is important to note that accessibility is not the same as openness, and that FAIR specifically makes the case that some data (e.g. related to human health) can and should remain private, or be selectively available through proper channels described in the metadata. Ultimately this decision should be up to the data owner.

For **interoperability**, the (meta)data should be presented in formats dictated in each discipline, so that humans and workflows can interact with it. Interoperability additionally has a contextual dimension, as specific connections (with unique identifiers) to other (meta)data not only enhance the descriptions, but can also be crucial for outlining dependencies between datasets that are crucial for interpretation.

Finally, for **reusability**, the data should be accompanied by clear conditions for its reuse (license) and adhere to domain-relevant community standards. Such standards include descriptions of data provenance and the procedures behind data generation, to reduce the likelihood that data will fail to meet the documentation requirements in a given field and delineate descriptions of datasets in a way that is suitable for both machines and humans (since one knows exactly where to look to find the value of an attribute). Reusability without findability and accessibility is not possible since one needs to be aware of the existence of a dataset, its location and its content in order to reuse it. Interoperability can also be necessary from the perspective of contextualization, but moreover it can also provide tremendous advantages as far as data integration is concerned.

It is important to mark the differences between the FAIR data principles and standards. As opposed to the latter, the FAIR principles have no single proposed implementation; in fact, it lacks any kind of technical specification. It is up to each community and data provider to determine what FAIR should translate to for their particular case and implement it as they see fit. However, intra- and inter-community alignment is not optional, as there should be a consensus on the metadata and the way that it is to be communicated. Another point that commonly causes confusion is the openness of the data, which FAIR actually makes no statement about: data can be FAIR and not publicly available itself, though the metadata should be (as specified under Accessibility).

## Concepts behind FAIR

Some particular terms related to FAIR are central to this work and are therefore introduced below in greater detail.

- **Metadata:** A set of data attributes that give information about another data point. In a relational database, a row in a table is often about a single data point with the key establishing its identity, and other columns either describing

its attributes (metadata), or connecting it to other rows, in the same table or otherwise.

Depending on one's domain and perspective, the same piece of information could be either data or metadata.

- **Identifiers:** In the context of FAIR, (meta)data identifiers have to be persistent and globally unique. Both conditions ensure not only that references to a (meta)data point are and remain valid but also that there is no ambiguity, as a single identifier always represents the same thing. More familiar examples of persistent, unique identifiers are DOIs and ORCIDs. Often, public repositories may also assign identifiers to their resources — not to be mixed up with regular URIs, which have no guarantee of persistence. An example of a scientific repository that attributes persistent identifiers is UniProt, where the Beta-carotene 3-hydroxylase 1 protein in *Arabidopsis thaliana* is represented by the identifier <https://www.uniprot.org/uniprot/Q9SZZ8>.
- **Ontologies and controlled vocabularies:** A controlled vocabulary is a list of specifically selected terms, each carrying a specific meaning in its context, that is used to resolve ambiguity, power search systems and provide machine understanding. Ontologies can go several steps beyond controlled vocabularies to support reasoning while requiring more formalization but, simply put, they can also model relationships, restrictions and axioms between terms. Both ontologies and controlled vocabularies (should) represent a consensus in a community, and perhaps an authority governing them.
- **Machine readability:** The FAIR principles promote machine actionability, which is a significant step toward automatic data discovery, interoperability and reuse given the large numbers of datasets and data volumes that are being generated. Machine readability is only part of that, and does not imply the existence of any infrastructure surrounding FAIR data as, for example, an indexing/search mechanism would for discoverability. FAIR cannot exist without the provision of machine-readable metadata. In order for a machine to be able to meaningfully read and “make sense” of metadata attributes, they need to adhere to expectations set by the community. By “make sense” we mean that a machine should be able to understand the significance of the metadata and make informed choices about the best way to use the data. Ontologies are a means by which to support machine readability, as the meanings defined inside are very specific, though use of an ontology does not automatically render a dataset/metadata machine readable. A simple example would be the definition of the steps and access protocols that can be used to retrieve a dataset, as described by its metadata.
- **Linked data:** The concept of linked data is closely connected to FAIR, though neither requires the other. Linked data refers to publishing structured data online and connecting it to other data. Just like FAIR, linked data relies on the use of unique persistent identifiers (URIs, more specifically) which can be dereferenced (resolved) and thereby provide information about the entities they have been assigned to. The (meta)data having a structure means that, unlike natural language content which is only intended for humans, it is suitable for

| Subject            | Standards | Databases |
|--------------------|-----------|-----------|
| biology            | 691       | 1009      |
| agriculture        | 70        | 77        |
| omics              | 104       | 318       |
| genetics           | 66        | 183       |
| molecular genetics | 17        | 52        |
| genomics           | 64        | 249       |

**Table 1.1:** The table holds numbers retrieved from the FAIRsharing repository (Sansone et al., 2019) indicating the number of standards and datasets in domains related to plant breeding (retrieved on February 26, 2021).

machine consumption. The Resource Description Framework (RDF) and its implementations are commonly used to express linked data in a structure that resembles a graph, where different entities (referred to by their URIs) represent the nodes, and the relationships between them, i.e. the “links” of linked data, mark the edges (W3C, 2020b).

## FAIR in plant phenotyping

Better adherence to the FAIR data principles could improve the landscape of data management in plant phenotyping and empower researchers to do more with fewer resources, while preserving the integrity of datasets and supporting experimental reproducibility. As mentioned, the root of most of the current issues lies with the vast heterogeneity, decentralized data storage and ambiguous documentation practices within the field, all of which are directly addressed by the FAIR principles. **Table 1.1** presents the numbers of standards and databases in domains related to plant breeding, as retrieved from the FAIRsharing repository (Sansone et al., 2019). The numbers for databases range from a few dozen to a few hundred for some domains, and the dozens of standards reported for each of these domains indicates that integrative approaches (for querying, data retrieval and processing) would be a demanding task in such a complicated landscape. An non-exhaustive list of prominent databases relevant to plant breeding in general, and to potato as a more specific example, can be seen in **Table 1.2**. A general overview of the principles proposed to alleviate this situation follows, with specific contexts for the plant phenotyping domain and the applicability of the principles in it.

- **Findability:** The results of phenotyping experiments are housed in institutional databases, locally and as online files accompanying publications. All of these would benefit from a uniform method of broadcasting the information they hold, even if the data points themselves are not readily available. Attributes such as identification of plant material (e.g. species, accessions, genotypes) involved and the plant traits observed would be great starting points for introducing searchability. As it stands, one has to navigate different access portals and be confronted with



|                 | Database   | Reference   | Content   |
|-----------------|--|---|---|
| General         | NCBI Taxonomy                                      | Federhen, 2012  | organism names and taxonomic lineages                                     |
|                 | GenBank  | Benson <i>et al.</i> , 2000   | publicly available DNA sequences  |
|                 | Ensembl Plants                                     | Bolser <i>et al.</i> , 2016   | genome sequence, gene models, functional annotation, and polymorphic loci |
|                 | UniProt  | Bairoch <i>et al.</i> , 2005  | protein sequence and functional information                               |
|                 | European Nucleotide Archive (ENA)                  | Leinonen <i>et al.</i> , 2010   | nucleotide sequences  |
|                 | ArrayExpress                                       | Brazma <i>et al.</i> , 2003   | functional genomics data  |
|                 | European Variation Archive (EVA)                   | <a href="https://www.ebi.ac.uk/eva/">https://www.ebi.ac.uk/eva/</a>                                 | genetic variation data  |
|                 | Sol Genomics Network                               | L. A. Mueller <i>et al.</i> , 2005  | phenotypic, genotypic, genomic data for <i>Solanaceae</i> species         |
|                 | The <i>Arabidopsis</i> Information Resource (TAIR) | Huala <i>et al.</i> , 2001  | genetic and molecular biology data for <i>Arabidopsis thaliana</i>        |
|                 | T-DNA Express                                      | <a href="http://signal.salk.edu/cgi-bin/tdnaexpress">http://signal.salk.edu/cgi-bin/tdnaexpress</a> | <i>Arabidopsis</i> t-DNA insertion mapping tool                           |
| Potato-specific | Expression Atlas                                   | Papatheodorou <i>et al.</i> , 2018  | gene and protein expression data  |
|                 | KEGG PATHWAY Database                              | Kanehisa <i>et al.</i> , 2004   | diagrams of molecular interactions, reactions and relations               |
|                 | Spud DB  | Hirsch <i>et al.</i> , 2014   | potato genome browser and tools   |
|                 | European Cultivated Potato Database (ECPD)         | <a href="http://www.europotato.org/">http://www.europotato.org/</a>                                 | potato variety descriptions   |
|                 | Potato Variety Database                            | <a href="http://varieties.ahdb.org.uk/">http://varieties.ahdb.org.uk/</a>                           | potato variety descriptions   |
|                 | Eastern Potato Variety Development Database        | Clough <i>et al.</i> , 2010   | potato variety descriptions   |
|                 | Potato Pedigree Database                           | Van Berloo <i>et al.</i> , 2007   | potato pedigree data  |
|                 | PoMaMo   | Meyer <i>et al.</i> , 2005  | potato genome data  |

**Table 1.2:** The top of table lists online databases and resources that are commonly used in the plant breeding domain. The bottom of the table list resources that, in addition to the previous ones, are especially useful in potato breeding. Both lists are non-exhaustive and only presented as an example of the heterogeneity in the plant domain.

disharmonious nomenclature to hopefully find a piece of information that may interest them, which can be very resource and time intensive.

- **Accessibility:** Different plant-related databases and other file storage systems each require a unique approach to navigate. Meanwhile scientific publications often point to datasets that have been rendered inaccessible due to lack of maintenance, privacy concerns, or poor interactions with the original providers. Even when desired datasets are obtainable, many of them lack metadata definitions, which diminishes their usefulness.
- **Interoperability:** The integration of data from different sources is almost always obstructed by syntax and file format variation, but unresolvable issues can also arise due to complexities in data modeling, i.e. lack of clarity in meaning for machines as well as humans. Data and workflows are, as a result, impossible to reconcile and connect. Some attributes of phenotypic datasets additionally have close ties to other domains. For example, genotype identifiers can be used for drawing connections to genomic/genotypic data and the locations of field experiments can be used for the discovery of relevant environmental datasets (weather, soil). Intra- and inter-domain interoperability require close contact between communities, making it all the more challenging.
- **Reusability:** Plant phenotyping data is difficult to reconcile into a single model. This is aggravated by general dissent in the community about acceptable ways to manage and communicate data, depending on the specific goals and context of a project. Development of the necessary “domain-relevant community standards” would be a significant first step to address this.

## Plant resources to support FAIR

Each domain needs to make different accommodations to promote the FAIR principles. Reusability in particular relies on “domain-relevant community standards”, and interoperability requires “vocabularies that follow FAIR principles”.

In plant phenotyping, a number of steps have already been taken to help the community advance toward the adoption of these principles:

- **MIAPPE (1.0)** aims to guide researchers in their efforts to document plant phenotyping experiments (Krajewski *et al.*, 2015). It has been composed as a checklist, with each of its sections listing an aspect of an experiment and related attributes, e.g. general metadata, environment, experimental design, observed variables and biosource. Additionally, this checklist includes recommended ontologies that can be used to enrich the description of these attributes. To give it a concrete format for better computer readability, an implementation in the ISA-Tab format has also been proposed (Ćwiek-Kupczyńska *et al.*, 2016).
- **Ontologies:** The community has been producing controlled vocabularies and more intricate ontologies to model plant knowledge. Some of the most prominent ones are:

- The **Crop Ontology**, which spans several species-specific ontologies, with each listing plant variables that are commonly observed in the plant in question (R. Shrestha *et al.*, 2012). Each variable is modeled as a triple consisting of a trait, a method and a scale.
- The **Plant Trait Ontology**, which takes the opposite approach to modeling plant variables. Instead of focusing on plants first, it places plant traits at the top and categorizes them, independently of their species (Arnaud *et al.*, 2012). It proposes an alternative model where a plant phenotype is composed of the plant trait, the plant anatomical entity that was the target of the observation, and a value.
- The **Plant Ontology**, which models the morphology and anatomical structure of plants (Jaiswal *et al.*, 2005). It presents a complex hierarchy, interrelating terms on multiple levels and including genetic associations.

The plant community has clearly made the first steps toward standardizing aspects of its data complexity, both terminologically and methodologically. MIAPPE 1.0 has been a great step forward, and successfully covers the aspects of an experiment that should be documented. However, this documentation is still in need of disambiguation, in particular with respect to its underlying data model. Without this in place, different interpretations can be given to every MIAPPE component and the connections drawn between them. Furthermore, even when the ontologies that do exist already cover their particular domains well, it is necessary to establish a means of connecting them to draw a more complete picture — otherwise the isolated components are of limited use. Other than MIAPPE, the individual ontologies/controlled vocabularies also need to be the target of further community-based development, so that they may cover a broader scope of plant traits, morphology, anatomy and experimental design elements.

## The potential of FAIR plant data

If the plant community advanced in implementing the FAIR data principles there would be benefits to research longevity and reproducibility as well as knowledge gains from resources that would have otherwise either remained obscured or required heavy investment to access. (Meta)data should be accessible and understandable to both humans and machines, with the latter being crucial for the effective handling of high volumes of data.

Currently, research published in journals targets exclusively human readers. Often, an effort is made on the author's part to present the process that led to their experimental results. However, due to individual documentation practices, another effort has to be made separately on the reader's part to interpret it. Although traditional search engines and journal databases can direct users to publications that may fit one or two of their search criteria, they cannot account for the complexity inherent in the domain and represented by differing experimental goals. Institutional databases, as structured data sources, are individually easier to look into. However, collectively, the task also requires significant effort, and should results be found in different databases, harmonizing them for common reuse is as laborious as parsing publications, evaluating the materials and

methods sections for suitability, and tracking down the actual data independently. For humans, organized metadata would at least facilitate understandability and reduce the number of datasets that would otherwise be deemed un-reusable due to a lack of information or datasets incorrectly deemed suitable for reuse.

Machine readability is no substitute for human readability, but it can offer a dramatic acceleration of every stage of data handling. Machine-readable metadata can be indexed and referenced by central community registries. With a powerful model to support metadata, information systems could respond to complex queries and act as a one-stop-shop for users. Moreover, data that is easy to integrate would be much more attractive, allowing researchers to spend more time investigating their hypotheses and less on inspecting provider-specific data syntaxes and formats. Tracking down contacts based on publications to simply request the data (or maybe not even acquire it) only to discover that it is not usable, due to its shape or experimental parameters, would be an issue of the past.

## Making resources FAIR

The ideal, fully FAIR scenario for any kind of data is clearly far removed from the current reality. It is important to note that being FAIR is not a binary state, but rather a gradient with each step a significant improvement on the last. Across communities, there are general steps that can be taken to make resources FAIR (Jacobsen *et al.*, 2020). Improving data management and metadata practices is an investment that will only obviously start paying off after a critical mass of adopters has been reached. Until then, the process holds more subtle, yet nevertheless important, benefits such as better data visibility, citability and credit attribution, which could incentivize reuse (Pierce *et al.*, 2019).

The investment required is generally higher when it comes to historical data. This is in part due to the generally poorer documentation practices of the past, especially because emerging standards now require the specification of attributes that were previously not considered important in the domain and were therefore never recorded. The process of making an existing dataset FAIR can be a daunting task even when it is relatively recent, as the volume of metadata annotations that have to be made can be considerable if documentation has not been an ongoing effort. Observing good (meta)data management practices for datasets from the conception of an experiment is simpler, and can therefore lead to better results.

## A FAIR example: the Personal Health Train

The FAIR data principles have already guided an application in the medical domain, the Personal Health Train (PHT). It successfully demonstrates how the FAIR principles can promote data integration and reuse (Beyan *et al.*, 2020). While the question of data accessibility exists in every domain of science and business, in this case confidentiality is of particular concern given the sensitive nature of patient data and legislation surrounding it. Because of this, medical institutes have been reluctant to share their datasets outside their infrastructure and internally resort to duplication of the multimodal datasets they receive. Duplication is an easier solution, since no concrete integrative modeling is

required for highly heterogeneous medical data, but at the same time it limits the linkability of records and hinders further analysis.

To overcome this issue, the PHT application has been proposed. Instead of opposing the distributed, heterogeneous nature of the domain, it embraces it, ensuring that data providers can assert full control over what is shared and with whom. This is achieved by sharing not the original datasets, but the results of specified, transparently auditable operations. This is achieved through data trains, which carry data operation requests. They are dispatched, matched to relevant data stations, and directed there, where their queries are audited by mechanisms defined by the data owners. The operations, if approved, are carried out, and the results returned with no personal information (Deist *et al.*, 2020; Shi *et al.*, 2019).

As far as architecture is concerned, the PHT comprises three main components: the station, the train and the track, each displaying multiple facets of FAIR. In short, the data provider hosts a station holding data and metadata descriptions, and ensures it is discoverable while also providing computational resources that can be exploited by trains. The train holds the (potentially distributed) question that an interested party may ask and has its own unique identifier and metadata. Finally, the track is the component that connects the other two, and restricts who is authorized to do what, directing trains (questions) to suitable stations based on their metadata and resources, respectively, and aggregating the results of multiple routes before transmitting them back to the train dispatcher. This modular architecture ensures that the station owners can control how and by whom data can be used while the track ensures that a train is directed to all stations that can accept it so that the user (who dispatched the train) can get a full response.

Currently, the PHT is the most prominent FAIR-based application. It is already in use and its privacy-shielding, integration-minded infrastructure is proving useful to researchers within the medical domain.

## The Farm Data Train

Instead of duplicating efforts toward a FAIR data infrastructure, the plant community can learn from the progress in the medical domain when it comes to data management, specifically from the PHT. A parallel ambition is therefore the Farm Data Train (FDT)<sup>(1)</sup>, which could connect plant data providers to facilitate better resource exploitation (Finkers, 2018). The sets of interested stakeholders include: farmers, who want to be informed of the best choices and options for their own enterprises; equipment manufacturers, who provide their services for agricultural operations; and researchers who are involved in investigations related to biology, food processes, economics and more. The scope of data that the FDT would transport extends far beyond the domain of plant phenotyping, and includes all types of -omics data) over the chain from breeder, producer, processor, retail, to consumer. However, as phenotypic data is so central of a data source and challenging to work with, it is the first major hindrance that needs to be overcome for meaningful data standardization and human as well as machine intelligibility.

The needs for data discovery and acquisition are not too different between the

<sup>(1)</sup><https://www.youtube.com/watch?v=1MZs5cb3pC8>

two contexts. For the FDT and the PHT alike, stations need to maintain records of their metadata that can be paired with requests from incoming trains with the help of tracks. Farmers and researchers, much like medical data providers, should be able to inspect the data queries directed to their data stores and maintain the right to refuse requests. The types of data integration may differ between the two cases. For example, plant genotypes may be observed in different environments where they need to be tracked, so the dimension of confidentiality for plant identities may be discussed in a different light since it is a central component of integration. Plant data is also more commonly associated with environmental information, coming in from weather stations, soil databases, satellites and drones. Investigations around crop yield stability, for example, call for the integration of data from multi-environment trials by necessity. This stands in contrast to the medical domain, where data trains may gain useful insights from analyzing the response returned from perhaps even a single station, provided that a usable dataset has been discovered. Another difference that makes the use of the FDT infrastructure imperative is the high number of stakeholders (e.g. farmers) that would want to make use of such data, and therefore need a robust mechanism to achieve high discoverability and transparency.

The needs for data processing may vary; both human and plant genomics both use extensive datasets that require significant computer resources. However, plant genomic datasets are often not as strictly guarded as their human counterparts, and therefore the researcher posing the question may wish, or be forced to, undertake the computing load in lieu of the data station. In that respect, for the FDT, it often makes less sense to bring the processing to the station and return only the results. Overall, however, this infrastructure can clearly aid integration and reuse in both domains.

## **A step toward FAIR scientific literature**

Heterogeneous databases and local files generally cannot skip a data transformation stage before they can be reused because they lack standardization. Still, utilizing such resources is much more efficient than diving into literature to assemble a corpus of information and determine future steps. The essential difference here is that, while all are textual, databases and more generally data files are structured information, whereas literature is not, as it is intended to be consumed exclusively by humans.

The intention behind the FAIR principles is to provide an answer to the high-speed, high-volume research environments in this era of big, fragmented data. Scientific publications should have a place in this answer as they report cutting-edge discoveries and collectively compose a landscape that is more powerful than any of its constituents. To combine FAIR and publications, the first step would be to make the information obscured in unstructured text somehow available in a computer-readable format. The steps that follow this structurization would be similar to the ones recommended for any other type of dataset found within the domain.

## **Building towards FAIR plant data: Scope of this thesis**

The FAIR data principles require, for each domain that follows them, conventions upon which the respective communities agree. In plant phenotyping, the foundations have

been laid with the establishment of the MIAPPE metadata standard, which serves as a checklist for researchers, and ontologies. In this thesis, we build toward making plant phenotyping data FAIR and explore the application and potential of the principles within the domain.

This effort revolves around improving the means for data exchange between plant databases, improving metadata standards and connecting existing resources. This marks a step toward all aspects of the FAIR principles, thereby bringing us closer to a scenario where phenotypic datasets can be readily discovered, acquired and integrated, even potentially with data outside of the plant domain. Data integration can enable more powerful meta-analyses and through reuse increase the efficiency of the scientific process by reducing the need for new experiments and increasing reproducibility.

**Chapter 2** builds upon the existing MIAPPE standard, elevating it above a flat checklist and improving aspects with two core goals in mind: increasing the scope of phenotyping studies that can be supported and boosting its compliance to the FAIR principles. To ensure machine readability, we provide an ontology that specifies the semantic model of MIAPPE and can be used to implement it, in addition to BrAPI and the popular ISA-Tab data format.

**Chapter 3** presents the plant Breeding API (BrAPI). Phenotyping data plays a central role in breeding applications but the fragmented global landscape of databases presents challenges when it comes to data integration. BrAPI aims to bridge the content-, structure- and syntax-related chasm between these databases by providing a common data exchange format.

**Chapter 4** follows the process of making FAIR a phenotyping dataset that has been used in previous studies and demonstrates its use for a simple data integration scenario. We identify core challenges and points where MIAPPE compliance could indeed help with both the preservation and the findability of essential metadata. The dataset in question is part of what was used in Hurtado-Lopez's doctoral thesis (Hurtado-Lopez, 2012), and comprises the relevant data from five independently conducted phenotypic experiments revolving around the CxE potato population. The data integration in this set requires a second component, environmental data corresponding to those experiments, to draw connections between the genotypes and their phenotypic responses to environmental conditions. Finally, we demonstrate how standardized (meta)data can be an asset to experimental replicability and transparency.

**Chapter 5** explores scientific literature as a treasure trove of unstructured (and therefore machine-inaccessible) information. Though literature may be accessible to researchers, keeping up with all new publications and contextualizing recent findings is a time-consuming task. To alleviate some of the mental load required on the researchers' part, we explore knowledge networks from scientific articles constructed with NLP methods, and retrospectively demonstrate that such networks can be used to reduce the time between the publication and constructive utilization of data. This example is based on literature around the flesh color of potato as a trait and its genetic associations.

**Chapter 6** takes a look back and discusses the recent steps that have been taken toward making plant phenotyping data FAIR and the implications for the future of the domain and plant scientists.

## Acknowledgements

**Figure 1.4** is based on an original diagram created by Patrick Hendrickx. The version shown here includes adjustments and additions, but it would not have existed without his efforts. His contribution is gratefully acknowledged and appreciated.



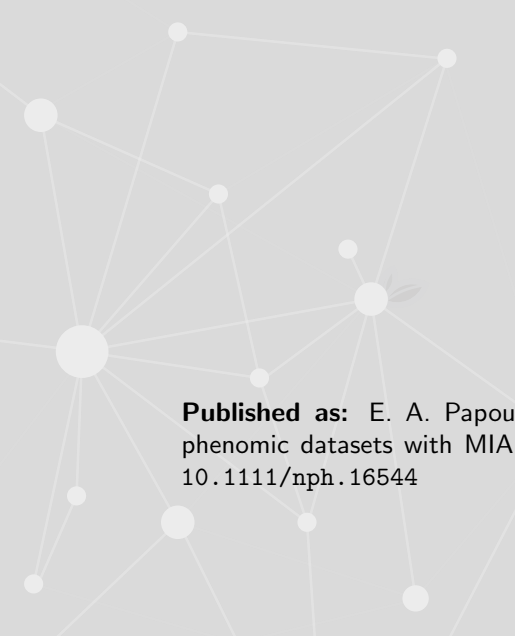




## Chapter 2

# Enabling reusability of plant phenomic datasets with MIAPPE 1.1

Evangelia A. Papoutsoglou<sup>1</sup>, Daniel Faria<sup>2</sup>, Daniel Arend<sup>3</sup>, Elizabeth Arnaud<sup>4</sup>, Ioannis N. Athanasiadis<sup>5</sup>, Inês Chaves<sup>6,7</sup>, Frederik Coppens<sup>8,9</sup>, Guillaume Cornut<sup>10</sup>, Bruno V. Costa<sup>6,11</sup>, Hanna Ćwiek-Kupczyńska<sup>12</sup>, Bert Driesbeke<sup>8,9</sup>, Richard Finkers<sup>1</sup>, Astrid Junker<sup>3</sup>, Graham J. King<sup>13</sup>, Paweł Krajewski<sup>12</sup>, Matthias Lange<sup>3</sup>, Marie-Angélique Laporte<sup>4</sup>, Celia Michotey<sup>10</sup>, Markus Oppermann<sup>3</sup>, Richard Ostler<sup>14</sup>, Hendrick Poorter<sup>15,16</sup>, Ricardo Ramírez-Gonzalez<sup>17</sup>, Jochen C. Reif<sup>3</sup>, Philippe Rocca-Serra<sup>18</sup>, Susanna-Assunta Sansone<sup>18</sup>, Uwe Scholz<sup>3</sup>, François Tardieu<sup>19</sup>, Cristobal Uauy<sup>17</sup>, Björn Usadel<sup>15,20</sup>, Richard G.F. Visser<sup>1</sup>, Stephan Weise<sup>3</sup>, Paul J. Kersey<sup>21</sup>, Célia Miguel<sup>6,11</sup>, Anne-Françoise Adam-Blondon<sup>10</sup> and Cyril Pommier<sup>10</sup>



**Published as:** E. A. Papoutsoglou *et al.* (2020a). Enabling reusability of plant phenomic datasets with MIAPPE 1.1. *New Phytologist* 227.1, pp. 260–273. DOI: 10.1111/nph.16544

<sup>1</sup>Plant Breeding, Wageningen University & Research, PO Box 386, 6700AJ, Wageningen, The Netherlands

<sup>2</sup>BioData.pt / Instituto Gulbenkian de Ciência, Oeiras, Portugal

<sup>3</sup>Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) Gatersleben, 06466 Seeland, Germany

<sup>4</sup>Information Technology Group, Wageningen University, Wageningen, The Netherlands

<sup>5</sup>Bioversity International, Parc Scientifique Agropolis II, 34397 Montpellier Cedex 5, France

<sup>6</sup>Instituto de Tecnologia Química e Biológica António Xavier, Universidade Nova de Lisboa (ITQB NOVA) Avenida da República, Oeiras, 2780-157, Portugal

<sup>7</sup>Instituto de Biologia Experimental e Tecnológica (iBET), 2780-157, Oeiras, Portugal

<sup>8</sup>Department of Plant Biotechnology and Bioinformatics, Ghent University, Technologiepark 71, 9052 Ghent, Belgium

<sup>9</sup>VIB Center for Plant Systems Biology, Technologiepark 71, 9052 Ghent, Belgium

<sup>10</sup>URGI, INRA, Université Paris-Saclay, 78026 Versailles, France

<sup>11</sup>BioISI – Biosystems & Integrative Sciences Institute, Faculdade de Ciências, Universidade de Lisboa, Lisboa, Portugal

<sup>12</sup>Institute of Plant Genetics, Polish Academy of Sciences, ul. Strzeszyńska 34, 60-479 Poznań, Poland

<sup>13</sup>Southern Cross Plant Science, Southern Cross University, Lismore, NSW 2477, Australia

<sup>14</sup>Computational and Analytical Sciences, Rothamsted Research, Harpenden, Hertfordshire AL5 2JQ, UK

<sup>15</sup>Plant Sciences (IBG-2), Forschungszentrum Jülich GmbH, D-52425 Jülich, Germany

<sup>16</sup>Department of Biological Sciences, Macquarie University, North Ryde, NSW 2109, Australia

<sup>17</sup>Department of Crop Genetics, John Innes Centre, Norwich Research Park, Colney, Norwich, NR4 7UH, UK

<sup>18</sup>Oxford e-Research Centre, Department of Engineering Science, University of Oxford, 7 Keble Road, Oxford, OX1 3QG, UK

<sup>19</sup>INRA, Laboratoire d'Ecophysiologie des Plantes sous Stress Environnementaux, UMR759, 34060 Montpellier, France

<sup>20</sup>Institute for Biology I, BioSC, RWTH Aachen University, Worringer Weg 3, 52074 Aachen, Germany

<sup>21</sup>Royal Botanic Gardens, Kew, Richmond, Surrey, TW9 3AE, United Kingdom



## Abstract

- Enabling data reuse and knowledge discovery is increasingly critical in modern science, and requires an effort towards standardising data publication practices. This is particularly challenging in the plant phenotyping domain, due to its complexity and heterogeneity.
- We have produced the MIAPPE 1.1 release, which enhances the existing MIAPPE standard in coverage, to support perennial plants, in structure, through an explicit data model, and in clarity, through definitions and examples.
- We evaluated MIAPPE 1.1 by using it to express several heterogeneous phenotyping experiments in a range of different formats, to demonstrate its applicability and the interoperability between the various implementations. Furthermore, the extended coverage is demonstrated by the fact that one of the datasets could not have been described under MIAPPE 1.0.
- MIAPPE 1.1 marks a major step towards enabling plant phenotyping data reusability, thanks to its extended coverage, and especially the formalisation of its data model, which facilitates its implementation in different formats. Community feedback has been critical to this development, and will be a key part of ensuring adoption of the standard.

## Introduction

The volume of data being generated in the life sciences demands good data management practices to enable reusability. While it is common practice to publish standardised sequencing data in public repositories, other data types are often only made available through scientific publications, and can be hard to find (Vines *et al.*, 2014), interpret or reuse. In a survey with over a 1000 participants, more than half agreed that lack of access to data is a 'major impediment to progress in science' and 'has restricted their ability to answer scientific questions', and most pointed out that data may easily be misinterpreted due to its complexity or poor quality (Tenopir *et al.*, 2011). In the plant phenotyping domain, data reuse is both pressing and challenging (Ćwiek-Kupczyńska, 2018; Spindel and S. R. McCouch, 2016; Tardieu *et al.*, 2017). On the one hand, the development of automated high-throughput and high-resolution technologies has contributed to a scale-up in the number, complexity and size of plant phenotyping datasets. This has been amplified by the increasing number of long-term, highly multilocal phenotyping networks aiming to decipher the interaction between genotype and environment (Millet *et al.*, 2019). Conversely, the reuse and meta-analyses of phenotyping data are particularly challenging due to the heterogeneity of this domain that encompasses many types of experimental sites (field, glasshouse, controlled environment), plants (crops, forest trees), collected data (images, physical measurements, chemical assays, molecular biology assays), and experimental designs (factors being tested, timing, field layouts, etc.). Furthermore, plant phenotype hinges not only on the interaction between genotype and environment, but also developmental stage and epigenome status (King *et al.*, 2010), which raises the challenges of integrating genotypic and phenotypic data (Pommier *et al.*, 2019b).

A successful example of data reuse in this domain is the study by Hurtado-Lopez, 2012, who reused field trial datasets and integrated them with quantitative trait locus (QTL) data to yield novel insights into genotype by environment (GxE) interactions in potato. Because the original experimental data followed no standardisation guidelines, the authors had to manually assemble detailed metadata during the preprocessing of the data from descriptions in unstructured text. To facilitate such studies, so that they may become the norm rather than the exception, it is essential that the scientific community adopt good data management and publication practices (Zamir, 2013).

The requirements for data reuse in science have been formalised in the FAIR data principles (Wilkinson *et al.*, 2016). They state the criteria that scientific data must fulfil to be findable, accessible, interoperable and reusable by both humans and machines, which hinge on having rich, harmonised, machine-readable, high-quality metadata describing the data as explicitly and objectively as possible.

Four key components are needed from research communities to meet these requirements: metadata standards which list the fields required for interpreting the data from a given experimental domain; machine-readable (meta)data exchange formats in which to express and share the (meta)data; ontologies or controlled vocabularies to describe (meta)data values and ensure that they are objective, consistent and unambiguous across datasets; and searchable data repositories with a well-established protocol for machine access.

The need for a metadata standard was first recognised in the life sciences by

the microarray community, who developed MIAME (Minimum Information About a Microarray Experiment) (Brazma *et al.*, 2001). This was soon followed by similar standards for other domains (e.g. Field *et al.*, 2008; Bustin *et al.*, 2009; Lapatas *et al.*, 2015) as can be seen on FAIRsharing (Sansone *et al.*, 2019). In the plant phenotyping domain, the need for metadata to document experiments was initially addressed independently by the developers of phenotyping databases, such as BreedBase (BreedBase team, 2020), GnplS (Steinbach *et al.*, 2013), (PIPPA team, 2020) and Plant Hybrid Information System (PHIS) (Neveu *et al.*, 2019), which resulted in a multitude of implicit, often database-specific standards. However, the need for an explicit consensus to enable interoperability between these databases brought this community together to develop MIAPPE (Minimum Information About a Plant Phenotyping Experiment), the first and so far only community metadata standard for the plant phenotyping domain Krajewski *et al.*, 2015.

MIAPPE had three guiding principles: to minimise the chance of a researcher missing important information in the documentation of an experiment; to support the annotation of content with community-relevant vocabularies; and to promote a data format implementation. MIAPPE marked a critical step towards the FAIRness of plant phenotyping data, as concluded in a survey of c. 50 citations of this standard in publications and web portals (Krajewski and Ćwiek-Kupczyńska, 2020). However, there were aspects to improve, such as the coverage, usability and clarity of the standard. In particular, MIAPPE lacked fields needed to capture experiments with woody plants, as it was conceived primarily with crop plants in mind, and it lacked an explicit data model, which left some researchers struggling to understand how to represent their experiments.

The microarray community was again among the first to produce a machine-readable (meta)data exchange format in the form of MAGE-Tab (MicroArray Gene Expression tabular) (Rayner *et al.*, 2006), a standardised format for MIAME. This gave rise to the broader-purpose ISA-Tab (Investigation/Study/Assay tab-delimited) format (Rocca-Serra *et al.*, 2010; Sansone *et al.*, 2012), which was adopted by more domains, including plant phenotyping, with an ISA-Tab implementation of MIAPPE (Ćwiek-Kupczyńska *et al.*, 2016)

The use of ontologies and controlled vocabularies in the life sciences dates back to Linnaeus's taxonomy, but they have witnessed a more recent boom after the creation of the Gene Ontology (The Gene Ontology Consortium, 2019), and currently number in the several hundred, as seen on BioPortal (Noy *et al.*, 2009). For the plant phenotyping domain, there are a number of ontologies that cover different key aspects. The Crop Ontology (R. Shrestha *et al.*, 2012) models plant traits and methods for assessing them in several species-specific ontologies. It merits special reference, in that it aims at standardising the methods used by data producers for phenotyping and the way they are reported, rather than only at terminological standardisation. It therefore includes an implicit metadata standard, the trait-method-scale trio, which was incorporated into MIAPPE. The Planteome project (Cooper *et al.*, 2018) developed three key ontologies: the Plant Trait Ontology (Arnaud *et al.*, 2012) modelling species-independent plant traits under a broader scope than the Crop Ontology and serving as a reference ontology for multispecies analyses; the Plant Ontology (Jaiswal *et al.*, 2005) covering plant anatomical structures and development stages and enabling interplant comparisons;

and the Plant Experimental Conditions Ontology (Cooper *et al.*, 2018) describing plant treatments. In addition to these, relevant ontologies include: the Agronomy Ontology (Aubert *et al.*, 2017) covering agronomic practices, techniques and variables; the Environment Ontology (Buttigieg *et al.*, 2013) describing natural environments; and the Statistics Ontology (Statistics Ontology Project, 2020) devoted to statistical methods. All these ontologies and several others are indexed in AgroPortal (Jonquet *et al.*, 2018) which serves as the reference repository and search service for plant-related ontologies.

Finally, while searchable data repositories have long been the norm in the life sciences, especially concerning gene and protein data, only in the last decade has it become common practice to enable machine access to their data via application programming interfaces (APIs). Currently all major databases, such as GenBank (Benson *et al.*, 2000) or UniProt (The UniProt Consortium, 2019), provide such access, but most smaller databases do not. This was the case for the plant phenotyping domain up until recently, with its numerous, independent and heterogeneous local databases. To address this problem and enable interoperability between databases, the plant community undertook the development of the Breeding API (BrAPI) (Selby *et al.*, 2019), a common API for data search and retrieval that can be implemented by plant breeding databases irrespective of their internal data model. Like the databases it aims to connect, BrAPI also has an implicit (meta)data model that aims to reconcile the metadata available in existing databases, spanning organisational metadata, plant phenotypic (meta)data and genotypic (meta)data. BrAPI was initiated independently from MIAPPE, so while there is substantial overlap between the two resources, there are also a few key differences in their metadata fields, as well as differences in terminology.

The way forward for enabling FAIR plant phenotyping data lies in bringing together all of the components described above. MIAPPE would be the cornerstone of such an architecture, specifying the metadata that is needed and connecting metadata fields to the ontologies recommended to fill them, as well as reconciling the several implicit metadata models of existing knowledge resources. BrAPI would serve as the means for federating the many independent plant phenotyping databases to enable findability and accessibility, and should enforce and validate the MIAPPE compliance of datasets. The MIAPPE ISA-Tab implementation would support data publication and exchange. And potentially all of the ontologies listed above would play a role in describing the data and metadata of plant phenotyping experiments in a standardised and unambiguous way. However, it is clear that further development effort on these resources is needed to attain such a goal.

In this paper, we detail the efforts of an international consortium to enhance the MIAPPE standard towards enabling FAIR plant phenotyping data. We describe the following refinements: (i) the extension of MIAPPE to accommodate a wider range of use cases (including those relevant to perennial and woody plants); (ii) the specification of a data model underlying the standard, to facilitate its interpretation and usage; (iii) the formalisation of MIAPPE in a computer-interpretable format (using the Web Ontology Language, OWL) to enable dataset validation and computational analysis; and (iv) the alignment of MIAPPE and BrAPI to enable the exposure of MIAPPE-compliant datasets via BrAPI endpoints.



# Materials and Methods

## Development of MIAPPE 1.1

To take on the challenge of improving MIAPPE, the community gathered both life and computer scientists. The former drove the documentation and description of the standard, ensuring that the terms and definitions are meaningful and not purely technical. The latter took the initiative for the technical aspects, involving data formalisation, organisation, integration, sharing and interoperability. This ongoing partnership ensures that MIAPPE bridges the domains of life and data science and addresses the needs of both communities.

The development of MIAPPE 1.1 was carried out collaboratively using simple and efficient protocols and format (spreadsheet). Throughout the process, drafts were presented and discussed with the international community through consultations by emails, the MIAPPE consortium GitHub issue tracker (MIAPPE contributors, 2020a) and during ‘bring your own data’ training sessions.

Like its predecessor, MIAPPE 1.1 is a metadata standard that formally organises the documenting of a phenotyping dataset, including environmental aspects. It primarily structures the metadata, imposing no constraints on the data itself (which may consist of images, other binary data, tabular files, etc.).

In comparison with MIAPPE 1.0, MIAPPE 1.1 introduces several new concepts while preserving most of those already present. The major change, however, is that it moved from a simple checklist to a fully formalised data model that makes explicit mandatory information, restrictions and expectations and thus represents a major improvement in clarity from MIAPPE 1.0.

The key changes in MIAPPE 1.1 fall into one of three categories, which are detailed in the following subsections: scope extension, interoperability and data model specification. In addition to these, the MIAPPE data model has been formalised in OWL as the Plant Phenotyping Experiment Ontology (PPEO). Note: throughout the rest of this document, we use *italics* to denote MIAPPE concepts, `<angle brackets>` to denote ontology concepts, and “double quotes” for MIAPPE field value examples.

## Scope extension

The scope of MIAPPE 1.0, mostly restricted to field crops, was extended in MIAPPE 1.1 to encompass woody plants, mainly by enabling the identification of plant materials by their geolocation coordinates, which are typically used to identify forest trees instead of plant identifiers (e.g. GenBank accession numbers) used in crop research. Two levels for plant material identification and description are available in MIAPPE 1.1: (i) its identification within the experiment (*biological material*); and (ii) its identification before the experiment (itmaterial source), allowing for individual plants, lots or progeny to be described and related to previously published or publicly accessible material. Furthermore, the *preprocessing* field (previously called *pretreatments*) can describe any type of action performed on the *material source* before it is used as the experimental *biological material* (for instance “tree transplantation” and “grafting”).

## Interoperability

MIAPPE 1.1 incorporates several metadata standards and practices that cover parts of its scope, in order to ensure interoperability and avoid remodelling and redefining aspects that are already well established: the generic metadata fields (e.g. identifier, title, version, date) in MIAPPE 1.1 are largely based on the DataCite metadata model of Dublin Core (DataCite Metadata Working Group, 2014); the fields for biological material identification are based on the Multi-Crop Passport Descriptors (MCPD) v.2.1 (Alercia *et al.*, 2015); the observation unit concept and its fields were imported from BrAPI and GnplS-Ephesis (Pommier *et al.*, 2019b); and the observed variable section is largely based on the data model of the Crop Ontology.

Additionally, to further foster interoperability, MIAPPE 1.1 includes precise definitions and examples for each of its fields, with recommendations for the use of controlled vocabularies, ontologies and ISO norms whenever appropriate. For example, the ISO 8601 norm is recommended for dates, Crop Ontology terms are recommended in the observed variable section, and Plant Ontology terms are recommended for characterising samples. These definitions and recommendations clarify the intended usage of MIAPPE 1.1 in a way that is accessible to biologists and breeders, while promoting compliance with the FAIR principles.

## Data model specification

The specification of a data model was essential to clarify MIAPPE's structure, and improves its internal consistency. The construction of a formal model helped: (i) ascertain the roles and relationships of the MIAPPE 1.0 checklist's main categories and concepts; and (ii) extend those concepts to a broader range of experiments.

Objects in the MIAPPE 1.1 data model correspond to sections in the MIAPPE 1.1 checklist. A schematic view of the data model is presented in **Figure 2.1**.

The MIAPPE data model is reconciled with the more generic data models underlying the ISA-Tab exchange format (Rocca-Serra *et al.*, 2010) through key objects such as Investigation and Study (ISA) and specialised representations such as BrAPI (Selby *et al.*, 2019) with entities such as Observation unit and Observation variable (BrAPI).

## Data model formalisation (PPEO)

While the specification of the MIAPPE data model addresses the concern of improving the clarity of the standard for users, it does not address machine readability, which is important to enable validation and facilitate implementation at scale. The need for the latter led us to encode the MIAPPE standard in OWL as the PPEO (Pommier *et al.*, 2020).

In PPEO, each MIAPPE section is encoded as an ontology class, with additional classes declared to group linked MIAPPE fields (e.g. `<method>` groups the linked fields *method description* and *method accession number*). Each MIAPPE field is encoded as an ontology data property, which specifies the type of value expected (e.g. `<has collection date>` for class `<sample>`, which must take a date-time value). The relations between classes are formalised through object properties (e.g. `<has biological material>` connects `<observation unit>` to `<biological material>`). Cardinality restrictions imposed by the

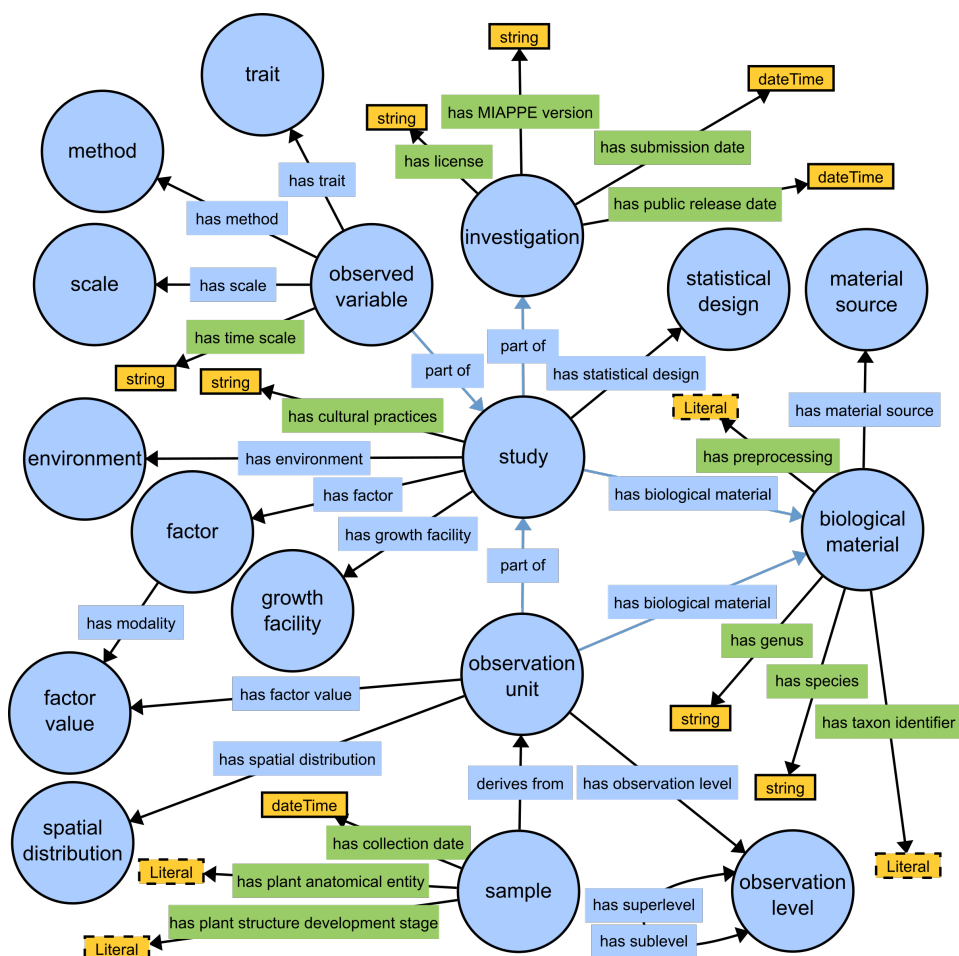
data model are encoded as ontology restrictions on the corresponding classes (e.g. an <investigation> must have at least one <study>). Ontology usage recommendations are encoded as annotations. Finally, to facilitate the implementation of MIAPPE in various forms, PPEO includes labels expressing corresponding names of each class in different resources (BrAPI, ISA-Tab).

**Figure 2.1** represents a subset of the PPEO.

## MIAPPE 1.1 overview

The MIAPPE sections, which correspond to objects in the data model, are the following:

- *Investigation* – the entry point of each MIAPPE dataset. It contains several general metadata fields (e.g. *title*, *description*, *submission/publication dates*), including some critical for FAIRness (*unique identifier*, *license*, *MIAPPE version*). One or more *publication* may be associated with the *investigation*.
- *Study* – corresponds to one experiment and defines its location (which by definition must be single per *study*) and duration. It lists general fields documenting the experiment (e.g. *experimental design*, *cultural practices*, *growth facility*). Like the *investigation*, it contains a *unique identifier* field. An *investigation* must have one or more *studies*.
- *Person* – contains contact details for each contributor of an entire *investigation* or an individual *study*, including the *role* of the *person*.
- *Data file* – references a data file of the MIAPPE dataset (e.g. a tabular file containing the results of observations, an image file), which may be attached to the dataset (referenced by name) or available in an online repository (referenced by URL). A *version* and a *description* must be provided for each *data file*. A *study* may have any number of *data files*.
- *Biological material* – identifies and describes the plant materials used in the *studies*. Plant materials must be identified through a *biological material ID* field, which can be institution-specific or platform-specific (e.g. seed lot number for annual plants, clone number for perennials or an experimental plant ID), and is recommended to follow the MCPD convention of holding institute identifier (FAO WIEWS code) plus a unique identifier of the individual plant material provided by that institute. They must also be identified through the *organism* field, which indicates the unique taxonomic identifier of the *biological material* in a standard such as the NCBI taxonomy. Optionally, they may be identified through the fields *genus*, *species* and *infraspecific name*, where textual names are expected (but should follow accepted standards). They may also be identified through geographical coordinates (i.e. *latitude*, *longitude*, *altitude*, and *coordinates uncertainty*), as is common for forest trees. The *biological material preprocessing* describes the *biological material* pretreatments, applied (e.g. to the seeds, or the tree cuttings) before the beginning of the experiment. Finally, the *material source* fields identify the origin or provenance of the *biological material* (e.g. gene bank accession, *in situ* material like an orchard, tree material provenance including forest wild site, laboratory-specific populations). These fields include the *material source*



**Figure 2.1:** Subset of the Plant Phenotyping Experiment Ontology representing the MIAPPE data model. Generated using WebVOWL (<http://editor.visualdataweb.org/>) and edited manually. Circles indicate classes. Object properties are shown in blue rectangles, and data properties are shown in green rectangles. Yellow rectangles represent literals.

*ID* (which follows the same recommendations as the *biological material ID*), the *material source DOI* (for referencing material sources listed in repositories), four geographical coordinates fields (same as for *biological material*), and finally a textual description. The *biological material* section thus covers a minimal subset of the MCPD standard used by gene banks, while also enabling interoperability and data linking through the use of identifiers (namely NCBI taxonomy identifiers) both between MIAPPE-compliant datasets and with external datasets. Moreover, through the provision of external identifiers to resources detailing their *biological material* (e.g. DOIs, accessions to gene banks or genome archives) researchers can encompass additional information, such as extended MCPD information (e.g.

synonyms, genealogy) and genotypic information. Last but not least, with these additions, MIAPPE 1.1 can handle cases such as forest tree clonal trials, where the plants identified solely through *biological material coordinates* in one *study* are used to generate new plant material for another *study*, in which their identification is done by specifying the location of the *material source*.

- *Environment* – describes a management practice parameter (e.g. sowing density, rooting medium composition) that was kept constant throughout the *study* across all *observation units* (to be described later). It applies to the whole *study* and has only a *type (parameter)* and a *value*. There can be discrepancies between intended environmental settings (e.g. target temperature in a glasshouse) and actual measurements of environmental *observed variables* (e.g. hourly temperature measured with four sensors). A *study* may have any number of *environments*.
- *Experimental factor* – describes a management practice that varied between *observation units* in a *study*, assessing the effect of which is the object of the *study*. *Experimental factors* can be biotic or abiotic (e.g. pest, disease interaction, cultural practice) and are characterised by a *type*, a *description* and a list of possible *values*. For instance, a “drought” *experimental factor* can discriminate “rainfed” and “irrigated” blocks, and a “nitrogen input level” can identify groups of plants under “high nitrogen input”, “low nitrogen input” and “no nitrogen inputs”. A *study* may have one or more *experimental factors*.
- *Event* – describes a discrete occurrence at a specific time that affected the whole *study* or one or more *observation units*, which can be the application of a field/glasshouse practice (e.g. planting, fungicide application) or an unpredictable happening (e.g. rainfall, pathogen attack). *Events* allow a general traceability of the conditions/events, and have been adopted since their usefulness was successfully demonstrated in the PHIS (Neveu *et al.*, 2019). *Events* include a *type*, ideally taken from an ontology such as the Crop Research Ontology (R. Shrestha, 2020) or the Agronomy Ontology (Aubert *et al.*, 2017), a *date* and a *description*, but no dedicated field for categorical or numerical values. *Events* can be repeated through time (e.g. to capture repeating cultural practices, such as adding fertiliser) by duplicating the *type* and *description* while providing a new *date*.
- *Observation unit* – is the experimentation object on which phenotypic and environmental parameters are measured and to which *experimental factors* are applied. It is characterised by a *type* or level, which can be a single “plant”, a group of plants (“pot”, “plot”, “block”), or the whole “study”. These *types* are hierarchical, meaning that we can have *observation units* and corresponding observations made from the *study* level down to the plant level. In some cases, an *observation unit* may contain no plant (e.g. raw plots after harvest or areas of a forest without tree), but can still be the object of environmental observations. Optionally, an *observation unit* can have a cross-reference to an external database, such as BioSamples (Courtot *et al.*, 2019). Also optionally, it can have one or more *spatial distribution key-value* pairs that locate the *observation unit* in the experimental

hierarchy (e.g. “block: 1”) or globally (e.g. “latitude: +43.619261”). A *study* should have one or more *observation units*.

- *Sample* – represents subplant material that was physically collected from an *observation unit* and was stored and processed before observations are made on it (e.g. in molecular studies). When traceability of *sample* processing is not needed, subplant observations can be assigned directly to the corresponding plant-level *observation unit* without the use of a *sample* as an intermediary. In such cases, the *observed variable* should describe that the observation is made on a plant part (e.g. “leaf chlorophyll content”, “grain protein content”) and include additional information on how the sampling was made in the textual description. The *sample description* field contains a free text description such as organism count, oxygenation, salinity or storage attributes. The *plant anatomical entity* and the *plant structure development stage* give more details on the *sample* properties at the time of sampling, which is specified with the field *collection date*. A *sample* must be derived from a single *observation unit*, but each *observation unit* may have any number of derived *samples*.
- *Observed variable* – documents a phenotypic or environment parameter that was observed and recorded as part of the *study*. It follows the Crop Ontology model of representing variables as combinations of a *trait*, a *method* and a *scale*. *Trait* details the characteristic being observed/measured (e.g. “plant height”). *Method* describes the procedure used in the observation/measurement (e.g. “with a measuring tape, starting at ground level”). *Scale* indicates the unit or scale with which observations/measurements were recorded (e.g. “cm”). *Observed variables*, *traits*, *methods* and *scales* are each identified by name, and may have a reference to the corresponding ontology concept (ideally from the Crop Ontology). *Observed variables* also have an *ID* by which they are referenced in the data file. *Methods* can also have a *description* plus an *additional reference*, usually from the literature. The *time scale* indicates the unit of time (e.g. “date-time”, or “growing degree days”) used to timestamp observations of this *observed variable*.

Note that in MIAPPE 1.1, the description of environment aspects is broken into several sections so as to allow flexibility in capturing and representing environment information: *environment* (fixed parameters throughout the *study*), *experimental factors* (fixed set of values for the *study* which vary between *observation units*), *observed variables* (measured during the *study*) and potentially *events* (discrete occurrences such as heavy rain).

## MIAPPE implementations

MIAPPE is a general specification that needs to be adopted and implemented by data repositories and exchange tools if it is to be easily usable. The MIAPPE 1.1 update encompasses four major implementations which are discussed in the following subsections: (i) an ISA file archive backed by an updated ISA-Tab configuration, developed in collaboration with the ISA Framework team; (ii) a web service implementation through BrAPI, developed in close collaboration between the MIAPPE and BrAPI communities; (iii) a spreadsheet template developed and used as training material to introduce

biologists to MIAPPE, which can also be used for simple metadata exchange; and (iv) finally, an RDF implementation based on the PPEO.

## MIAPPE ISA-Tab

The ISA Framework encompasses a model and a set of serialisations (TAB, JSON and RDF) to describe the experimental metadata with links to data files, code, articles and other digital objects. It comes with a suite of associated tools, is extensively used in the life sciences (The ISA Team, 2020), and among the endorsed resources of the ELIXIR Interoperability Platform.

MIAPPE 1.0 already included an ISA-Tab implementation in the form of a configuration file. This implementation has been revised in view of the changes in MIAPPE 1.1, and the configuration file has been updated accordingly. This configuration (ISA-Tab for plant phenotyping contributors, 2020) can be used with the ISA Creator tool, to more easily produce MIAPPE-compliant ISA-Tab archives.

The overview of the mapping between MIAPPE 1.1 and ISA-Tab sections is shown in **Table 2.1**. The Investigation and the Study express the same concepts in both MIAPPE and ISA-Tab, and many of their fields are listed under the corresponding sections. There are also direct correspondences for *experimental factors* (Study Factors), *biological material* (Source), *observation units* (Samples) and *samples* (Extracts). The remaining MIAPPE-specific fields are stored as ISA-Tab Comments. ISA Protocols must include a protocol named “Growth” holding the MIAPPE *cultural practices* field and *environment parameters*, one protocol for the “Phenotyping” process, plus an optional list of protocols with Type “Event” to handle MIAPPE *events*, with specific occurrences listed in an external Events file. Finally, the ISA Sampling Protocol indicates the derivation of a MIAPPE *sample* from an *observation unit*. Each ISA-Tab Assay represents one data file measured at one observation level. *Observed variables* are listed in the trait definition file, and referenced in the data files. The data files are formatted according to the common practices of the domain and contain references to that *Variable ID*, the measured values and times plus any information which researchers might deem useful.

## Breeding API (BrAPI)

The Breeding API (BrAPI) (Selby *et al.*, 2019) is a RESTful API developed by an international open-source community for querying plant breeding data, already implemented by several databases, and selected by the European biological data infrastructure ELIXIR as the cornerstone of its plant data search service. It is therefore critical that BrAPI and MIAPPE be compatible.

The collaboration between the BrAPI and MIAPPE teams has aimed at ensuring the compatibility of the two schemas, and the latest BrAPI (v.1.3) covers most of MIAPPE. The main sections correspond to the same concepts and either share the same name or have direct correspondence, such as investigation (BrAPI Trial) and biological material (BrAPI Germplasm). Some MIAPPE sections are currently absent from BrAPI (e.g. environment, event) but have already been proposed as additions to BrAPI and are under consideration for the next major release. The mapping between MIAPPE and BrAPI is overviewed in **Table 2.2**.

| MIAPPE section      | ISA-Tab section  | ISA-Tab section specification                    |
|---------------------|--|--|
| Investigation       | Investigation/<br>investigation publications           |  |
| Study               | Study/<br>study design descriptors/<br>study protocols |  |
| Person              | Investigation contacts/<br>study contacts              |  |
| Data file           | Study  | With comment fields                              |
| Biological material | Source   |  |
| Environment         | Study protocols  | Growth type protocol                             |
| Experimental factor | Study Factors  |  |
| Event               | Study protocols  | Event type protocols and<br>external Events file |
| Observation unit    | Sample   |  |
| Sample              | Extract/study protocols                                | Sampling type protocol                           |
| Observed variable   | Observed variable                                      | In external trait definition<br>file             |

**Table 2.1:** Mapping between MIAPPE and ISA-Tab sections. The table lists the MIAPPE sections with the ISA-Tab sections holding their fields. MIAPPE-exclusive fields have been added as comments in the corresponding sections. The detailed mapping can be found in **Supplementary table S2.1**, and in the MIAPPE repository ([https://github.com/MIAPPE/MIAPPE/tree/master/MIAPPE\\_Checklist-Data-Model-v1.1/MIAPPE\\_mapping](https://github.com/MIAPPE/MIAPPE/tree/master/MIAPPE_Checklist-Data-Model-v1.1/MIAPPE_mapping)).

Finally, BrAPI datasets can be exported as MIAPPE-compliant ISA-Tab archives using the BrAPI2ISA tool (BrAPI2ISA contributors, 2020).

Spreadsheet template

The spreadsheet template for MIAPPE (MIAPPE contributors, 2020b) was developed mainly for training purposes, as a simpler alternative to ISA-Tab. It is an explicit representation of MIAPPE, where each section has been placed in a separate worksheet. This template facilitates the understanding of the connections between documentation, data model and actual data. For training, it is important that the data model and one-to-many relationships (Fig. 1) be explicitly presented and comprehensively explained to the users (e.g. biologists or data managers).



| MIAPPE              | BrAPI object          |
|---------------------|-----------------------|
| Investigation       | Trial                 |
| Study               | Study                 |
| Person              | Contact               |
| Data file           | Data link             |
| Biological material | Germplasm             |
| Environment         | Environment parameter |
| Experimental factor | Treatment             |
| Event               | Events                |
| Observation unit    | Observation unit      |
| Sample              | Samples               |
| Observed variable   | Variable              |

**Table 2.2:** Mapping between MIAPPE sections and BrAPI objects. The table lists the MIAPPE sections with the BrAPI objects holding their fields (in the current and future versions). The detailed mapping for each field can be found on the MIAPPE GitHub repository and in **Supplementary table S2.1**, and in the MIAPPE repository ([https://github.com/MIAPPE/MIAPPE/tree/master/MIAPPE\\_Checklist-Data-Model-v1.1/MIAPPE\\_mapping](https://github.com/MIAPPE/MIAPPE/tree/master/MIAPPE_Checklist-Data-Model-v1.1/MIAPPE_mapping)).

## RDF based on the PPEO

PPEO was conceived to enable the direct expression of MIAPPE datasets in RDF (W3C, 2020b), by instantiating the ontology. Moreover, because PPEO explicitly maps MIAPPE to its implementations, it should be straightforward to convert MIAPPE datasets expressed in any of them to RDF.

MIAPPE and BrAPI are also connected through PPEO, which not only maps the two resources, but also includes classes exclusive to BrAPI, such as the `jobervation_i` class (which is outside of the scope of MIAPPE, as it pertains to data). A proof of concept has demonstrated the feasibility of producing linked data through BrAPI using the JSON-LD format (W3C, 2020a), with PPEO enabling the semantic mapping (see this dataset: Oury *et al.*, 2020a).

Having MIAPPE datasets in RDF enables the use of a wide range of available tools for reasoning and analysis, and facilitates data integration (by enabling data linking and cross-referencing at the semantic level) and validation.

## Results

To evaluate the applicability of the standard and the functionality of its implementations, plant scientists were asked to describe their phenotyping experiments using MIAPPE 1.1. The datasets were provided by: the Instituto de Biología Experimental e Tecnológica

(iBET), Portugal; the Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Germany; the Genetic and Genomic Information System (GnplS) of the Institut National de la Recherche Agronomique, France; and the Vlaams Instituut voor Biotechnologie (VIB), Belgium. The datasets are summarised in **Table 2.3**, and described in detail in **Supplementary Notes S2.1**. All of them are listed under (Papoutsoglou *et al.*, 2020b), and their files can be retrieved through the repositories listed there (Baute *et al.*, 2019a,b,c,d; Chaves *et al.*, 2020a,b; Junker, 2020; Junker and M. Li, 2020; Michotey *et al.*, 2020a,b; Oury *et al.*, 2020b,c; Pea *et al.*, 2019a,b).

The datasets span model, crop and perennial plants in a variety of experimental settings, as well as various MIAPPE 1.1 implementations. They demonstrate the ability of MIAPPE to handle diverse experimental designs, including automated glasshouses (IPK and VIB datasets), field networks for crops (GnplS) and forest trees (iBET and GnplS) with multiple scales and repetitions. Perennial plant use cases feature time series data, that is several observations across time for the same *observed variable* on the same plant. Field networks (GnplS wheat) demonstrate the use case of a multilocal and multiannual dataset where each location represents one *study* over several years. Several datasets demonstrate also the use of *experimental factors* such as cultural practices (nitrogen level in GnplS wheat) or experimental questions (covered or uncovered plants in IPK *Arabidopsis*). The *observed variables* proved to be well suited for very diverse destructive and nondestructive measurements adapted to agronomic (e.g. yield, grain weight), morphological (e.g. plant height), stress (e.g. disease or game), molecular (e.g. protein content) and physiological (e.g. photosynthetic efficiency) data. The data types covered by the *observed variables* are mostly numeric or textual, but also include images (IPK barley). These *observed variables* were described using references to ontologies (Michotey and Chaves, 2020; Michotey *et al.*, 2019; Pommier *et al.*, 2019a) whenever possible, but ad hoc variables were also used in specific cases not covered by ontologies. One of the most challenging aspects addressed by MIAPPE and successfully demonstrated by the datasets is the documentation of the *biological material*. The datasets clearly demonstrate how to organise information for model plants (IPK *Arabidopsis*), mutants (IPK barley), recombinant inbred line and population (VIB maize), GenBank reference accessions (GnplS wheat) and perennial plants including *in situ* material (iBET cork oak stands) or dedicated experimental locations acting as experimental tree fields with populations or crosses (GnplS poplar).

While the datasets showcase the applicability of MIAPPE to diverse experimental settings, they by no means represent the full extent of its coverage. Additional settings that were contemplated in the conception of MIAPPE but are not covered by the examples include: high-throughput phenotyping facilities with plants manipulated by conveyor belts, which produce large volumes of data with respect to the positions of plants and their development; precision agriculture field studies with drones and sensors capturing a wealth of data both about plant development and the environment; and cases where tracing the identity of plant materials is more complex. In the interest of demonstrating MIAPPE's coverage, **Table 2.4** presents additional examples of settings and details their modelling in MIAPPE.

| Publication                        | Biological Material  | Through-put | Plant Type  | Setting   | Dataset                                       |
|------------------------------------|--|-------------|-------------|---|---|
| Inácio <i>et al.</i> , 2017        | natural population<br>trees identified by<br>geographical location;<br><i>material source</i> not identified | low         | forest tree | field;<br>three locations   | Cork oak                                      |
| Junker <i>et al.</i> , 2015        | mutant;<br>multiple replicates   | high        | model plant | automated greenhouse,<br>controlled environment,<br>four experimental factors | <i>Arabidopsis</i>                            |
| M. Li <i>et al.</i> , 2019         | mutant and wildtypes   | high        | crop        | automated greenhouse  | Barley  |
| Oury <i>et al.</i> , 2018          | genebank material  | low         | crop        | multiyear multilocal field<br>network   | Wheat   |
| Monclus <i>et al.</i> , 2012       | clonal material with material<br>source traceability;<br>includes crosses and<br>populations                 | low         | forest tree | field   | Poplar  |
| Baute <i>et al.</i> , 2015         | recombinant inbred line (RIL)<br>population  | high        | crop        | greenhouse  | Maize<br>(Baute <i>et al.</i> , 2015)         |
| Dell'Acqua<br><i>et al.</i> , 2015 | recombinant inbred line (RIL)<br>population  | low         | crop        | field   | Maize<br>(Dell'Acqua<br><i>et al.</i> , 2015) |
| Baute <i>et al.</i> , 2016         | recombinant inbred line (RIL)<br>population  | high        | crop        | greenhouse  | Maize (Baute<br><i>et al.</i> , 2016)         |

**Table 2.3:** Overview of some characteristics of the example datasets. The experiments on the table encompass different experimental settings, plant types and throughput.

| No. | Scenario  | MIAPPE modelling   |
|-----|---|--|
| (1) | Heterozygous parent genotypes are used to derive a crossing population exhibiting significant phenotypic segregation.<br>Genotype tracing is necessary. | The cross of the parents is mentioned in the <i>material source</i> . Each of the progeny is treated as a <i>biological material</i> derived from the same <i>material source</i> , and is attributed a unique ID.   |
| (2) | Each tree in a field is observed through several sensors, at the roots and near its top.  | <i>Observation unit levels</i> : "plant". Each tree is a single <i>observation unit</i> . Each sensor measures one or many <i>observed variables</i> , (e.g. "Canopy temperature", "Cork thickness", ...)<br>An <i>observation unit</i> is created for the sensor.   |
| (3) | A sensor is placed in the middle of the field.  | No plants have to be present for an <i>observation unit</i> to be valid, as long as that <i>observation unit</i> is used to produce measurements or express <i>experimental factor values</i> .<br><i>Observation unit levels</i> :<br>"study" > "genotype" > "plot".  |
| (4) | Multilocal, multiyear field phenotyping network   | The whole network is an <i>investigation</i> .<br>Each location is a <i>study</i> over several years.<br>The <i>biological material</i> list is shared for the whole <i>investigation</i> . The list of <i>observed variable</i> definitions is also shared by all <i>studies</i> .<br>The measured data and observations can be at the "plant" or "plot" level, or as a per-genotype average within each <i>study</i> .<br>Study-level observations can be measurements from a meteorological station.<br><i>Observation unit levels</i> : any. |
| (5) | Time series of event or observation.  | <i>Study</i> type: any.<br><i>Observed variables</i> list the <i>time scale</i> they use.<br>In the data file, a single <i>observed variable</i> is measured several times, each value being timestamped in julian days, growing degree days or any other time scale.<br>The same applies with <i>events</i> with a given <i>event type</i> recorded several times at different time stamps.   |

**Table 2.4:** *Modelling possibilities for complicated experiment details. The table shows more specific scenarios that may be necessary to accommodate in MIAPPE, and the proposed modelling for them inside the standard.*

## Discussion

A global metadata standard is a key component for enabling FAIR data in any research domain, by providing a common framework under which researchers can describe their datasets with the necessary information for their interpretation, thus promoting interoperability and reusability. MIAPPE aims at serving such a role for the plant phenotyping community, and the first version of the standard took ample strides in that direction. In this work, we summarise the steps taken to extend and improve the usability of the standard.

The datasets presented in the Results demonstrate the broader applicability of MIAPPE 1.1, which was one of the main goals behind the update. The datasets span a variety of settings (e.g. woody, crop and model plants; glasshouses, single fields and field networks; single-year and multiannual experiments) and include aspects that could not be modelled under MIAPPE 1.0 (the most critical being the identification of *biological materials* using geographical coordinates).

MIAPPE 1.1 also has improved in flexibility and usability compared to the previous version. It has clearer definitions, examples, and when applicable, ontology recommendations for all fields. It has an explicit data model available in schematic form and encoded in OWL as PPEO. It has fewer mandatory fields, since not all of them are applicable to all experiments. It allows different strategies for modelling aspects such as environmental parameters: under the *environment* section, as *events* or as *observed variables*. The improved *biological material* description and the new *material source* can now handle gene banks and experimental collections with either the bare minimum identification or very detailed information, including infraspecific description, provenance, complex processing or identification mechanism. We received evidence for the improved usability of MIAPPE from the community, during two open requests for feedback and in several training sessions. The specification of the data model and the enriched definitions and examples were highlighted as clear improvements.

While MIAPPE promotes interoperability and reusability, the other two FAIR principles (findability and accessibility) rely mainly on BrAPI, which enables data search and retrieval through machine access. However, for these two resources to be part of a common scheme for enabling FAIR plant phenotyping data, it is necessary to ensure that BrAPI calls adequately cover MIAPPE and enable searches by all key MIAPPE fields. The process of reconciling MIAPPE and BrAPI was undertaken in parallel with the MIAPPE 1.1 update, through a collaboration between the BrAPI team, the ELIXIR Interoperability platform and Plant Sciences community, the EMPHASIS Plant Phenotyping Infrastructure and the CGIAR. BrAPI will be fully MIAPPE 1.1 compliant once its (currently beta) 2.0 release is finalised.

This reconciliation and the interoperability between the various MIAPPE 1.1 implementations is demonstrated in our Results, as most datasets are available in two different implementations (including BrAPI), in many cases through automatic conversion. This is critical, as MIAPPE aims to support a wide range of users and applications, from data submission by life scientists to data exchange, validation and even reasoning by machines. Formats supporting all these applications must not only be available but also be interconvertible.

While, from a technical standpoint, we believe that the merits of MIAPPE 1.1 speak

for themselves, we are well aware of the many hurdles ahead of getting any standard widely adopted by the community it seeks to serve. Indeed, there are several dozen standards currently deprecated in the FAIRsharing portal and surely many more have been lost to history.

One of the main factors behind wide adoption is having community engagement throughout the development process. Indeed, GO (The Gene Ontology Consortium, 2019) has been so successful because it emerged from the communities involved in gene function annotation for several model organisms and has remained open to input from the community throughout its history. The story behind MIAPPE is curiously similar, as its development gathered several researchers involved in plant phenotyping repositories, the process of updating it had extensive direct engagement with the community, and it remains open to community input through its GitHub repository (MIAPPE contributors, 2020a). MIAPPE 1.1 therefore gathers as close to a community-wide consensus as is possible to get and distil into a clear and well-organised standard, especially considering the heterogeneity and complexity of the plant phenotyping domain, and the difficulty in reconciling the perspectives of its different subdomains and experiment types. We will further foster its adoption through constant efforts of outreach and dissemination, to gather new communities and ensure the long-term usefulness of the standard.

Also critical for adoption is demand: when funders and/or publishers require compliance with a standard or data publication practice – such as depositing sequencing data in one of the public gene banks – it tends to be widely adopted. In the case of MIAPPE, the demand consists of the increasing pressure from funding agencies towards compliance with the FAIR data principles. Researchers working in plant phenotyping and seeking FAIR data solutions will be pointed towards MIAPPE thanks to its presence in the FAIRsharing portal (FAIRsharing.org: MIAPPE, 2020) and above all to the endorsement of ELIXIR, which led the MIAPPE 1.1 update and is helping shape the policies and lay the foundations needed for enacting the FAIR principles.

Equally critical is usability, as researchers tend to view the need for standardisation and reusability as a burden and often do as little effort as they can get away with when submitting a dataset, unless the process is virtually effortless. While MIAPPE's usability was improved with the 1.1 update, it is still missing an easy-to-use submission interface. For this reason, we are engaging with popular data management tools such as COPO (The COPO team, 2020) or FAIRDOME (Wolstencroft *et al.*, 2017) to incorporate MIAPPE and thus enable user-friendly MIAPPE-compliant dataset submission.

Last but not least, in order to persist, a standard must constantly evolve to keep up with technical and scientific advances. In this regard, the 1.1 update demonstrates that MIAPPE is very much a living standard, and we are already starting the next phase of development of MIAPPE. It will concentrate on two main aspects: extending its coverage of environmental aspects (initiated by the EMPHASIS members of the MIAPPE community) and facilitating the recording of technical aspects of material and data processing (e.g. sensors, cameras, software, configurations, calibrations), which are becoming increasingly important. Within this scope, another possible improvement could be to establish a formal complementarity with the ICASA standard, which is used by the agronomic and modelling communities, and provides variables for agronomic management practices, treatments, environmental conditions and measurements of crop responses (White *et al.*, 2013). There is overlap between these two standards –

with MIAPPE dedicated to plant phenotyping as used by geneticists, biologists and some agronomists and ICASA to 'any field experiment or crop production situation' – but the scope of each extends far beyond that of the other and there are obvious complementarities between them. User feedback from these endeavours will steer further developments, by revealing areas where improvement is desired by the community.

## Acknowledgements

This work was based on extensive reviews from and interactions with the broader MIAPPE community. We are grateful for all feedback. This work was supported by the European Union's Horizon 2020 programme through the ELIXIR-EXCELERATE project, Grant Agreement no. 676559, the EMPHASIS-Prep project, Grant Agreement no. 739514, and the research infrastructure for life science data ELIXIR. It was also funded by: the National Research Institute for Agriculture, Food and Environment (INRAE); the French Infrastructure en Biologie Santé 'Phenome-FPPN' supported by the French National Research Agency (ANR-11-INBS-0012); the Portuguese Fundação para a Ciência e Tecnologia through project BioData.pt (Grant no. 22231, co-financed by FEDER); the German-Plant-Phenotyping Network (DPPN) and the German Network for Bioinformatics Infrastructure (de.NBI), both funded by the German Federal Ministry of Education and Research (BMBF) (project identification number: 031A053, 031A536A,C). Philippe Rocca-Serra and Susanna-Assunta Sansone are funded by UK Research Councils (BB/L024101/1, BB/L005069/1), Wellcome Trust (212930/Z/18/Z, 208381/A/17/Z), European Union (H2020-EU.3.1, 634107, H2020-EU.1.4.1.3, 654241, H2020-EU.1.4.1.1, 676559), IMI (116060) and NIH Data Common Fund. In-kind contribution was made by the Crop Ontology project team supported by the Integrated Breeding Platform, by Planteome (NSF award IOS:1340112), and by the CGIAR Platform on Big Data for Agriculture. Some developments were supported by the RDA Europe project funded by the European Commission. Inês Chaves acknowledges DL 57/2016/CP1351/CT0003. Paweł Krajewski and Hanna Ćwiek-Kupczyńska acknowledge funding from EPPN2020 grant agreement no. 731013 and National Science Centre grant 2016/21/N/ST6/02358. Evangelia A. Papoutsoglou acknowledges a grant from WUR Plant Breeding. The authors declare that they have no conflicts of interest.



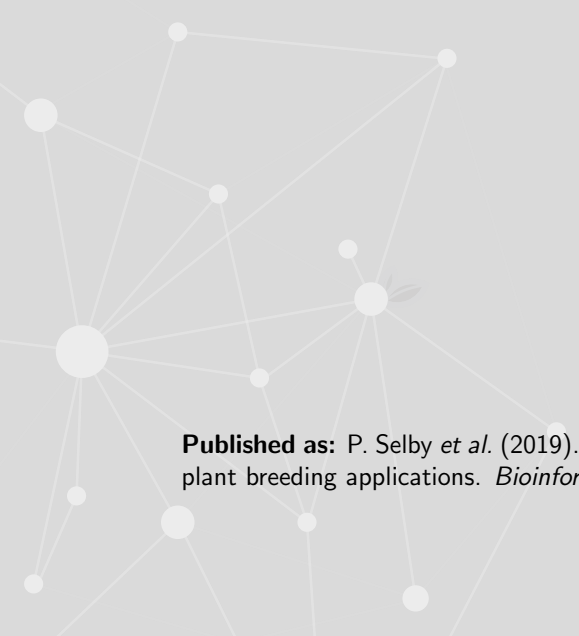


# Chapter 3

## BrAPI — An application programming interface for plant breeding applications

Peter Selby<sup>1</sup>, Rafael Abbeloos<sup>2</sup>, Jan Erik Backlund<sup>3</sup>, Martin Basterrechea Salido<sup>4</sup>, Guillaume Bauchet<sup>5</sup>, Omar E. Benites-Alfaro<sup>6,7</sup>, Clay Birkett<sup>8</sup>, Viana C. Calaminos<sup>9</sup>, Pierre Carceller<sup>10</sup>, Guillaume Cornut<sup>11</sup>, Bruno Vasques Costa<sup>12</sup>, Jeremy D. Edwards<sup>13</sup>, Richard Finkers<sup>14</sup>, Star Yanxin Gao<sup>15</sup>, Mehmood Ghaffar<sup>4</sup>, Philip Glaser<sup>15</sup>, Valentin Guignon<sup>16</sup>, Puthick Hok<sup>17</sup>, Andrzej Kilian<sup>17</sup>, Patrick König<sup>4</sup>, Jack Elendil B. Lagare<sup>8</sup>, Matthias Lange<sup>4</sup>, Marie-Angélique Laporte<sup>16</sup>, Pierre Larmande<sup>18</sup>, David S. LeBauer<sup>19</sup>, David A. Lyon<sup>5</sup>, David S. Marshall<sup>20,21</sup>, Dave Matthews<sup>8</sup>, Iain Milne<sup>20</sup>, Naymesh Mistry<sup>22</sup>, Nicolas Morales<sup>5</sup>, Lukas A. Mueller<sup>5</sup>, Pascal Neveu<sup>23</sup>, Evangelia Papoutsoglou<sup>14</sup>, Brian Pearce<sup>17</sup>, Ivan Perez-Masias<sup>6</sup>, Cyril Pommier<sup>11</sup>, Ricardo H. Ramírez-González<sup>24</sup>, Abhishek Rathore<sup>25</sup>, Angel Manica Raquel<sup>9</sup>, Sebastian Raubach<sup>20</sup>, Trevor Rife<sup>26</sup>, Kelly Robbins<sup>1</sup>, Mathieu Rouard<sup>16</sup>, Chaitanya Sarma<sup>25</sup>, Uwe Scholz<sup>4</sup>, Guilhem Sempéré<sup>10,27</sup>, Paul D. Shaw<sup>20</sup>, Reinhard Simon<sup>28</sup>, Nahuel Soldevilla<sup>3,29</sup>, Gordon Stephen<sup>20</sup>, Qi Sun<sup>15</sup>, Clarysabel Tovar<sup>3,29</sup>, Grzegorz Uszynski<sup>17</sup>, Maikel Verouden<sup>30</sup> and The BrAPI consortium<sup>2</sup>

In alphabetical order.



**Published as:** P. Selby *et al.* (2019). BrAPI — an application programming interface for plant breeding applications. *Bioinformatics*. DOI: 10.1093/bioinformatics/btz190

<sup>1</sup>Plant Breeding and Genetics Section, School of Integrative Plant Science, Cornell University, Ithaca, New York, USA

<sup>2</sup>VIB, Ghent, Belgium

<sup>3</sup>Integrated Breeding Program (IBP), CIMMYT, Texcoco, Mexico

<sup>4</sup>Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Gatersleben, Germany

<sup>5</sup>Boyce Thompson Institute, Ithaca, NY, USA

<sup>6</sup>International Potato Center (CIP), Lima, Peru

<sup>7</sup>International Food Policy Research Institute (IFPRI), Washington DC, USA

<sup>8</sup>USDA ARS, Ithaca, NY, USA

<sup>9</sup>International Rice Research Institute (IRRI), Los Baños, Laguna, The Philippines

<sup>10</sup>AGAP, Univ Montpellier, CIRAD, INRA, Montpellier SupAgro, Montpellier, France

<sup>11</sup>URGI, INRA, Université Paris-Saclay, Versailles, France

<sup>12</sup>Instituto de Biologia Experimental e Tecnologica (iBET), Oeiras, Portugal

<sup>13</sup>USDA ARS, Stuttgart, AR, USA

<sup>14</sup>Department of Plant Breeding, Wageningen University & Research, Wageningen, The Netherlands

<sup>15</sup>Institute of Biotechnology, Cornell University, Ithaca, New York, USA

<sup>16</sup>Bioversity International, Montpellier, France

<sup>17</sup>Diversity Arrays Technology, Bruce, Australia

<sup>18</sup>DIADE, IRD, University of Montpellier, Montpellier, France

<sup>19</sup>College of Agricultural and Life Sciences, The University of Arizona, Tucson, AZ, USA

<sup>20</sup>Information & Computational Sciences, The James Hutton Institute, Dundee, UK

<sup>21</sup>SRUC, Edinburgh, UK

<sup>22</sup>LeafNode Technology, Auckland, New Zealand

<sup>23</sup>MISTEA, INRA, Montpellier SupAgro, Université de Montpellier, Montpellier, France

<sup>24</sup>John Innes Centre, Norwich Research Park, Norwich, UK

<sup>25</sup>International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Hyderabad, India

<sup>26</sup>Department of Plant Pathology, Kansas State University, Manhattan, KS, USA

<sup>27</sup>INTERTRYP, Univ Montpellier, CIRAD, IRD, Montpellier, France

<sup>28</sup>Patranca E.I.R.L., Lima, Peru

<sup>29</sup>LeafNode Technology, Buenos Aires, Argentina

<sup>30</sup>Wageningen University & Research, Biometris, Wageningen PB, The Netherlands



## Abstract

### Motivation

Modern genomic breeding methods rely heavily on very large amounts of phenotyping and genotyping data, presenting new challenges in effective data management and integration. Recently, the size and complexity of datasets have increased significantly, with the result that data are often stored on multiple systems. As analyses of interest increasingly require aggregation of datasets from diverse sources, data exchange between disparate systems becomes a challenge.

### Results

To facilitate interoperability among breeding applications, we present the public plant Breeding Application Programming Interface (BrAPI). BrAPI is a standardized web service API specification. The development of BrAPI is a collaborative, community-based initiative involving a growing global community of over a hundred participants representing several dozen institutions and companies. Development of such a standard is recognized as critical to a number of important large breeding system initiatives as a foundational technology. The focus of the first version of the API is on providing services for connecting systems and retrieving basic breeding data including germplasm, study, observation, and marker data. A number of BrAPI-enabled applications, termed BrAPPs, have been written, that take advantage of the emerging support of BrAPI by many databases.

### Availability and implementation

More information on BrAPI, including links to the specification, test suites, BrAPPs, and sample implementations are available at <https://brapi.org/>. The BrAPI specification and the developer tools are provided as free and open source.

## Introduction

Plant breeding is widely recognized as crucial to feeding a rapidly growing population, especially in developing countries (Flavell, 2017), ([http://www.fao.org/fileadmin/templates/wsfs/docs/expert\\_paper/How\\_to\\_Feed\\_the\\_World\\_in\\_2050.pdf](http://www.fao.org/fileadmin/templates/wsfs/docs/expert_paper/How_to_Feed_the_World_in_2050.pdf)). To meet this demand, it is necessary to breed new varieties that maintain high productivity with reduced inputs and are adapted to new eco-agricultural environments resulting from climate change. Plant breeding is a complex undertaking that necessarily integrates many interrelated disciplines, each with their own conventions for data structure and storage, and increasingly large, multi-faceted datasets. To address the challenges in the size and complexity of breeding data, a number of database systems have been designed over the years to solve specific problems. Although the power and insights that can be gleaned from large datasets increase with a greater volume and diversity of data sources, these separate systems make data integration difficult. Breeders need seamless access to all relevant data, but each system tends to keep its data siloed with *ad hoc* formats that hinder the ability to exchange, compare and combine data across research teams.

To meet these requirements, numerous groups have been working together to create an Application Programming Interface (API) for breeding data (Ghouila *et al.*, 2018). An API specification describes the functions and services available in an application which can be accessed in an automated way by a computer program. It describes what services are available, what inputs are allowed, what the structure of the output data will be, and the protocol used to pass data to a service, often on the web. In recent years, web services have become the major paradigm for information exchange on the web, and web service standards have also been defined and implemented successfully by the bioinformatics community. Examples of such systems include the Distributed Annotation System (DAS) (Dowell *et al.*, 2001), BioMOBY (Wilkinson and Links, 2002), and the EMBRACE (Pettifer *et al.*, 2010) Web Service collection.

Most of the modern web service infrastructure follows the REST standards (Fielding and R. N. Taylor, 2002). REST stands for 'Representational State Transfer' and defines a stateless client/server communication architecture, built on the HyperText Transfer Protocol (HTTP) (<https://tools.ietf.org/html/rfc7231>). In a RESTful API, HTTP is the communication protocol and the available services are defined as Unified Resource Locators (URLs). Typically, the inputs are defined by constructing a URL with query parameters defined by the API (or HTTP request body objects for more complex inputs), the output data are usually returned in a defined structure. For the output, historically, XML was used, but newer APIs typically prefer the Javascript Object Notation (JSON) format.

Data exchange requires solutions on many levels, including the semantic level and the syntactic level (Doan *et al.*, 2004). For breeding data, standardization of the semantic level has made significant progress over the last few years through the definition of ontologies for describing plant structure and development (Cooper *et al.*, 2018), and for describing traits in popular crops (R. Shrestha *et al.*, 2012). However, the breeding community still needs to standardize data at the syntax level. This can be achieved by defining a standardized Application Program Interface.

Here, we report on the design and implementation of a standard RESTful Breeding API (BrAPI), as a specification with a focus on common plant breeding data requirements.

The interface was designed by members of the BrAPI consortium. A complete list of contributors is given in the consortium description and a continuously updated list can be found on the BrAPI website (<https://brapi.org/>).

## Results

The Breeding API is a practical tool to help solve problems in accessing, exchanging, and integrating data across systems and applications. Given the multidisciplinary nature of plant breeding, there is a broad range in the particulars of the possible data operations that could be considered. Since a complete list of BrAPI related use cases would grow unmanageably large, we decided to focus on a small number of main use cases to design the primary API elements with a view towards reusability in other use cases.

### Use cases

These are the main use cases we considered:

#### Field phenotyping apps

Trials are often performed in fields that have limited internet connectivity, requiring special solutions for collecting phenotypic data. A popular approach is to collect data using handheld devices paired with custom mobile apps (Rife and J. A. Poland, 2014). Information about the field, the plot and accession identifiers needs to be loaded on the device before phenotypic data collection. After completion, collected data need to be uploaded to the database, when internet connectivity is available. Currently available solutions require custom files to be transferred, often involving significant user intervention. However, a simpler method would be to use an API to retrieve and store the data directly from the database.

#### Sample tracking

For both phenotyping and genotyping applications, analyses may need to be run by service providers, such as analytical labs and genotyping centers, that use different tracking mechanisms. The sample tracking use case describes the hand-off of the sample information to the service provider, and the subsequent retrieval of the results. In practice, tracking samples can be complex because the identifiers from several different systems must be correlated.

#### Genome visualization and analysis

Genome-based breeding requires extensive genotyping, which can be helpful to visualize in different ways to aid in breeding decisions. An example of such a tool is Flapjack (Milne *et al.*, 2010), which can display a number of genotypes and run analyses on the data. BrAPI standardizes the interfaces for such tools, hence they can be used with a much wider range of data sources and without the need for special adaptations for each source.

## FAIR data portals

One of the challenges of big data is identifying datasets of interest and ensuring their long term availability. This can be addressed by building federations of Findable, Accessible, Interoperable and Reusable (FAIR) data repositories (Wilkinson *et al.*, 2016). Interfaces such as BrAPI can help such efforts by standardizing access to the data repositories, thereby creating federations. Portals to the federated data can then be deployed to provide general or community specific data access. This increases the visibility of all datasets and therefore reduces the risk of losing isolated datasets over time. The portals should implement simple searches on standard metadata, such as MCPD or MIAPPE (Ćwiek-Kupczyńska *et al.*, 2016; Krajewski *et al.*, 2015; Milne *et al.*, 2010).

## Data integration and exchange

In this use case, two databases exist with overlapping data as well as specific data in each database. Database A would like to access data in database B. For example, database A may contain information about accessions, such as phenotypic and trial metadata, while database B contains genotypic information. Using a BrAPI call, database A can extract the genotyping data from database B and use that data in breeding decision support.

## API definition

The BrAPI definition is kept in the 'API' repository of the 'plantbreeding' organization on GitHub (<https://github.com/plantbreeding/API>), with all changes to the definition managed using GitHub's 'issues', 'projects' and 'pull requests' facilities.

## API organization

BrAPI calls are organized into categories that reflect the major domains needed for exchanging data between plant breeding information management systems and client applications. Some example categories include Studies, Germplasm, Traits, Trials, MarkerProfiles and Authentication. (A full list of the categories is presented in **Table 3.1.**)

## URL structure

All BrAPI calls follow a common URL structure. The URL starts with a domain name (and optional base path of the implementation server) followed by '/brapi/' and the major version number. Next, the call name appears with optional object ids and other parameters. Most calls use the HTTP request method 'GET', but some require 'POST' and 'PUT', as specified in the documentation. For security, the use of SSL (HTTPS) is highly recommended for all BrAPI endpoints.

Examples:

```
https://example.com/brapi/v1/locations
```

```
https://example.com/brapi/v1/trials?programDbId=abc123
```

```
https://example.com/maize-db-01/brapi/v1/studies-search
```

| Category              | Comments  | # of calls |
|-----------------------|---|------------|
| Calls                 | Meta information about which BrAPI calls are available on a server implementation.  | 1          |
| Crops                 | Provides the common names for the crops available on a server implementation.   | 1          |
| Germplasm             | Provides search capabilities and details for germplasm data. Includes MCPD, pedigree and breeding method data.  | 8          |
| Germplasm Attributes  | Germplasm Attributes are simply inherited characterization descriptors that are inherent in the germplasm line but not environment-dependent.   | 3          |
| Markers               | Provides search capabilities and details for genetic marker metadata.   | 3          |
| Marker Profiles       | Provides search capabilities and details for genomic data. Includes allele matrices.  | 5          |
| Programs              | Provides search capabilities and details for breeding programs. A program may contain multiple trials.  | 2          |
| Trials                | Provides search capabilities and details for breeding trials. A trial may contain multiple studies. Used also for any large phenotyping dataset like multilocal phenotyping networks.                                       | 2          |
| Studies               | Provides search capabilities and details for genotyping and phenotyping studies and support for observation data gathering. Includes germplasm, observation, plot layout, and season details related to a particular study. | 17         |
| Phenotypes            | Provides search capabilities for phenotyping observation data across studies, trials, and programs.   | 5          |
| Traits                | Provides details for trait ontology data which are available for observation variables.   | 2          |
| Observation Variables | An Observation Variable is combination of a trait, a method and a scale. Phenotyping data are collected for observation variables. Fully aligned to the Crop Ontology.  | 5          |
| Genome Maps           | Provides summaries and details for stored genome maps.  | 4          |
| Location              | Provides details of geographical locations of studies.  | 2          |
| Samples               | Provides support for storing and retrieving plant sample metadata.  | 4          |
| Vendor Samples        | Provides support for sending sample metadata to an external vendor for processing (i.e. gene sequencing).   | 5          |

**Table 3.1:** *Categories of BrAPI calls*

## Return object structure

We have defined a standard JSON formatted response structure that is common across all calls. The standard response consists of a JSON object with a 'metadata' key and a 'result' key. The 'metadata' key provides the pagination information, an array of status information, and an array of data files. If the response data contain an array of entities which could possibly grow large, the 'pagination' object will be populated with the keys 'pageSize', 'currentPage', 'totalCount', 'totalPages' containing the appropriate values. If the response is a single entity that does not require pagination, then the 'pagination' object still must be returned, but all data elements within it should be set to zero. All pages are zero indexed, so the first page will always be page zero. The 'status' array contains a list of objects with the keys 'code' and 'message'. These status objects should be used to provide additional status or log information about the call. If the call was completed successfully and there are no status objects reported, an empty array should be returned. The 'datafiles' array contains a list of URLs to any extra data files generated by the call. For example, this could be images related to the data returned, or large data extract files which contain more data than that returned in the response payload, see **Figure 3.1** for an example.

The data payload 'result' contains the specific model object for the given call response. There are three basic patterns that response objects follow. The 'master' pattern is used for returning all the data associated with a single entity. The 'details' pattern is used to return an array of entities. In the 'details' pattern, the 'result' object always contains a single array called 'data' and no other fields. The 'master/details' pattern is a combination of the 'master' and 'details' patterns. It is used to represent a parent object which has an array of child entities. The 'result' object contains some data associated with the parent as well as the 'data' array with all the child entities. Whenever the 'data' array is present, the response is assumed to be paginated. This means the size of the 'data' array is always limited by the 'pagination' object in the 'metadata'.

In most cases, all the data will be contained within the JSON response. For large 'data' arrays, several requests might need to be made to retrieve every page of the array. In the event that the size of the data package exceeds what could reasonably be handled using the HTTPS protocol and the client, the service provider can place the data in a file and provide a link in the 'datafiles' array, to be downloaded later.

## Authentication

A user or system may have to authenticate to a server to access protected data. BrAPI is a data communication specification, so the authentication scheme used to protect that data is considered outside the scope of the BrAPI specification. However, authentication and authorization are important topics to address whenever any kind of data is moved or presented. In order to facilitate communication of data between tools in a standardized way, the BrAPI community has developed a set of best practices using the OAuth2 architecture for implementing proper authentication with any BrAPI enabled tools and databases. In its most basic form, the OAuth2 architecture is a sessionless, token based architecture. OAuth2 allows users to sign in with user credentials they already have, and provides a token. This token can then be used to authenticate that user within



```

{
  "metadata": {
    "pagination": {
      "totalCount": 20,
      "pageSize": 3,
      "totalPages": 7,
      "currentPage": 0
    },
    "status": [{
      "code": "200",
      "message": "Success"
    }]
  },
  "datafiles": ["/mnt/local/matrix_01.csv",
    "/mnt/local/matrix_02.csv"]
},
"result": {
  "key0": "master",
  "key1": 20,
  "key2": ["foo", "bar", "baz"],
  "data": [{
    "detailKey0": "detail0",
    "detailKey1": ["foo", "bar"]
  }, {
    "detailKey0": "detail1",
    "detailKey1": ["bar", "baz"]
  }, {
    "detailKey0": "detail2",
    "detailKey1": ["baz", "foo"]
  }],
}
}

```

**Figure 3.1:** An example BrAPI response object. This object shows a generic response with 'metadata', a 'master' result record and a set of 'data' records.

different tools and databases. The token should be added as a header in every BrAPI request.

## Versioning

All software projects need the ability to evolve to reflect changing requirements, to cover new use cases, and to incorporate user feedback. A well defined and rigorous versioning scheme is essential for BrAPI to ensure that client and server communication is well defined and the community can keep track of the changes. In BrAPI, there are major versions and minor versions. The major version is currently 'v1', which is reflected in the URL scheme. Minor versions are incremented about every three months, reflecting changes in the API that have been accepted by the BrAPI community and reviewed by the BrAPI coordinator. To help maintain consistency, all changes in minor versions are backward compatible with earlier minor versions within the same major version. The

'calls' call provides meta-information about each BrAPI call available on a given server. The response of the 'calls' call includes all the supported version numbers for each call, so external clients can easily check for compatibility with that server.

## Community

For a communication standard like BrAPI to be successful, there must be people and organizations willing to contribute and use it. Early on in the development of BrAPI, we recognized the need to foster and develop a strong community of users. This community has grown rapidly over the past few years and it now has representatives from several dozen different organizations from around the world.

The development of BrAPI is a community effort. Work on the API is mainly organized around regular 'hackathons', where BrAPI contributors gather for a week of discussions and API design work. BrAPI community institutions take turns organizing and hosting the hackathons. This has proven very effective for collaborative development and capacity building (Ghouila *et al.*, 2018). Between the hackathons, the proposed APIs are implemented at the different sites, and problems encountered during implementation are fed back into the design at the following hackathon. An important role in the community is played by the BrAPI coordinator, who helps to organize the hackathons and workshops, reviews and coordinate proposals for new or updated calls, provides support for implementers, and maintains the documentation and the BrAPI website.

### Brapi.org

To serve the developer community, a website (<https://brapi.org>) was created as a nexus of all BrAPI related tools and information. It provides the official documentation for the API as well as information on meetings, hackathons, community news, testing tools, development libraries, BrAPPs, and a community forum.

## Server implementations

BrAPI server implementations have been created for a number of popular breeding, genebank and plant genomics databases. A variety of languages and database systems have been used to develop BrAPI-compliant systems. Web frameworks' languages include Drupal/Tripal (PHP), Catalyst (Perl), Java Spring (Java), NodeJS (JavaScript), Django (Python) whereas databases and data query systems include Postgres, MongoDB, Elasticsearch, HDF5, and MySQL. Many of these systems are open source, so their code may be adapted for other systems with similar implementation parameters. A list of current BrAPI server implementations is given in **Table 3.2**.

| Database name   | URLs  | Organization, Reference   |
|---|---|---|
| Breeding Management System (BMS)  | <a href="https://www.integratedbreeding.net">https://www.integratedbreeding.net</a>         | CGIAR<br><a href="https://cgiar.org">https://cgiar.org</a>  |
| Description: comprehensive breeding management system with trial design, data collection, and analyses.   |   |   |
| Cassavabase   | <a href="https://cassavabase.org">https://cassavabase.org</a>                               | Boyce Thompson Institute (BTI),<br><a href="https://btiscience.org">https://btiscience.org</a>  |
| Musabase  | <a href="https://musabase.org">https://musabase.org</a>                                     |   |
| Yambase   | <a href="https://yambase.org">https://yambase.org</a>                                       |   |
| Sweetpotatobase   | <a href="https://sweetpotatobase.org">https://sweetpotatobase.org</a>                       |   |
| Solanaceae Genomics Network   | <a href="https://solgenomics.net">https://solgenomics.net</a>                               |   |
| Description: comprehensive breeding management system, including trial design management, phenotyping sample and data collection; with a focus on genomic breeding technologies such as Genomic Selection |   |   |
| B4R   | <a href="https://b4r.irri.org">https://b4r.irri.org</a>                                     | International Rice Research Institute (IRRI),<br><a href="http://irri.org">http://irri.org</a>  |
| Description: comprehensive breeding management system tailored for rice and other grains  |   |   |
| Germinate   | <a href="https://ics.hutton.ac.uk/get-germinate">https://ics.hutton.ac.uk/get-germinate</a> | The James Hutton Institute,<br><a href="http://hutton.ac.uk">http://hutton.ac.uk</a>  |
| Description: breeding database and analysis tools   |   |   |
| GOBii   | <a href="http://gobiiproject.org">http://gobiiproject.org</a>                               | Cornell University,<br><a href="https://cornell.edu">https://cornell.edu</a><br><br>BTI,<br><a href="https://btiscience.org">https://btiscience.org</a> |
| Description: large scale and efficient genotyping storage system including data analysis workflows  |   |   |
| T3  | <a href="https://triticeaetoolbox.org">https://triticeaetoolbox.org</a>                     | USDA,<br><a href="https://usda.gov">https://usda.gov</a>  |
| Description: comprehensive breeding management system designed for wheat  |   |   |

| Database name   | URLs  | Organization, Reference   |
|---|---|---|
| Musa Germplasm Information System (MGIS)  | <a href="https://www.crop-diversity.org/mgis">https://www.crop-diversity.org/mgis</a>                   | Bioversity International, <a href="https://bioversityinternational.org">https://bioversityinternational.org</a> , (Ruas <i>et al.</i> , 2017) |
| Description: information system on banana germplasm   |   |   |
| Gigwa   | <a href="http://gigwa.southgreen.fr">http://gigwa.southgreen.fr</a>                                     | CIRAD, IRD (South Green)  |
| Description: Gigwa (Sempéré <i>et al.</i> , 2016) is a web-application that aims at storing and exposing genotypic datasets and provides a web interface for filtering them in real time. It is able to interoperate with genome browsers and export results into several formats.  |   |   |
| EU-SOL Database   | <a href="https://www.eu-sol.wur.nl">https://www.eu-sol.wur.nl</a>                                       | Wageningen University & Research, <a href="https://wur.nl">https://wur.nl</a>   |
| Description: this site contains information about a collection composed of ~7000 domesticated ( <i>S. lycopersicum</i> ) lines, along with representative wild species, collected with the scope of the european project EU-SOL. This germplasm was generously provided by different international genebanks and by donations from private collections. This Integrated Project is supported by the European Commission through the 6th framework program. Contract number: FOOD-CT-2006-016214 |   |   |
| GnplS   | <a href="https://urgi.versailles.inra.fr/gnplS">https://urgi.versailles.inra.fr/gnplS</a>               | INRA, <a href="https://www.inra.fr">https://www.inra.fr</a>   |
| Description: French national archive for plant phenotyping data. It provides any type of PGR and Phenotyping data. Used for instance by Perphecim for climate change adaptation studies and as a data repository in the Elixir federation which is under construction. It contains almost a thousand Phenotyping trials over thousands of woody and annual plant varieties.   |   |   |
| KDDart  | <a href="https://kddart.diversityarrays.com/brapi/v1/">https://kddart.diversityarrays.com/brapi/v1/</a> | DART, <a href="http://www.kddart.org">http://www.kddart.org</a>   |
| Description: genotype and phenotype database, linked to genotyping service  |   |   |
| Crop Ontology   | <a href="http://www.croponontology.org/">http://www.croponontology.org/</a>                             | Bioversity, <a href="https://bioversityinternational.org">https://bioversityinternational.org</a>   |
| Description: database of available trait ontologies for diverse crops in the CGIAR system   |   |   |
| PIPPA   | <a href="https://pipppa.psb.ugent.be">https://pipppa.psb.ugent.be</a>                                   | VIB <a href="https://www.psb.ugent.be/">https://www.psb.ugent.be/</a>   |
| Description: PSB Interface for Plant Phenotype Analysis   |   |   |
| PHIS  | <a href="http://www.phis.inra.fr">http://www.phis.inra.fr</a>   | INRA, <a href="https://www.inra.fr">https://www.inra.fr</a>   |

| Database name  | URLs  | Organization, Reference                         |
|--|---|---|
| Description: ontology-driven Information System designed for Plant Phenomics. PHIS is designed to store, organize and manage highly heterogeneous and multi-spatial and temporal data from multiple sources (field, greenhouse).   |   |   |
| GBIS/I   | https://fair-ipk.ipk-gatersleben.de/public/breedingapi.html | IPK-Gatersleben, https://www.ipk-gatersleben.de |
| Description: among other, FAIR-IPK offers access to IPK genbank information system GBIS. This comprises passport data (information on the identity, history, geographical origin and botanical classification of the material) of the 150, 780 accessions in Gatersleben (as of 30 June 2016), including the Satellite Collections North in Gross Lüsewitz (potatoes) and Malchow/Poel (oil and fodder crops). |   |   |
| TERRA REF  | https://terraref.ncsa.illinois.edu/bety                     | https://terraref.org                            |
| Description: an open access reference database for high throughput phenomics. Crops include sorghum and wheat.   |   |   |

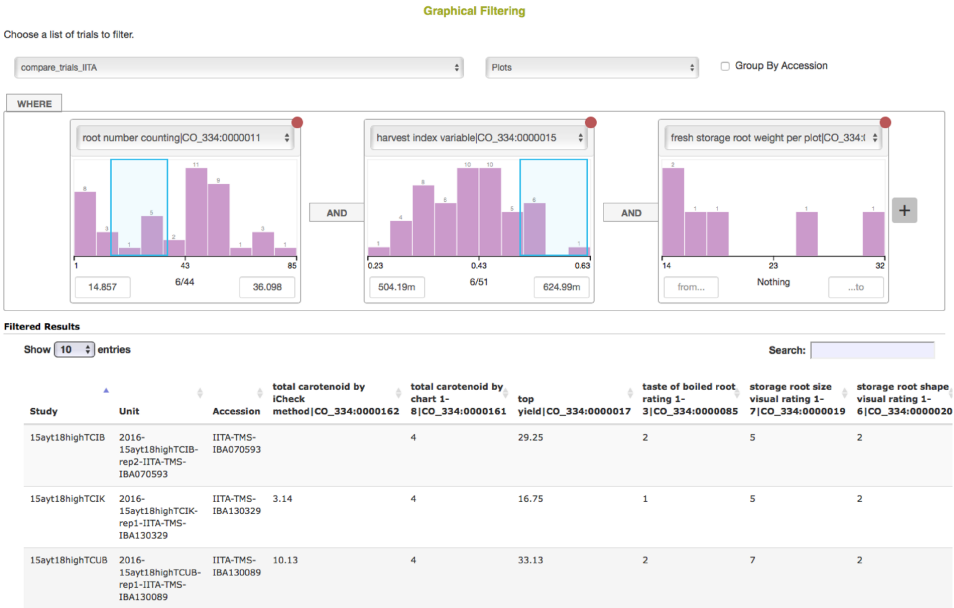
**Table 3.2:** *Server implementations*

### Client implementations

BrAPI client code libraries have been created in several languages, such as Java (<https://github.com/imilne/jhi-brapi>), the BrAPI R package (<https://github.com/CIP-RIU/brapi>), Brapi Drupal for PHP, and brapi.js for Javascript (<https://github.com/solgenomics/BrAPI-js>). A non-exhaustive list of current client applications is given in **Table 3.3**. It is possible for service providers to use BrAPI for the implementation of native website features. Some of these features have been implemented as reusable BrAPI compliant widgets, which we call BrAPI Apps or ‘BrAPPs’ for short. The available BrAPPs are listed on the BrAPI website (<https://brapi.org/brapps.php>). **Figure 3.2** shows a screenshot of an example BrAPP which performs graphical filtering of phenotypic values.

### Test suites and fixtures

Comprehensive testing is very important for any software project. Testing tools are available for both BrAPI server implementations and BrAPI enabled clients. For testing BrAPI enabled clients, a BrAPI test server is available at the brapi.org site (<https://test-server.brapi.org/brapi/v1>). The BrAPI Test Server has a complete implementation of the BrAPI specification and returns a consistent sample set of data. This allows developers of clients to build tests which are appropriate for their tool, while calling a live BrAPI server implementation. The sample data reported by the test server are completely fabricated, and can be updated at any time upon request.



**Figure 3.2:** A screenshot of an example web application that retrieves information through BrAPI. Such applications are often referred to as 'BrAPPs'. This application, called 'Graphical Filtering', allows to filter accessions by phenotypic data, by interactively selecting ranges of trait values for different traits in the dataset. Data from Cassavabase (<https://cassavabase.org/>) are shown, but BrAPPs seamlessly integrate with any BrAPI-enabled database.

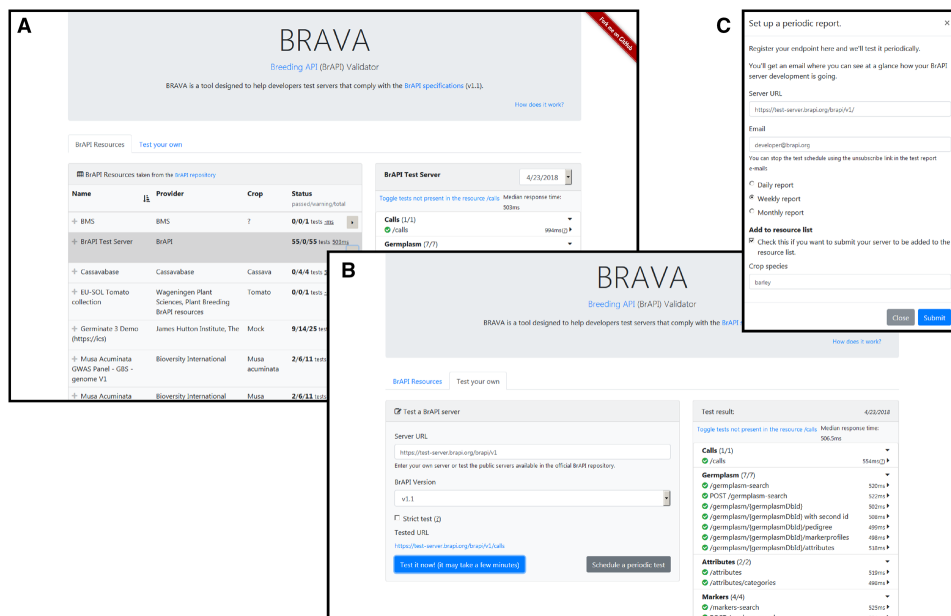
BrAPI validator (BRAVA) test tool

For testing server implementations, the BRAVA test client is available for testing compliance with the BrAPI specification (<http://webapps.ipk-gatersleben.de/brapivalidator/>). Available as a web frontend, BRAVA enables developers to check the compliance of their BrAPI endpoints against the specification and the referential integrity of input and output parameters of dependent endpoints. The frontend, as shown in **Figure 3.3**, enables testing of BrAPI server implementations. A user can also schedule tests and generate periodic reports of the overall status and details of BrAPI endpoint compliance. The compliance tests and results are grouped by and aggregated per REST resource. Using the BrAPI meta-endpoint '/calls', BRAVA is able to detect the available endpoints on the server and will only test those endpoints.

A given endpoint might be tested multiple times with different inputs or HTTP methods. Each test checks the HTTP status code, content type, validity of response body, and response data types. Each test will also compare the response to the expected JSON schema which defines the structure of a JSON object and acceptable types. Some tests check the compatibility of response data to a corresponding parameter. For example, a test will call '/germplasm-search' and will use the first 'germplasmDbld' from the response to make the call '/germplasm/germplasmDbld'. Some tests will

| Program name   | URL   | Institution(s)   |
|--|---|--|
| Flapjack   | <a href="https://ics.hutton.ac.uk/flapjack">https://ics.hutton.ac.uk/flapjack</a>                           | The James Hutton Institute,<br><a href="https://hutton.ac.uk">https://hutton.ac.uk</a> |
| Highly Interactive Data Analysis Platform (HIDAP)    | <a href="https://apps.cipotato.org/hidap_sbase/">https://apps.cipotato.org/hidap_sbase/</a>                 | International Potato Center (CIP)  |
| brapi R package: Implementation of Breeding API in R | <a href="https://github.com/CIP-RIU/brapi">https://github.com/CIP-RIU/brapi</a>                             | International Potato Center (CIP),   |
| Wageningen University & Research, Patranca           |   |  |
| brapixR package                                      | <a href="https://github.com/c5sire/brapix">https://github.com/c5sire/brapix</a>                             | Patranca   |
| brapiui R package                                    | <a href="https://github.com/c5sire/brapiui">https://github.com/c5sire/brapiui</a>                           | Patranca   |
| Pedigree Viewer                                      | <a href="https://github.com/solgenomics">https://github.com/solgenomics</a>                                 | BTI  |
| Graphical Phenotype Filtering                        | <a href="https://github.com/solgenomics">https://github.com/solgenomics</a>                                 | BTI  |
| Trial Comparison                                     | <a href="https://github.com/solgenomics">https://github.com/solgenomics</a>                                 | BTI  |
| Comparative Map Viewer                               | <a href="http://maps.solgenomics.net/">http://maps.solgenomics.net/</a>                                     | BTI  |
| ISMU   | <a href="https://github.com/icrisatSbdm/ismu">https://github.com/icrisatSbdm/ismu</a>                       | ICRISAT  |
| Gigwa  | <a href="http://gigwa.southgreen.fr">http://gigwa.southgreen.fr</a>   | CIRAD, IRD (South Green)   |
| Beegmac  | <a href="http://webtools.southgreen.fr/BrAPI/Beegmac/">http://webtools.southgreen.fr/BrAPI/Beegmac/</a>     | CIRAD (South Green)  |
| GnplS  | <a href="https://urgi.versailles.inra.fr/gnpis">https://urgi.versailles.inra.fr/gnpis</a>                   | INRA   |
| Variable Ontology Widget                             | <a href="https://github.com/gnpis/trait-ontology-widget">https://github.com/gnpis/trait-ontology-widget</a> | INRA   |
| Drupal BrAPI Implementation                          | <a href="https://www.drupal.org/project/brapi">https://www.drupal.org/project/brapi</a>                     | Bioversity   |

**Table 3.3:** Client implementations



**Figure 3.3: BRAVA portal.** (A) List of publicly available endpoints and their compliance status according to BRAVA. An expanded report panel shows the individual test results for the selected resource. (B) 'Test your own' panel where the user can test a custom URL or (C) subscribe to get periodic reports.

compare a response value to a previously stored value. For example, an entity accessed by calling '/germplasm/1' must have a 'germplasmDbld' of '1'.

When a test run is complete, the test suite result is sent to the web client and a report is generated. The report can be inspected in the client and has a tree-like structure to analyse results for individual calls. The scheduled and public resource test reports are stored for future assessment.

## Discussion and Outlook

We have defined a first version of a plant breeding API that defines the key calls needed to exchange information about germplasm, phenotypes, experiments, studies, geographic locations, samples, and genetic markers. This opens the door to a rich set of possibilities for building client applications that can work with any BrAPI-compliant data provider.

Since 2015, a diverse group of data providers and client application programmers have been building BrAPI into their software. Client applications can rely on the standard interface to enable integration with any BrAPI data source. Building software using standard interfaces is an efficient and sustainable coding practice which enables the reuse of software components. As the public plant breeding software community is relatively small, this will be essential for creating a feature-rich breeding software



ecosystem. A good example of the efficient reuse of components can be seen in the community developed BrAPPs, which are tools that make extensive use of BrAPI and can be widely shared and deployed on different BrAPI enabled systems. This framework is useful to commercial plant breeding software development efforts and we welcome more engagement with that community.

We are continuing our efforts and have initiated work on improved versions of the API. We recognize that the types of data relevant to plant breeding are expanding, and BrAPI will continue to evolve in response.

One aspect of the API that we would like to enhance is the ability to handle linked data (Xin *et al.*, 2018). For example, linking between datasets can rely on standard variables, vocabularies or ontologies, such as the Crop Ontology for Agricultural Data (R. Shrestha *et al.*, 2012). To fully enable this, current research and developments are based on adding semantic capabilities to BrAPI, especially through the JSON-LD standard, and some support will likely be included in the next major version of BrAPI. It is also important to improve the clarity and understandability of the BrAPI data for both human and machine. Future development will include documentation of the mapping between BrAPI and other common data specifications, such as MIAPPE and MCPD. This will provide a human friendly documentation of BrAPI formats and concepts. Furthermore, it will also provide reference concepts and schemas necessary to integrate BrAPI with other initiatives such as bioschemas.org.

Beyond breeding applications, BrAPI has also found a niche in gene bank applications, such as MGIS (Ruas *et al.*, 2017), through compatibility with the Multi-Crop Passport Data standard (MCPD). Although the initial intent was to enable interoperability between breeding management resources, BrAPI can also be used with other types of databases, such as plant genetic resources databases [i.e. MGIS (Ruas *et al.*, 2017)] and plant genome databases (i.e. SGN, MaizeGDB, etc.). BrAPI offers a way to link genetic resources distributed by gene banks with materials used in breeding programs. Improved integration between gene banks and plant breeding management databases, and genomic databases has the potential to greatly enhance the management and utilization of plant germplasm collections (Ruas *et al.*, 2017; Spindel and S. R. McCouch, 2016). Efficient and smart use of genetic diversity is a key for continued progress in plant breeding efforts to address the challenges of increased productivity and adaptation (Halewood *et al.*, 2018).

As the needs and technologies of our community continue to evolve, we expect BrAPI to grow to meet those needs.

## Getting involved

We invite the reader to join our community and contribute to the future of BrAPI. To start, please visit <https://brapi.org/> to learn more, contact the BrAPI coordinator at [brapicoordinatorselfby@gmail.com](mailto:brapicoordinatorselfby@gmail.com) to join the mailing list, Slack channel, and community forum.

## Acknowledgements

The BrAPI consortium gratefully acknowledges The Bill and Melinda Gates Foundation for providing funds and support for an organizational meeting in Seattle and three hackathons. We would also like to thank the following organizations for hosting BrAPI Hackathons: GOBii, The Boyce Thompson Institute, INRA, CGIAR Bioversity International, and CGIAR Research Program on Roots, Tubers, and Bananas. We acknowledge Elixir (European Infrastructure for bio-informatics, supported by the ELIXIR-EXCELERATE project funded by the European Commission within the Research Infrastructures program of Horizon 2020, grant agreement numbers 676559 ) and Phenome EMPHASIS.fr (French Phenotyping Infrastructure funded by the Infrastructure Biologie Santé 'Phenome-FPPN' supported by the French National Research Agency, ANR-11-INBS-0012) for the adoption and support of the BrAPI. We also acknowledge the Excellence in Breeding Platform (EiB) for their ongoing support of the BrAPI effort.

## Funding

This work was supported by the Bill and Melinda Gates Foundation, Bioversity International, German BMBF (FKZ 031A536A and 031B0190A), the Boyce Thompson Institute for Plant Research, the CGIAR Research Program on Roots, Tubers and Bananas, Cornell University, and the Excellence in Breeding Platform.





## Chapter 4

# Using the MIAPPE standard to improve reusability of plant phenotyping data: Lessons learned from reusing multi-location potato field trial data

Evangelia A. Papoutsoglou<sup>1</sup>, Ioannis N. Athanasiadis<sup>2</sup>, Richard G.F. Visser<sup>1</sup> and Richard Finkers<sup>1</sup>

<sup>1</sup>Plant Breeding, Wageningen University & Research, Wageningen, the Netherlands

<sup>2</sup>Geo-information Science and Remote Sensing, Wageningen University & Research, Wageningen, the Netherlands



*Submission in preparation*

## Abstract

Plant phenotyping data poses a challenge for reuse, as the experiments that produce it follow a variety of experimental parameters and settings and are often insufficiently and heterogeneously documented. At the same time, science needs to tackle reusability challenges to rise up to the societal needs (nutrition, crop adaptation and stability), which can be done more efficiently by means of meta-analyses and data integration. Although the plant phenotyping community has recently made progress toward reusable resources with the MIAPPE metadata standard and the Breeding API, there are currently no examples demonstrating the establishment or use of such resources for a scientific process from beginning to end, or evaluating the significance of such processes. In this work, we take an existing example of non-FAIR data reuse and establish data and metadata, and infrastructure to make it available in a FAIR way. We assume the role of a scientist discovering a phenotypic dataset on a FAIR data point, verifying the existence of related datasets with environmental data, acquiring both and integrating them. We discuss the challenges and the potential for reusability and reproducibility of FAIRifying existing datasets, that were encountered in this process.

## Introduction

Plant phenotyping is an important process underlying breeding, where plant varieties with improved attributes are developed as a necessity to meet the needs of our growing population (FAO *et al.*, 2018). Phenotyping produces data that is as intricate as the process itself, reflecting the heterogeneity not only in experimental goals, designs and settings and agronomic management practices, but also in human choices and data collection and documentation procedures. Further complexity is observed because of the many types of data it can involve, assembled by humans and/or machines.

The overall complexity of phenotypic data is a hindrance to data reuse because of the general heterogeneity and lack of coordination and standardization in data management practices across a myriad of existing data holder information systems, without a central global repository. Like most domains of science, though perhaps to an even greater extent because of the many different species involved as well as new high-throughput technologies, plant phenotyping has undergone an explosion in the size and number of available datasets. However useful each of them may be to the original producer, their potential to deepen our understanding of plant biology remains unmaterialized because meta-analyses across independently generated datasets remain epistemologically (ambiguities, missing documentation) and logistically (undiscoverable data, different data types) difficult (Coppens *et al.*, 2017; Pieruschka and Schurr, 2019).

To improve the global data landscape, managing the challenges of heterogeneity and attempting to bridge distributed resources, the FAIR (Findable, Accessible, Interoperable, Reusable) data principles have been proposed (Wilkinson *et al.*, 2016). Some of their requirements are common across scientific domains, whereas others are based on reaching a community-wide consensus. In terms of FAIRness, the plant phenotyping community has progressed with MIAPPE, a metadata standard aiming to improve the reusability of plant phenotyping data (Papoutsoglou *et al.*, 2020a). MIAPPE is a Minimum Information standard, ensuring that no “essential” descriptors are overlooked and that metadata is readable to humans and computer tools alike. These descriptors aid three aspects of FAIR: data findability, as attributes of experiments can be indexed commonly by a service; interoperability, as there is a common vocabulary for the metadata; and reusability, as the minimum information necessary for interpretation is supplied. Accessibility needs to be tackled through concrete implementations of the standard such as, for example, the Breeding API (Selby *et al.*, 2019).

Without elements of FAIRness and supporting standards like MIAPPE, data reuse can be challenging. An example of successful reuse can be found in Hurtado-Lopez’s work (Hurtado-Lopez, 2012). The meta-analyses conducted relied on five experiments that were conducted in four different locations across different latitudes, as the environmental variation across those locations is known to affect developmental processes in potato. With respect to the genotypes in those studies there were significant overlaps, as the CxE population (a diploid backcross mapping population extensively studied at Wageningen University & Research) (Jacobs *et al.*, 1995) was chosen for all experiments (including (C. Celis-Gamboa *et al.*, 2003; Hurtado-Lopez, 2012; Zaban *et al.*, 2006)). Although there were differences in the specific genotype sets involved in each of them, there was a large overlap between the experiments. These genotypes were evaluated with respect to traits of morphological (e.g. tuber size), developmental (e.g. flowering), and agronomic

(e.g. yield) nature. On top of the phenotypic data, this work integrated molecular and environmental data for a multi-environment QTL analysis, laying a basis for in-depth meta-analyses reusing primary data from a variety of domains, including genetic, phenotypic, molecular and environmental. Among the insights in Hurtado-Lopez's work, which range from QTL associations and trait correlations, to better comprehension of the effect of environmental conditions on potato, one concerns the effect of temperature and photoperiod on agronomic traits (Chapter 5).

It is important to note that the experiments were conducted by different, uncoordinated parties in different time periods (over 11 years), though there was a direct line of communication with them to establish sufficient understanding for data reuse. In some cases, this can be the critical difference warranting that datasets are indeed compatible, as the omission of experimental design or other details in published materials and protocols may be prohibitive. Still, a key conclusion made clear in P. Hurtado-Lopez's Discussion chapter is that more needs to be done toward improving data practices. In particular, she identifies three elements that should be part of proper documentation for multi-environment studies, if reuse is to be successful: content, origin/source and structure. All of these elements are part of the MIAPPE standard (Papoutsoglou *et al.*, 2020a). It is also key to remember that, although Hurtado-Lopez's work is an example of successful reuse, it reflects work that was logistically complicated for entirely unscientific reasons.

Hurtado-Lopez's work was completed before the establishment of the FAIR data principles, though standardization had already taken root in some scientific domains (e.g. MIAME for microarray data). Therefore, the data she received was often disorganized, lacking important details and sometimes messy, resulting in communications that cost time to incrementally resolve ambiguities. In some cases, data could not be used as there was insufficient information. When all the information was there, the harmonization of the variety of formats and data file structures was another point to be tackled. In spite of that, Hurtado-Lopez was able to successfully reuse a significant part of the different datasets, integrating them and extracting new, biologically significant conclusions.

Sufficiently organized documentation alone would have resolved issues pertaining to implicit details about the experiments, and would have reduced the time spent tracking down bits of information in largely unstructured chains of literature. Moreover, the ability to tackle data integration and manipulation uniformly for all experiments would have been a practical, time-saving asset. These benefits would have had an impact on the final outcome of that work, if more data had been usable. Nevertheless, eliminating (or at least relieving) some of the hindrances to this particular instance of data reuse would have saved resources, and enabled the researcher to focus on more biological, and fewer data handling challenges.

We envision a scenario where different data handling stages (acquisition, integration, analysis/reuse) can be presented as fully streamlined and reproducible. To that end, we use a case study approach and focus on showcasing an example of data integration from different datasets (5 field trial datasets and as many environmental ones) from different sources, and on investigating the challenges and benefits of the approach. We present and evaluate a workflow illustrating an alternative process that could have been followed had the initial data been formatted according to the FAIR principles. The process starts from the phenotyping datasets and moves on to ones holding data



about the environmental attributes desired. The data is used to generate descriptive visualizations, which can be invaluable for exploratory data analysis.

## Results

This work relies on the datasets that Hurtado-Lopez used in her doctoral thesis. As shown below, it comprises 5 phenotypic experiments and the respective photoperiod and temperature conditions at those locations.

### Data

An overview of the datasets can be found in **Table 4.1**.

|                    | Content description   | Experiment ID | Types of source files   |
|--------------------|---|---------------|---|
| Phenotypic dataset | Data from the 1999 field trial in the Netherlands   | 1999NL        | Excel files with experimental measurements                                  |
|                    | Data from the 2003 field trial in Venezuela   | 2003VE        | Excel files with experimental measurements                                  |
|                    | Data from the 2004 field trial in Finland   | 2004Fin       | Excel files with experimental measurements, genotype name translation files |
|                    | Data from the 2005 field trial in Finland   | 2005Fin       | Excel files with experimental measurements, genotype name translation files |
|                    | Data from the 2005 field trial in Ethiopia  | 2010ET        | Excel files with experimental measurements                                  |
| Weather dataset    | Photoperiod data (per day) for each location, covering at minimum the time period in question       | N/A           | Website ( <a href="http://www.timeanddate.com">www.timeanddate.com</a> )    |
|                    | Temperature data (daily average) for each location, covering at minimum the time period in question | N/A           | Excel files with measurements   |

**Table 4.1:** *Summary of the two datasets with their constituents and other attributes.*

### Phenotypic data

No phenotypic data was generated for the present work. The phenotyping experiments that were considered for this work were conducted in: the Netherlands (1999) (C. Celis-Gamboa *et al.*, 2003), Venezuela (2003), Finland (2004 and 2005) (Zaban *et*

*al.*, 2006), and Ethiopia (2010) (Hurtado-Lopez, 2012). All experimental data was retrieved from P. Hurtado-Lopez's PhD thesis supervisor after her departure from the university. Hurtado-Lopez was not the original data collector with the exception of the one experiment in Ethiopia. The rest of the data was relayed to Hurtado-Lopez by its original creators, and was subsequently organized by her. It is unknown which, if any, steps were taken to disambiguate and harmonize the contents of these files.

To us, the data was available predominantly in the form of spreadsheets. The measurements for each experiment were recorded by different people, which is evident in the internal structure of the files.

## Weather data

P. Hurtado-Lopez used weather data to conduct multi-environment and multi-trait analyses in her work (Hurtado-Lopez, 2012). Temperature and weather data was available through her files. However, in some cases, retrieval of finer-resolution data points for the sunlight hours over the span of an experiment was done online ([www.timeanddate.com](http://www.timeanddate.com)), for reasons of homogeneity, reliability and accuracy of this source.

## Process

We retrieved the field trial datasets, each involving a different file/folder structure comprising, among others, spreadsheet files. We selected a trait (tuber weight) and placed the relevant data in text (csv) files. Metadata could be located in publications and local text documents, slidesets, and P. Hurtado-Lopez's thesis text. The relevant metadata were assembled according to MIAPPE 1.1, and everything was transformed to RDF using PPEO (Papoutsoglou *et al.*, 2020a). Weather data was also retrieved and similarly transformed using the AEMET weather ontology (Atemezing *et al.*, 2013). To help aspiring FAIR data owners expose their data and their attributes, and users to find and re-use those datasets, development of FAIR Data Points (FDPs) has been proposed (Kuzniar *et al.*, 2020). They present a hierarchy of levels (FDP, catalog, dataset, distribution) to help organize and present their contents. Humans as well as machine agents can navigate FDPs in either direction in the hierarchy, progressively harvesting the (meta)data therein as necessary. We have constructed and used a simple FDP for this case, which is available on GitHub (FAIR-CxE contributors, 2020).

An observation was made during the configuration of the FDP metadata, especially for the dataset level. The FDP specification recommends the use of specific attributes to describe the dataset (Kuzniar *et al.*, 2020), and the information given mainly describes the resource. However, no concrete information about the dataset content can be given this way. The existing recommendations cannot accommodate our use case, since we need to know essential attributes about a dataset to reuse it, without having to look inside it (which is inefficient). To support meaningful indexing and searchability on FAIR data portals, the contents of datasets should be somehow described on the FDP. Our suggested solution is to extend the metadata descriptors that the FDP provides on the Dataset level. We did this by embedding the MIAPPE metadata into the Dataset level. The principle of the function of the FDP remains the same, but with this additional layer of metadata we can decide whether we are interested in a dataset without the additional step of accessing it. A schematic of this issue and our solution can be seen on

**Figure 4.1.** The detailed steps of the FDP exploration can be seen in **Supplementary Notes S4.1**.

With the FDP in place, we can navigate to its topmost level and start exploring the FAIRified data. The role we assume is of a researcher who knows of the existence of the CxE datasets (though not their details), and wants to run a multi-environment, multi-trait analysis. We will not conduct a proper biological analysis in this work, as the methodology for plant phenotyping analyses and their intricacies are outside the scope of this work. Instead, we want to showcase what FAIR data discovery, acquisition, and integration could look like, focusing on six milestones (**Figure 4.2**), each reflecting a step that a researcher would take in exploring the data.

For this scenario, the FDP lists a version ("distribution") of the phenotypic dataset (including data and metadata) that is available on a triple store and directly queriable with the SPARQL language.

This process is described in greater detail in the Jupyter notebook available in our repository, and everything (code, original and derived data files, FDP, triple store, scripts, queries) is also available there (FAIR-CxE contributors, 2020).

## Finding relevant phenotypic datasets on the FDP

To find relevant phenotypic datasets, a user has to navigate to the FDP of their institute. The top-level of the FDP exposes metadata, among others, about the host institute, and the data catalogs it contains (**Figure 4.3**). In this case we have chosen to separate the data by types: phenotypic, genotypic and genomic, and since we are looking for the CxE phenotypic datasets, we navigate to the Phenotypic catalog. We examine each dataset, and we see that, out of the three present, one is relevant. To be able to explore the dataset by querying it, we select the SPARQL endpoint distribution, and we get an URL for it. For this, they only need to know the address of the FDP.

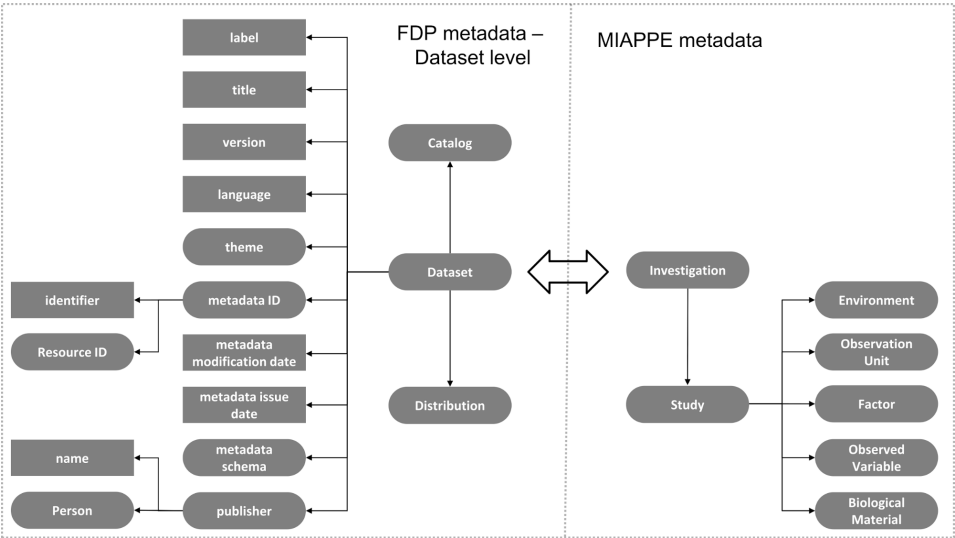
## Exploring the Investigation and Study metadata

To begin with, we can write a query to retrieve the investigation in that SPARQL endpoint, and the properties (from the PPEO ontology) that connect it to further information. Assuming one has limited knowledge of a dataset, this step of exploration may present us with an arrangement as seen on **Figure 4.4**.

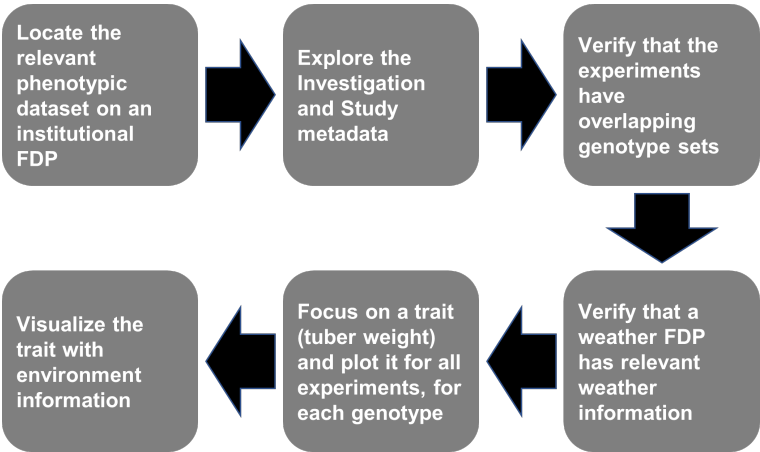
With the same approach, we can discover the attributes that connect the studies to their details. We can also produce more structured tables, holding each study attribute in a specified column (**Figure 4.5**).

## Verifying that the experiments have overlapping genotypes

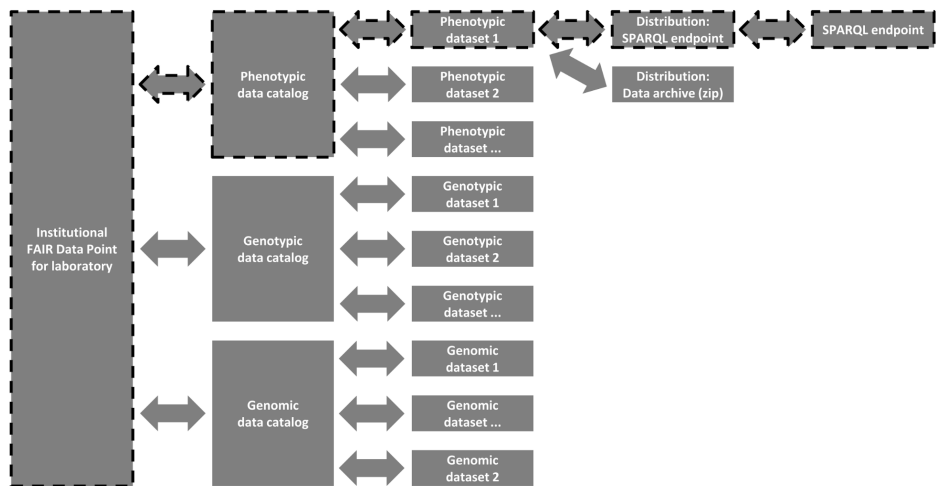
A necessary prerequisite in the scenario we are developing (multi-environment analysis) is the existence of an overlap in the genotypes in the studies. We can write a query to verify that they indeed exist, and we can have a list. In total, there are 101 common genotypes across all experiments.



**Figure 4.1:** An illustration of the FDP Dataset specification (left), and our extension of it with MIAPPE metadata (right). Without concrete metadata about the dataset contents, the FDP cannot support meaningful, content-oriented indexing and searchability. Note only the connection that is made in the picture is between the Investigation and the Dataset. Other alignments between the two sides (e.g. Distribution and Study) have no meaning.



**Figure 4.2:** A diagram showing the steps that a researcher in this situation would follow to locate, acquire, examine, and reuse data in a FAIR way.



**Figure 4.3:** *The structure of our FDP. One can start at the left and move toward the right, along the path indicated by the arrows, following the shapes with the dashed outlines. It is also possible to move backwards, as each level links to both its parent and its child.*

### Finding weather data for the location and time period of interest

Through the FDP, we previously discovered the address of a SPARQL endpoint holding the data for the phenotyping experiments we were interested in. Though the process is not shown here, we assume that we have discovered a similar SPARQL distribution for weather data. We can jointly query this endpoint and our phenotyping data endpoint, to make sure that the weather data we are interested in is indeed there.

The weather data is structured around weather stations, each located at certain coordinates, producing measurements at certain intervals for certain weather variables. More information about this structure is given in the Methods section. For each experimental location, we have constructed a weather station. In real life scenarios, these points would probably not overlap with the experimental fields. Nonetheless, we assume this to be a perfect case. The MIAPPE metadata gives us the location coordinates for the studies, from which we can calculate the distance to each weather station. Sorting in ascending order, we see the lowest values indicating the matches (Figure 4.6).

### Focusing on a trait and plotting it for all experiments, for each genotype

We now know that we have access to phenotypic data and weather information for our experiments. We will focus on the correlation between temperature and photoperiod, as P. Hurtado-Lopez's thesis does (among many other things), to demonstrate data integration. One of the choices to be made revolves around the trait to examine. A few traits are available for each investigation, and they can be examined more closely

|    | investigationProperty   | target  | type  |
|----|---|---|---|
| 0  | <a href="http://purl.org/pppeo/PPEO.owl#hasAssociatedPublication">http://purl.org/pppeo/PPEO.owl#hasAssociatedPublication</a> | <a href="https://library.wur.nl/WebQuery/wurpubs/fulltext/240586">https://library.wur.nl/WebQuery/wurpubs/fulltext/240586</a>               | NaN   |
| 1  | <a href="http://purl.org/pppeo/PPEO.owl#hasDescription">http://purl.org/pppeo/PPEO.owl#hasDescription</a>                     | 2012 thesis of Paula Ximena Hurtado Lopez   | NaN   |
| 2  | <a href="http://purl.org/pppeo/PPEO.owl#hasDescription">http://purl.org/pppeo/PPEO.owl#hasDescription</a>                     | FAIRified, partial data from the 2012 thesis of Paula Ximena Hurtado Lopez  | NaN   |
| 3  | <a href="http://purl.org/pppeo/PPEO.owl#hasIdentifier">http://purl.org/pppeo/PPEO.owl#hasIdentifier</a>                       | WUR_inv_CE2020  | NaN   |
| 4  | <a href="http://purl.org/pppeo/PPEO.owl#hasLicense">http://purl.org/pppeo/PPEO.owl#hasLicense</a>                             | CC-BY 4.0   | NaN   |
| 5  | <a href="http://purl.org/pppeo/PPEO.owl#hasMIAPPEVersion">http://purl.org/pppeo/PPEO.owl#hasMIAPPEVersion</a>                 | 1.1   | NaN   |
| 6  | <a href="http://purl.org/pppeo/PPEO.owl#hasName">http://purl.org/pppeo/PPEO.owl#hasName</a>                                   | Investigating genotype by environment and QTL by environment interactions for developmental traits in potato                                | NaN   |
| 7  | <a href="http://purl.org/pppeo/PPEO.owl#hasPart">http://purl.org/pppeo/PPEO.owl#hasPart</a>                                   | <a href="http://localhost:3030/phenoi/data#study/WUR_inv_CE2020_1999NL">http://localhost:3030/phenoi/data#study/WUR_inv_CE2020_1999NL</a>   | <a href="http://purl.org/pppeo/PPEO.owl#study">http://purl.org/pppeo/PPEO.owl#study</a> |
| 8  | <a href="http://purl.org/pppeo/PPEO.owl#hasPart">http://purl.org/pppeo/PPEO.owl#hasPart</a>                                   | <a href="http://localhost:3030/phenoi/data#study/WUR_inv_CE2020_2003VE">http://localhost:3030/phenoi/data#study/WUR_inv_CE2020_2003VE</a>   | <a href="http://purl.org/pppeo/PPEO.owl#study">http://purl.org/pppeo/PPEO.owl#study</a> |
| 9  | <a href="http://purl.org/pppeo/PPEO.owl#hasPart">http://purl.org/pppeo/PPEO.owl#hasPart</a>                                   | <a href="http://localhost:3030/phenoi/data#study/WUR_inv_CE2020_2004Fin">http://localhost:3030/phenoi/data#study/WUR_inv_CE2020_2004Fin</a> | <a href="http://purl.org/pppeo/PPEO.owl#study">http://purl.org/pppeo/PPEO.owl#study</a> |
| 10 | <a href="http://purl.org/pppeo/PPEO.owl#hasPart">http://purl.org/pppeo/PPEO.owl#hasPart</a>                                   | <a href="http://localhost:3030/phenoi/data#study/WUR_inv_CE2020_2005Fin">http://localhost:3030/phenoi/data#study/WUR_inv_CE2020_2005Fin</a> | <a href="http://purl.org/pppeo/PPEO.owl#study">http://purl.org/pppeo/PPEO.owl#study</a> |
| 11 | <a href="http://purl.org/pppeo/PPEO.owl#hasPart">http://purl.org/pppeo/PPEO.owl#hasPart</a>                                   | <a href="http://localhost:3030/phenoi/data#study/WUR_inv_CE2020_2010ET">http://localhost:3030/phenoi/data#study/WUR_inv_CE2020_2010ET</a>   | <a href="http://purl.org/pppeo/PPEO.owl#study">http://purl.org/pppeo/PPEO.owl#study</a> |

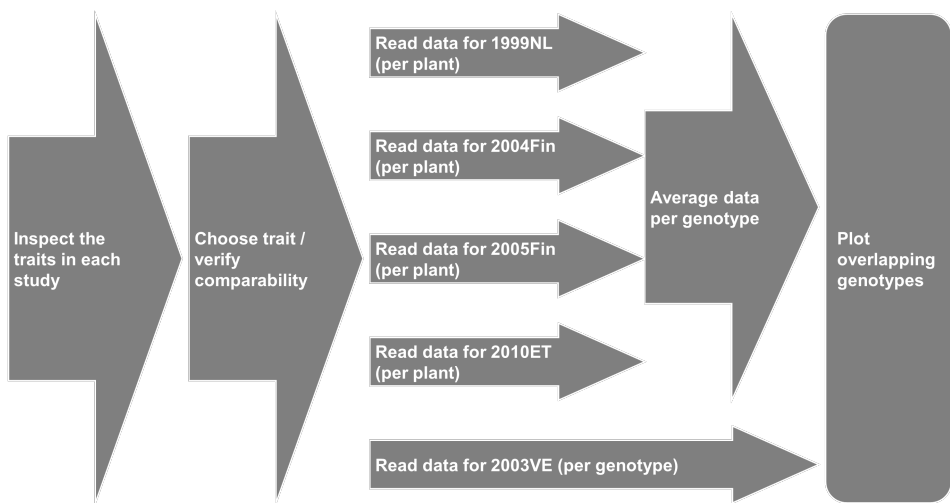
**Figure 4.4:** Retrieving the investigation properties, their values and their class type, when applicable. NaN values in the type column indicate that the target of that MIAPPE attribute is a literal.

|   | studyId | cntName | locName    | addr  | startDate  | endDate    | lat  | long   | alt   |
|---|---------|---------|------------|---|------------|------------|------|--------|-------|
| 0 | 1999NL  | NL      | Wageningen | WUR field                                       | 1999-03-12 | 1999-11-14 | 52.0 | 5.63   | 11m   |
| 1 | 2003VE  | VE      | Merida     | Mucupiche field in La Fresa                     | 2003-06-24 | 2003-11-30 | 8.6  | -71.13 | 2200m |
| 2 | 2004Fin | FI      | Ruukki     | Field in North Ostrobothnia Research Station    | 2004-04-16 | 2004-11-10 | 64.7 | 25.0   | 48m   |
| 3 | 2005Fin | FI      | Ruukki     | Field in North Ostrobothnia Research Station    | 2005-05-03 | 2005-11-14 | 64.7 | 25.0   | 48m   |
| 4 | 2010ET  | ET      | Holeta     | Field in Holeta Agricultural Research Institute | 2010-07-16 | 2010-12-06 | 9.12 | 38.13  | 2400m |

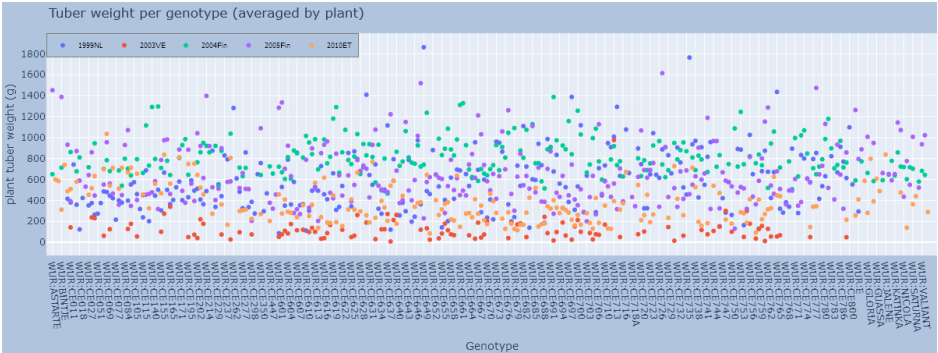
**Figure 4.5:** Fetching specific study values for each study: the study ID, the country abbreviation, the location name and address, the start and end dates of the study, its GPS coordinates and altitude.

|   | studyId | cntName | stu_locName | stu_lat | stu_long | w_name         | w_lat | w_long | dist_sq     |
|---|---------|---------|-------------|---------|----------|----------------|-------|--------|-------------|
| 0 | 2004Fin | FI      | Ruukki      | 64.7    | 25.0     | Ruuki station  | 64.7  | 25     | 0.0         |
| 1 | 2005Fin | FI      | Ruukki      | 64.7    | 25.0     | Ruuki station  | 64.7  | 25     | 0.0         |
| 2 | 1999NL  | NL      | Wageningen  | 52.0    | 5.63     | WUR station    | 52    | 5.63   | 0.0         |
| 3 | 2003VE  | VE      | Merida      | 8.6     | -71.13   | Merida station | 8.6   | -71.13 | 0.0         |
| 4 | 2010ET  | ET      | Holeta      | 9.12    | 38.13    | Holeta station | 9.12  | 38     | 0.016900279 |
| 5 | 2004Fin | FI      | Ruukki      | 64.7    | 25.0     | WUR station    | 52    | 5.63   | 536.4868    |
| 6 | 2005Fin | FI      | Ruukki      | 64.7    | 25.0     | WUR station    | 52    | 5.63   | 536.4868    |
| 7 | 1999NL  | NL      | Wageningen  | 52.0    | 5.63     | Ruuki station  | 64.7  | 25     | 536.4868    |
| 8 | 1999NL  | NL      | Wageningen  | 52.0    | 5.63     | Holeta station | 9.12  | 38     | 2886.5112   |
| 9 | 2010ET  | ET      | Holeta      | 9.12    | 38.13    | WUR station    | 52    | 5.63   | 2886.5112   |

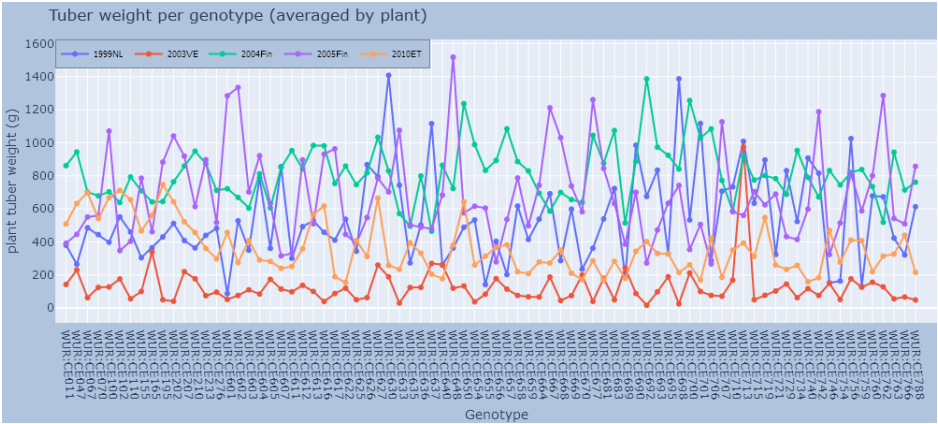
**Figure 4.6:** A query calculating the differences (squared) between the coordinates of each experiment and each weather station. Sorting the differences in ascending order confirms that there is a weather station that is suitable for each experiment.



**Figure 4.7:** Necessary steps to be taken if a trait is to be compared/summarized in a plot across experiments.



**Figure 4.8:** A chart showing the total tuber weight produced, on average, by each genotype in each experiment, for a total of 292 genotypes. Not all genotypes have been used in all experiments. The x-axis on this chart lists the genotype name and the y-axis the average tuber weight per genotype (averaged by plant). The different dot colors show which experiment the data point corresponds to.



**Figure 4.9:** A chart showing the total tuber weight produced, on average, by each genotype in each experiment. Unlike the previous figure, this one includes the genotypes that not only were studied in all 5 experiments (101), but also had a value for our trait of interest (80). The x-axis on this chart lists the genotype name (though not all are labeled on the x-axis), and the y-axis the average tuber weight per plant per genotype. The different dot colors show which experiment the data point corresponds to.



through the Observed variables. We will focus on data for Tuber weight per genotype for each experiment. The process for plotting a trait across all experiments is illustrated in **Figure 4.7**.

The queries required to assemble the data from the SPARQL endpoint are detailed in the notebook available on the GitHub repository for this work (FAIR-CxE contributors, 2020). With the responses, we can see that there are sharp differences in the performance of the genotypes in each experiment, and great variability even for each genotype. **Figure 4.8** includes all genotypes (292), whereas **Figure 4.9** only the ones that were present in all five studies (101). Of those, not all have a value for the trait in question.

### Making plots combining this trait with the weather information

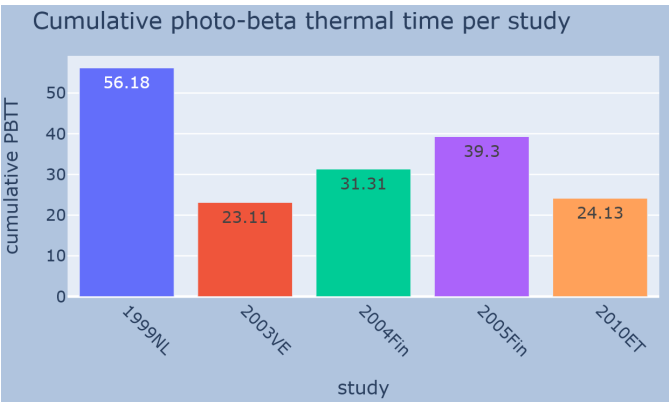
Literature indicates that there is an association between the yield of potato plants and day length. In Hurtado-Lopez's thesis, she also includes a measure that combines the effects of temperature as well as day length, (cumulative) photo-beta thermal time (PBTT) (Hurtado-Lopez, 2012). We have calculated these values for each experiment, as shown on **Figure 4.10**.

We finally want to create a figure where tuber weight is plotted against the photo-beta thermal time. In the figure below, we have combined three different types of data: a) data from the phenotyping experiments (tuber weight values); b) data from the experiment metadata (to get the genotypes associated to the observation unit IDs); and c) weather information, for the average hours of sunlight at each location, for the duration of its experiment. In **Figure 4.11**, a line is drawn for each genotype based on two data points: one for the lowest average tuber weight yielded by the plants of that genotype in an experiment, and one for the highest. Some lines rise and others fall; the difference between the minimum and maximum values can be seen as an indicator of the performance stability for that genotype.

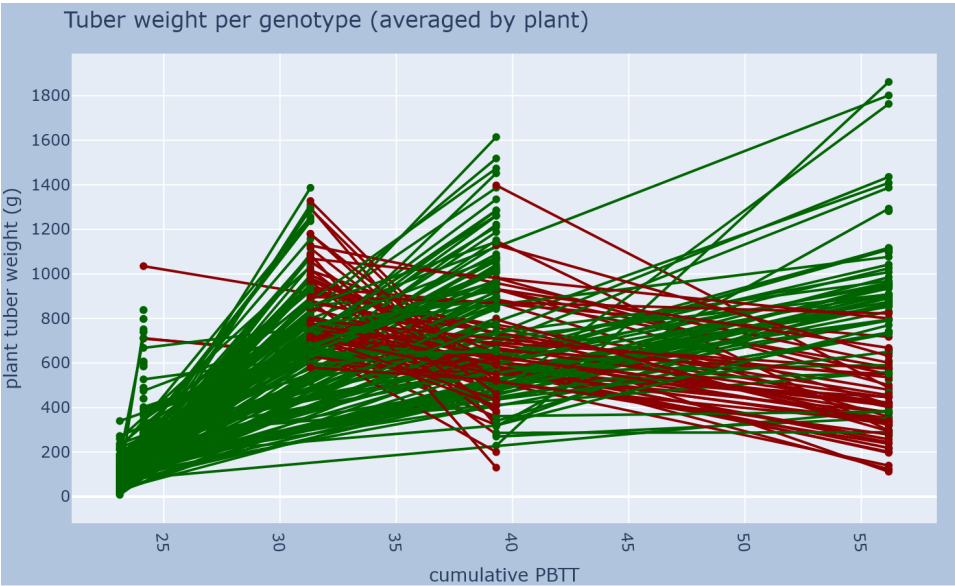
With these results, we have shown that the data and the metadata are queryable and homogeneous. The datasets are discoverable on the FDP. MIAPPE ensures that key aspects are documented, and their components are easier to understand as their information is assembled in an organized way. Furthermore, all of the data can be processed uniformly. This means that it is easier to integrate it not only with other datasets of the same type (phenotyping experiments), but also with different types (e.g. weather). Although no new biological insight is given in these exploratory visualizations, this type of summary can be useful to indicate where particular attention is warranted by a domain expert.

## Discussion

The FAIR data principles encompass a broad collection of data attributes and methods related to management practices to increase data gain. Reusability is a broader challenge that affects science, and some fronts can be tackled in singular ways across domains (e.g. provision of licenses for reuse). In this work, we evaluate the domain-specific requirements for FAIRness and the process to fulfill them in plant phenotyping. We consider a pipeline which starts with data that is already enriched with computer-readable documentation, and present a transparent pipeline that combines heterogeneous data. The pipeline



**Figure 4.10:** The cumulative PBTT calculated for each experiment, based on daily average temperatures and photoperiod.



**Figure 4.11:** This figure shows the best and the worst performance for each genotype, depending on the environment. Each environment/study corresponds to a specific value of cumulative PBTT on the x-axis. The y-axis shows the average tuber weight per plant per genotype. A line that is rising (green) indicates that the genotype performed worst in an environment where the days were, on average short and the temperatures low, and better where days were long and temperatures high, whereas a falling line (red) indicates the opposite.

is composed of different stages for data discovery (F - Findability), acquisition (A - Accessibility / I - Interoperability), analysis and integration (R - Reusability) for exploratory purposes.

Most of the challenges encountered in this work are exclusive to the FAIRification stage and in particular revolve around interpreting the data and metadata. The parties involved in data generation are frequently the only ones to possess various details about experimental settings and details, as they are not deemed important enough to document, or to document in an organized manner. In this use case, the metadata was predominantly present as free text (in Hurtado-Lopez's thesis and referenced papers). The authors clearly knew that; in order to publish their work and make it reproducible, the experiments had to be well-documented, and made great efforts to provide all necessary information. In some cases, this work proves that, in spite of those efforts, the results fell short of the ambition.

Without standards to guide the authors, and possibly because of their continued immersion in their work, this documentation (metadata) is not always comprehensive. This can be observed in the Environment and Events section of the MIAPPE spreadsheet that was filled in this experiment, and how little information is available in this section. For example, planting dates are mostly there, but the presence of only some actions around watering or fertilizer use highlights the information which was not recorded, and which could have been useful to know for future reuse. After all, the original data generator has no way of predicting which attributes may turn out significant for future re-users. Furthermore, whereas traits with complex evaluation procedures or scales are described in great detail, more "humble" ones don't enjoy the same benefit. E.g., for plant height: where from and where to exactly was this trait measured? Data from a wheat (*Triticum aestivum*) trial where plant height was measured "from the ground level to the tip of the spike" (Abebe *et al.*, 2020) may not be directly comparable to height measured "from the first node to the tip of the spike (excluding awns)" (Niu *et al.*, 2020), or may require transformations and other considerations if integration of such datasets is required. When such details are unknown, reuse of data can be impossible or irresponsible.

Conflicts and ambiguities also arise when the actual data is examined alongside the metadata. In some cases, the thesis/articles stated that, alongside a number of CxE genotypes, commercial cultivars were planted and evaluated. The data itself may fail to include those (though in some cases may state that they have been removed), or fail to match the number of reported genotypes for that experiment. In many cases, the spreadsheet files where data was logged have labels in different languages (presumably because of the authors' native languages), use undocumented abbreviations (which renders the data invalid for reuse), or not specific enough (is this value a sum? or an average? is it per plant or per plot? how many plants were averaged?). More confusingly, some abbreviations had different meanings across different files (e.g. "tottub" or presumably "total tubers": in some cases it was the number per plant, in other cases the average for that genotype across plants) - which is crucial when one needs to be sure that the data points they wish to compare, from different experiments, are indeed compatible. Another complication that arose was the use of symbols or categories outside the ones declared for certain traits (e.g. for a 0-7 flowering scale, undefined dots or asterisks).

In some cases, the same information was allegedly given in more files, but the numbers did not match up. This could be attributed to use of the same abbreviations, or similar labels, with different meanings across files. Ultimately, to resolve some of these questions and proceed with this work, it was necessary to consult with coordinators of the original experiments on some details. Finally, even when the data files and the metadata was comprehensive and unambiguous, cleaning the original data files and reshaping them to a common format was a time consuming process, as different traits for each experiment were organized differently, and significant effort had to be put into locating all relevant details in an ocean of free text.

In this work, the process of metadata collection and (meta)data harmonization was undertaken by a doctoral student with an engineering background who had spent three years interacting in the plant phenotyping community, helping shape standards and improve data sharing methods in the domain. It should therefore be noted that the time investment may have been markedly different for a plant biologist. However, for the current requirements, it took on average two weeks per experiment to read the relevant literature (inside and outside the core, Hurtado-Lopez's thesis), examine the pre-organized data files and reliably draw conclusions about the attributes that were collected. Furthermore, only a small part of the data (for the traits used in this work) was harmonized. A much larger volume was ignored because of time constraints, as understanding and sufficiently documenting all aspects would have been prohibitively time consuming. This work would have benefited appreciably from organized, structured documentation, the likes of which have been highly unlikely in plant phenotyping until the establishment of MIAPPE: it would have effectively reduced this time investment to mere hours, as only a brief examination of the attributes and data transformation would have been required. The investment would have been lower for the data collectors, as they are inherently familiar with their work.

Having been through the process of FAIRifying and using FAIR data, an open question has arisen with respect to findability. The FDP specification includes a "theme" that can be listed for a dataset, but that can be as general as "science" or "biology". More details need to be present alongside that, and there is probably some overlap with the minimum information and Minimum Information standards for different domains. The realization is that, depending on the purpose, the definition of what should be "minimum" varies: whereas all MIAPPE details are relevant when it comes to describing a phenotyping experiment, not all of them need to be there to serve findability (e.g. observation unit details). Consideration should be given to that, and in particular with respect to scalability issues. The scenario illustrated here, with a human manually navigating a FDP, will ideally not translate to reality; instead, indexing sites should provide search functions over larger dataset collections, relying on minimum information.

From this work, it is obvious that there is a high knowledge barrier to entry when it comes to the technologies chosen here to implement FAIR. As this is a proof of concept with respect to the possibilities, user friendliness was not a priority. With graphical user interfaces, instead of the html-based FDP used here, the steps of data discovery and acquisition could be facilitated greatly, though such approaches may sacrifice some flexibility that the more tech-savvy users could enjoy. The technical know-how required may dissuade a significant portion of users from utilizing powerful FAIR resources, so it should be one of the first points to be considered for improvement.

The pipeline we present here embraces all elements of the FAIR principles. Findability is achieved with unique identifiers for every (meta)data element, connections between them and rich descriptors. Although the metadata is currently not findable through any community repository, the FDP structure combined with the (partial) metadata provides a clear path to doing so (by indexing data catalog / dataset metadata). For accessibility, we use a combination of established protocols (HTTP), data models (RDF) and formats (TTL), and ensure that metadata are available independently of the data. Reusability is achieved through the explicit use of a license, the MIAPPE metadata standard (which includes provenance, though the FDP does too), and data that follows (admittedly still developing) community data formatting recommendations. Finally, MIAPPE and the use of a community-promoted implementation for it are responsible for (meta)data interoperability, thanks to the explicit data model of MIAPPE.

It is imperative that the science community as a collective start the process of data FAIRification as soon as possible, even if that means making strong compromises on the technical side. A balance needs to be achieved between assembling datasets that are FAIR-ready (i.e. equipped with good documentation and some identifiers) even though the technical provisions (FDPs, registries, curation, versioning and querying specifications) may not be in place. At least, the scientific communities should be exposed to the idea that standards are beneficial and worth some investment. Producing sufficient standards and actually using them to annotate one's own datasets is time-consuming, other benefits may have to be made clear, such as enhanced visibility and citability of one's own datasets. Furthermore, standards need to be promoted more strongly through entities such as journals, to create a basis for future incremental improvements.

## Methods

### Standards and data formats

MIAPPE has been in development by the plant phenotyping community since 2016, and recently received enhancements to its specifications, formats and scope (Ćwiek-Kupczyńska *et al.*, 2016; Krajewski *et al.*, 2015; Papoutsoglou *et al.*, 2020a). It is the state-of-the-art standard for metadata provision in this community, and now answers the explicit need for a domain-relevant community standard required by the FAIR data principles (R1.3).

The MIAPPE implementations that are currently mature enough to fully implement it are the ISA-Tab format (Rocca-Serra *et al.*, 2010), the Breeding API (Selby *et al.*, 2019) and the Plant Phenotype Experiment Ontology (PPEO) (Pommier *et al.*, 2020). A spreadsheet implementation exists, though it is intended to play a secondary role to the others and serve as an easy way to introduce the standard to newcomers. For this work, a combination approach was deemed best: the spreadsheet has been used to assemble the metadata and as a means of intuitive inspection, and the PPEO to support machine readability.

As far as weather data is concerned, there is no clear preference for a specific ontology in the global community. The Linked Open Vocabularies portal (Vandenbussche *et al.*, 2017), which aggregates ontologies from the web so that they may be reused,

presents three results to a “weather” query: the Home Weather ontology, the Smart Home Weather ontology, and the Air Traffic Data ontology. As they pertain to either smart home or air traffic contexts, they were not deemed appropriate. Further search indicated the BIMERR Weather Ontology <sup>(1)</sup> and the SEAS Weather Ontology <sup>(2)</sup> as possibilities, but no examples of actual use were found to support its adoption for this work. Contrary to that, the AEMET weather ontology (Atemezing *et al.*, 2013), of the Spanish Meteorological Office, is currently used to expose its data as linked data <sup>(3)</sup>, and for this reason it was chosen.

While MIAPPE does not specify a data format, it states that “The data files are formatted according to the common practices of the domain and contain references to that Variable ID, the measured values and times plus any information which researchers might deem useful”. We made the choice to use common tabular files with columns for the observation unit ID and measurement date, followed by the variable ID for the traits in question. In some cases, especially when it comes to time series, this results in sparse files. For the weather data, the same data format was followed.

## Technologies

### FAIR Data Point

To approach this work in a way that is consistent with the FAIR principles, it was necessary to produce a FAIR data point (FDP). The metadata was assembled according to the specifications developed by the Dutch Techcentre for Life Sciences (Kuzniar *et al.*, 2020). It includes a tree structure with the metadata for the FDP itself at the root, followed by a Catalog (a collection of datasets), the dataset itself, and distributions of it (in different formats). Each level, on top of the metadata about itself, includes links to the ones below and above it, enabling navigation. The metadata is given in RDF / turtle format (ttl) , and exposed by a python server script.

Part of the given details of a dataset, whether inside or outside a FDP, concern its contents or its description, so that interested parties may be aware of its topic prior to accessing it. In the general FDP specifications, a general way to accomplish this is indicated.

### Resource Description Framework (RDF)

We chose to provide all data as linked RDF data. RDF is a W3C standard for data interchange, and one of the pillars powering the semantic web (Lassila, Swick, *et al.*, 1998). It is used to describe any kind of resource by making statements about it, each composed of three parts: a subject, a predicate and an object. Subjects are referred to by Unique Resource Identifiers (URIs) and statements can be made about them, connecting them to plain literals or other subjects with their own URIs. Semantic schemas specifying, for a given domain, possible ways to lay out and request information about subjects can be provided as ontologies. In turn, other parties can make statements

<sup>(1)</sup><https://bimerr.iot.linkeddata.es/def/weather/>

<sup>(2)</sup><https://ci.mines-stetienne.fr/seas/WeatherOntology>

<sup>(3)</sup>[http://aemet.linkeddata.es/index\\_en.html](http://aemet.linkeddata.es/index_en.html)

about those subjects, eventually assembling a web of linked data. RDF data can be queried using the SPARQL query language.

## FAIRification process

The process started with the FAIRification of the phenotyping datasets. Note that, for this work, no genotypic information is included. For this proof of concept only a few traits were chosen, with the base goal being a showcase of FAIR data discovery, acquisition and integration.

The materials used are available in this GitHub repository: FAIR-CxE contributors, 2020.

## Assembling the MIAPPE metadata

MIAPPE requires information for a number of categories. For each, an effort was made to locate the relevant information in P. Hurtado-Lopez's thesis and related publications referenced therein. The information was compiled on MIAPPE's spreadsheet format, using a different tab for each section.

- **Investigation:** We chose to represent the Investigation as a collection of the 5 experiments contained in P. Hurtado-Lopez's thesis. The details reflect this content, and were chosen freely.
- **Study:** Each of the 5 phenotyping experiments comprises a study. General details, such as the location, start and end date and experimental design information, are listed here.
- **Person:** The coordinators of this work and P. Hurtado-Lopez are listed.
- **Data file:** 5 data files were composed, each holding the data for one study. They are listed here.
- **Biological material:** For this section, we create one biological material for each CxE genotype (including the parents), and other cultivars. In particular, the CxE cross is listed as the material source for each of the genotypes produced from it. It is important to maintain traceability of each genotype through the experiments, as we want to compare their performance in different environments.
- **Environment:** Details supplied in the thesis about the conditions (e.g. average temperature, type of soil) are present here, though this list is far from comprehensive.
- **Experimental factor:** There were no factors.
- **Events:** Planting dates, where known, are listed here, as well as applications of fungicides and known water treatments. Note that a date is necessary for the creation of an event, which prohibits the inclusion of un-timestamped occurrences.

- **Observation unit:** For most experiments, we have information about individual plants, and their organization into plots and blocks. The exception is the experiment in 2003 in Venezuela, where the data was first recorded (by the person who conducted this experiment) per plant, and then averaged for each genotype - which are the only data records communicated. In this case, the observation unit type “genotype” is used. There are some rare data points in other studies where the same unit type is used, as some genotypes were precluded from the records for not being a CxE clone.
- **Sample:** There were no samples to list.
- **Observed variable:** For the most part, the thesis gives a general description about the way that plant traits were evaluated. In some cases, assumptions had to be made as the were not unambiguous.

The resulting spreadsheet can be found on this GitHub repository (FAIR-CxE contributors, 2020).

|    | A                   | B          | C         | D         | E         | F         | G                    | H                    |
|----|---------------------|------------|-----------|-----------|-----------|-----------|----------------------|----------------------|
| 1  | Observation unit ID | Date       | TubN_LT20 | TubW_LT20 | TubN_GE20 | TubW_GE20 | TubN_total_per_plant | TubW_total_per_plant |
| 2  | WUR:CE-ETH10-1-1-1  | 2010-12-01 | 10        | 140       | 10        | 410       | 20                   | 550                  |
| 3  | WUR:CE-ETH10-1-1-2  | 2010-12-01 | 2         | 46        | 3         | 145       | 5                    | 191                  |
| 4  | WUR:CE-ETH10-1-1-3  | 2010-12-01 | 5         | 75        | 9         | 460       | 14                   | 535                  |
| 5  | WUR:CE-ETH10-1-1-4  | 2010-12-01 | 3         | 65        | 4         | 185       | 7                    | 250                  |
| 6  | WUR:CE-ETH10-1-2-1  | 2010-12-01 | 5         | 86        | 6         | 255       | 11                   | 341                  |
| 7  | WUR:CE-ETH10-1-2-2  | 2010-12-01 | 7         | 80        | 16        | 705       | 23                   | 785                  |
| 8  | WUR:CE-ETH10-1-2-3  | 2010-12-01 | 10        | 143       | 9         | 340       | 19                   | 483                  |
| 9  | WUR:CE-ETH10-1-2-4  | 2010-12-01 | 4         | 85        | 4         | 280       | 8                    | 365                  |
| 10 | WUR:CE-ETH10-1-2-1  | 2010-12-01 | 16        | 104       | 9         | 305       | 25                   | 409                  |

**Figure 4.12:** Part of the experimental data file for the 2010 Ethiopia experiment. The first column lists the observation unit ID, which can be cross-referenced with the metadata present in MIAPPE for more information, such as the observation level of that unit (in this case: plant), or its genotype. The date column is followed by columns labeled with the observation unit IDs, which again can be cross-referenced with MIAPPE metadata for a comprehensive explanation.

## Assembling the experimental data

For each experiment, a tab-delimited text file was composed, holding all data for it. This includes data that was only measured once (e.g. total tuber weight for a plant), or as part of a time series. An example of such a file can be seen in **Figure 4.12**.

## Generating RDF

The MIAPPE spreadsheet holding the information about the experiments was processed using a Python script. This script used PPEO to produce RDF from the spreadsheet data, capturing all of its contents. The process can be inspected in a Jupyter notebook. Note that certain formatting assumptions were made (for example, for the location details cells of the spreadsheet) to make this possible. This process is fully reproducible,



however this is not a general-purpose converter as it does not cover the parts of MIAPPE (sections, attributes) that were not used in this work.

PPEO was also used to generate RDF for the actual experimental data. The URIs generated by the script in this process are not resolvable online. An example of the MIAPPE RDF file and of a data RDF file can be seen in **Figure 4.13** and **Figure 4.14** respectively.

### Exposing RDF

The lightweight Jena Fuseki triple store was chosen to serve as a SPARQL endpoint. In it, two datasets were created: one for the phenotypic data, and one for the weather data. In practice, these can be treated as two separate SPARQL endpoints and they can be queried individually. In a real life scenario, it would be realistic to have a long list of SPARQL endpoints, but two are sufficient for the purposes of an integration showcase. With federation, these two endpoints can collaborate and respond to questions that address data on both - whereas each individual one cannot.

For the FDP, TTL files were composed with relevant metadata, and then exposed as plain text using a Python script (with the Flask library).

## Data and code availability

All associated data and code are available on the GitHub repository: FAIR-CxE contributors, 2020.

```

@prefix ppeo: <http://purl.org/pp eo/PPEO.owl#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xml: <http://www.w3.org/XML/1998/namespace> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

<#institute/96549626> a ppeo:institution ;
  ppeo:hasLocation <#location/57285852> ;
  ppeo:hasName "Laboratory of Plant Breeding, Wageningen University and Research" .

<#investigation/WUR_inv_CE2020> a ppeo:investigation ;
  ppeo:hasAssociatedPublication "https://library.wur.nl/WebQuery/wurpubs/fulltext/24
  ppeo:hasDescription "FAIRified, partial data from the 2012 thesis of Paula Ximena .
  ppeo:hasIdentifier "WUR_inv_CE2020" ;
  ppeo:hasLicense "CC-BY 4.0" ;
  ppeo:hasMIAPPEVersion "1.1"^^xsd:float ;
  ppeo:hasName "Investigating genotype by environment and QTL by environment interac
  ppeo:hasPart <#study/WUR_inv_CE2020_1999NL>,
    <#study/WUR_inv_CE2020_2003VE>,
    <#study/WUR_inv_CE2020_2004Fin>,
    <#study/WUR_inv_CE2020_2005Fin>,
    <#study/WUR_inv_CE2020_2010ET> ;
  ppeo:hasPersonWithRole <#person_role/WUR_inv_CE2020_2010ET_coordinator>,
    <#person_role/WUR_inv_CE2020_2010ET_corresponding_author>,
    <#person_role/WUR_inv_CE2020_2010ET_data_author>,
    <#person_role/WUR_inv_CE2020_2010ET_scientist>,
    <#person_role/WUR_inv_CE2020_2010ET_scientist_coordinator> ;
  ppeo:hasPublicReleaseDate "2020-01-01"^^xsd:date ;
  ppeo:hasSubmissionDate "2020-01-01"^^xsd:date .

```

**Figure 4.13:** Part of the generated file with RDF for the MIAPPE metadata, showing the Investigation declaration.

```

@prefix ppeo: <http://purl.org/pp eo/PPEO.owl#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xml: <http://www.w3.org/XML/1998/namespace> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

<#data_file/WUR_inv_CE2020_data_2010ET_csv> ppeo:hasDigitalLocation "data_2010ET.csv" ;
  ppeo:hasObservation <#observation/WUR_inv_CE2020_TubN_GE20_10004>,
    <#observation/WUR_inv_CE2020_TubN_GE20_10010>,
    <#observation/WUR_inv_CE2020_TubN_GE20_10016>,
    <#observation/WUR_inv_CE2020_TubN_GE20_10022>,
    <#observation/WUR_inv_CE2020_TubN_GE20_10028>,
    <#observation/WUR_inv_CE2020_TubN_GE20_1003>,
    <#observation/WUR_inv_CE2020_TubN_GE20_10034> .

```

**Figure 4.14:** Part of the generated file with RDF for the 2010 Ethiopia phenotyping experiment, showing the declaration of a data file with its location, and some of the observation (IDs) listed in it.





## Chapter 5

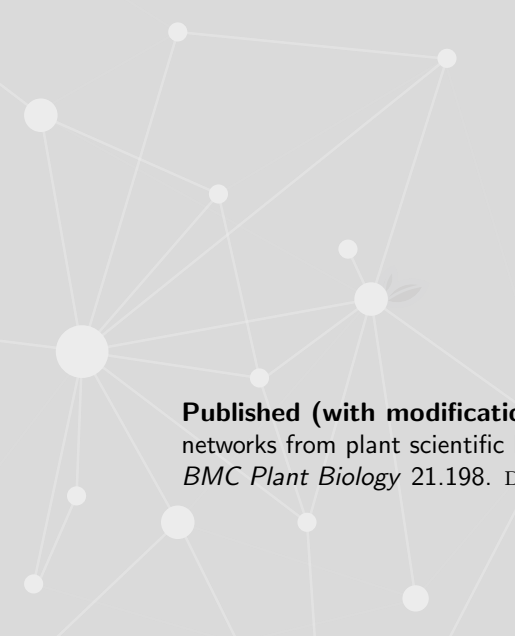
# Extracting knowledge networks from plant scientific literature: Potato tuber flesh color as an exemplary trait

Gurnoor Singh<sup>1,\*</sup>, Evangelia A. Papoutsoglou<sup>1,\*</sup>, Frederique Keijts-Lalleman<sup>2</sup>, Bilyana Vencheva<sup>2</sup>, Mark Rice<sup>2</sup>, Richard G.F. Visser<sup>1</sup>, Christian W.B. Bachem<sup>1</sup> and Richard Finkers<sup>1</sup>

\*Both authors contributed equally to this manuscript.

<sup>1</sup>Plant Breeding, Wageningen University & Research, Wageningen, the Netherlands

<sup>2</sup>IBM Netherlands, Amsterdam, the Netherlands



**Published (with modifications) as:** G. Singh *et al.* (2021). Extracting knowledge networks from plant scientific literature: Potato tuber flesh color as an exemplary trait. *BMC Plant Biology* 21.198. DOI: 10.1186/s12870-021-02943-5

## Abstract

### Background

Scientific literature carries a wealth of information crucial for research, but only a fraction of it is present as structured information in databases and therefore can be analyzed using traditional data analysis tools. Natural language processing (NLP) is often and successfully employed to support humans by distilling relevant information from large corpora of free text and structuring it in a way that lends itself to further computational analyses. For this pilot, we developed a pipeline that uses NLP on biological literature to produce knowledge networks. We focused on the flesh color of potato, a well-studied trait with known associations, and we investigated whether these knowledge networks can assist us in formulating new hypotheses on the underlying biological processes.

### Results

We trained an NLP model based on a manually annotated corpus of 34 full-text potato articles, to recognize relevant biological entities and relationships between them in text (genes, proteins, metabolites and traits). This model detected the number of biological entities with a precision of 97.65% and a recall of 88.91% on the training set. We conducted a time series analysis on 4023 PubMed abstract of plant genetics-based articles which focus on 4 major Solanaceous crops (tomato, potato, eggplant and capsicum), to determine that the networks contained both previously known and contemporaneously unknown leads to subsequently discovered biological phenomena relating to flesh color. Analysis of these networks indicates a connection between our trait and a candidate gene (zeaxanthin epoxidase) already two years prior to explicit statements of that connection in the literature.

### Conclusions

Our time-based analysis indicates that network-assisted hypothesis generation shows promise for knowledge discovery, data integration and hypothesis generation in scientific research.

## Introduction

Scientific publications accumulate knowledge and developments in any field of research. One of the most important tasks in a researcher's work and career is keeping up to date with the ever-increasing volume of scientific literature, placing new outputs into context, and investigating the implications in their field. However, as the number of scientific publications is growing at an exponential rate, there is a need to use artificial intelligence to enable a machine to read, extract, and analyze the information in textual sources.

Potato (*Solanum tuberosum* L.) is one of the most important staple crops for human nutrition. In addition to its culinary versatility, potato is a cost-effective product and plays a major role in meeting the ever-increasing food demands of the world. Its tubers are a good source of starch, proteins, vitamin C, folate, and provitamin A in the form of beta-carotene (Sulli *et al.*, 2017). Different potato genotypes produce tubers of different properties, like shape, size, color, starch content, and nutritional value.

One of the most extensively studied traits in potato is tuber flesh color. Potato tubers can have a wide range of colors, from orange to white and purple. Carotenoids are considered to be the primary determinant of tuber flesh color (C. Brown *et al.*, 2006). Carotenoids play essential roles in photosynthesis, while in non-photosynthetic tissues, they exert a broad range of functions acting as pigments, antioxidants, and precursors of signaling molecules, including volatiles (Giuliano, 2014). Previous studies have shown that beta-carotene and zeaxanthin are the components that predominantly determine potato tuber flesh color. In recent years, several candidate genes like beta-carotene hydroxylase (BCH/CHY2) and zeaxanthin epoxidase (ZEP) have been found to relate to tuber flesh color. BCH/CHY2 are the genes related to the production of beta-carotene while ZEP is considered responsible for the accumulation of zeaxanthin (Acharjee *et al.*, 2011). Scientific evidence for the association of tuber flesh color with genetic and molecular entities is found in the scientific literature or biological databases. For example, Acharjee *et al.* previously published networks of experimentally found biological entities that relate to tuber flesh color in the years 2011 and 2016 (Acharjee *et al.*, 2016, 2011). In this research, we automate the process of extracting knowledge of molecular entities (genes/proteins/metabolites) that influence changes in tuber flesh color from scientific publications.

Compared to structured information (as in databases), textual information is huge, noisy, and redundant. Artificial intelligence can help automate the processing of textual information and the discovery of new knowledge. Natural Language Processing (NLP) is a field of artificial intelligence that focuses on enabling machines to understand and analyze (unstructured) data in the form of text (Hirschberg and Manning, 2015). Despite the availability of various data repositories for plant research, a wealth of information currently remains buried within the scientific literature. Hence, information extraction via NLP is of growing interest and importance. NLP can render scientific texts computationally accessible, support information extraction, knowledge network (KN) construction and hypothesis generation.

In the past years, many NLP based research studies have been conducted on the literature from molecular biology (Harmston *et al.*, 2010; C.-C. Huang and Z. Lu, 2015). These focused primarily on rule-based named entity recognition (NER) i.e.

identifying and annotating biological entities such as genes or proteins (Baran *et al.*, 2011; Ding *et al.*, 2015), metabolites (Choi *et al.*, 2016; Galea *et al.*, 2018), traits (Endara *et al.*, 2018), QTLs (G. Singh *et al.*, 2018), diseases (Cho *et al.*, 2017), and drugs (Jang *et al.*, 2018) in literature. A few NLP studies paid attention to extracting associations (relationships and events) between these biological entities, using NER systems under the hood (Ding *et al.*, 2015; Hahn *et al.*, 2012; Van Landeghem *et al.*, 2013). Automated approaches to mining knowledge concerning the association of an entity to its phenotypes are required to further advance the field of precision breeding (Sharma *et al.*, 2017).

Rule-based NLP is more widely used in mining knowledge from biological context than machine learning-based NLP (Cook and Jensen, 2019; Kim *et al.*, 2019). However, construction and formalization of rules is a complex task in rule-based NLP. Often the rule-based NLP user tends to overfit the rules to the training set, which affects performance in the test set. Dictionaries and ontologies are used as building blocks in rule-based NLP. In supervised machine learning-based NLP, on the other hand, a domain specialist annotates the training set of documents manually. These manually annotated documents, supported by dictionaries and ontologies, are used by an algorithm to produce context-specific rules. Finally, these rules are used to perform NLP on the unannotated test set.

In this research, we investigated whether the latent knowledge in scientific literature can be harnessed with NLP, and if new leads for gene-trait associations can be highlighted for hypothesis generation in a timely manner. We chose to focus on the flesh color of potato tubers, an agronomically important trait with known associations. This enabled us to compare the relationships that we distilled from the literature with established facts, serving as a metric for the performance of our pipeline. It was necessary to validate more secondary hypotheses before we could focus on the time dimension of this question, namely 1) whether the NLP model is able to extract the expected relationships from the free text in literature; and 2) whether abstracts alone can act as high-certainty, information-dense proxies for their corresponding articles.

Our pipeline started with the NLP model, which was customized based on domain-relevant literature to find biological entities (genes, proteins, metabolites, and traits) and general relationships between them. We chose to use the commercial IBM Watson software suite, as it has been previously used to successfully mine knowledge from large corpora of texts available online (Y. Chen *et al.*, 2016; Ferrucci, 2012). Watson Knowledge Studio is a proprietary cloud-based application to train an NLP model based on the context and linguistic nuances of a specific literature domain. In addition to annotating entities of interest in a given text (named entity recognition), Watson also performs relationship extraction; that is, labeling the connections between the detected entities of interest. The relationships extracted by Watson were used to build KNs. After a normalization step, we were able to integrate these, and produce visualizations of the distilled knowledge from a set of texts.

We composed a primary corpus of 34 selected articles, mainly concerning potato flesh color, which we used to train our NLP model. Later we deployed it on a subset of these 34 (abstracts only) and a broader-spectrum corpus comprising 4023 PubMed abstracts, published from 2000 to 2016. For the former, we compared the nodes and the edges of the networks to test our secondary hypotheses. For the latter, we also



performed a time-based analysis, tracking the closeness of our trait of interest to other relevant entities, marking the time points where significant developments occurred, to evaluate whether this approach is indeed helpful for research.

This proof of concept (although limited in size) is an example of how literature mining can help plant scientists obtain a clearer “big picture” about specific areas in their field of expertise. Elusive findings in the expanding body of literature could come to light, be automatically organized into KNs, and ultimately help accelerate research in a process with little human intervention.

## Results

First, to confirm that our domain-specific NLP model performed as intended and extracted knowledge networks (KNs) with the focus on tuber flesh color from scientific literature, we deployed it on 2 different corpora, i.e. the training set with full-text articles and the test set with PubMed abstracts only. This was followed by a time analysis on the test set, to investigate whether the knowledge in these KNs could really be used in the way we envision, to generate new hypotheses.

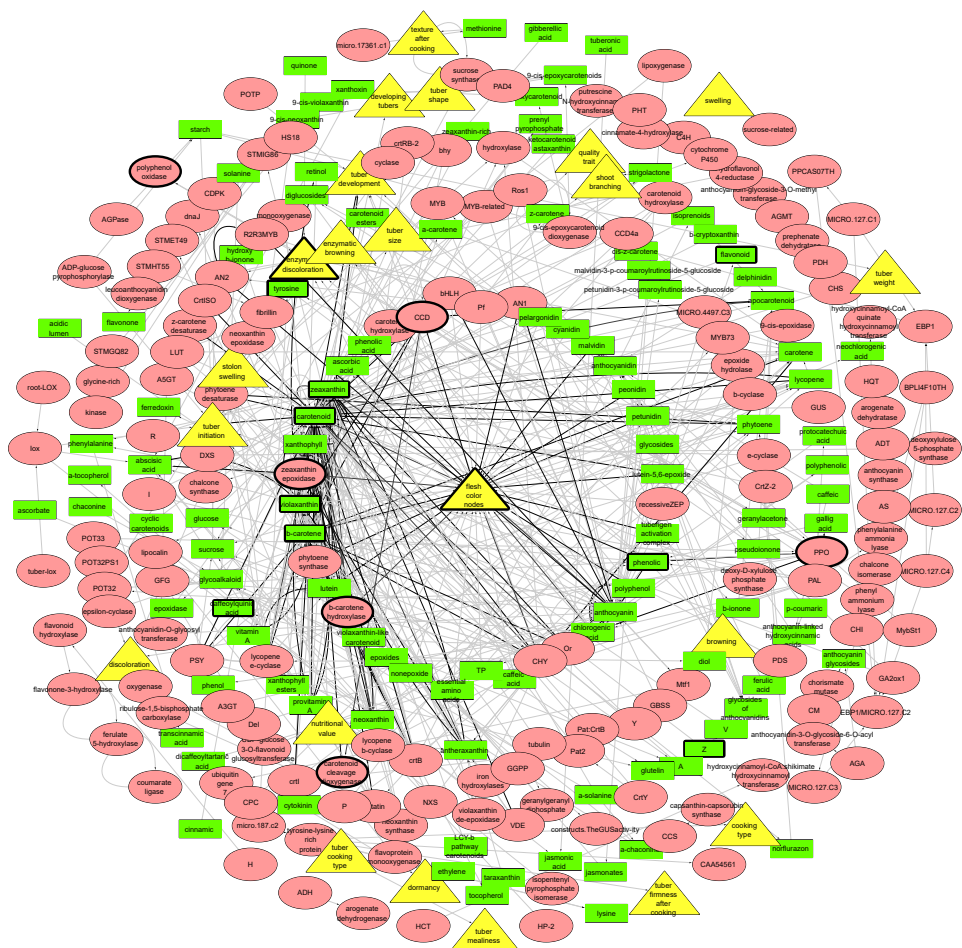
### Case 1: Analysis of training corpus (full-text articles)

We built a KN on the training set of 34 articles, with a total of 293 nodes and 551 unique edges. Out of these 293 nodes, there are a total of 159 genes/proteins, 112 metabolites and 22 traits (**Figure 5.1**). Carotenoids (an entity of the type metabolites) was the primary centroid of this network having 76 first-order neighbors. To evaluate the nodes and connections of this KN, we analyzed the overall structure based on the currently known experimental knowledge of tuber flesh color. Our KN contains scientifically credible links between nodes and the trait of interest, tuber flesh color. Most genes/proteins and metabolite entries in this network are part of the carotenoid biosynthesis pathway, which includes beta-carotene biosynthesis, xanthophyll cycle, abscisic acid biosynthesis, lutein biosynthesis, etc.

The trait under study, tuber flesh color, has 38 first-order neighbors, comprising 11 genes/proteins and 27 metabolites (the Cytoscape network can be found at (G. Singh and Papoutsoglou, 2019a)). These genes/proteins and metabolites are also listed in **Table 5.1**. Previously conducted research studies have found that ZEP and BCH/CHY are associated with white, yellow and orange flesh color. AN1, a gene responsible for the production of anthocyanin, is associated with purple flesh color. All these genes occur as direct neighbors of tuber flesh color in our network.

Our NLP model retrieved the entities in the training set with a precision of 97.65%, a recall of 88.91% and an F1 score of 93.07%. **Supplementary table S5.1** presents a confusion matrix showing the total number of entities per document, number of true positives (TP), number of false negatives (FN) and number of false positives (FP). Precision and recall were calculated as  $TP / (TP + FP)$  and  $TP / (TP + FN)$  respectively.

Additionally, to compare the difference in volume and quality of information extracted from abstracts vs. full-text versions of articles, our NLP model was applied separately on only the abstracts of the training corpus.



**Figure 5.1:** A KN representing knowledge triples found in the training set of 34 full articles. Yellow nodes refer to a trait entity, red nodes refer to gene/protein entities, and green nodes represent the metabolite entities. The centroid of this network is tuber flesh color. Nodes with bold outlines indicate that these entities have an experimentally proved association with tuber flesh color (trait of interest). This experimental evidence of these entities with tuber flesh color is reported in the articles (Acharjee et al., 2016, 2011).

| Set A                        | Set B                        | Set A - Set B                |
|------------------------------|------------------------------|------------------------------|
| AN1                          | anthocyanin                  | AN1                          |
| anthocyanidin                | ascorbic acid                | anthocyanidin                |
| anthocyanin                  | b-carotene                   | carotene hydroxylase         |
| ascorbic acid                | b-carotene hydroxylase       | cyanidin                     |
| b-carotene                   | bHLH                         | epoxides                     |
| b-carotene hydroxylase       | caffeic acid                 | essential amiacids           |
| bHLH                         | carotenoid                   | glycosides                   |
| caffeic acid                 | CCD                          | lutein                       |
| carotene hydroxylase         | chlorogenic acid             | lutein-5,6-epoxi             |
| carotenoid                   | CHY                          | malvidin                     |
| CCD                          | Or                           | nonepoxide                   |
| chlorogenic acid             | phenolic                     | pelargonidin                 |
| CHY                          | TP                           | peonidin                     |
| cyanidin                     | tuberigen activation complex | petunidin                    |
| epoxides                     | xanthophyll                  | Pf                           |
| essential amino acids        | zeaxanthin                   | phenolic acid                |
| glycosides                   | zeaxanthin epoxidase         | phytoene synthase            |
| lutein                       |                              | polyphenol                   |
| lutein-5,6-epoxide           |                              | recessiveZEP                 |
| malvidin                     |                              | violaxanthin                 |
| nonepoxide                   |                              | violaxanthin-like carotenoid |
| Or                           |                              |                              |
| pelargonidin                 |                              |                              |
| peonidin                     |                              |                              |
| petunidin                    |                              |                              |
| Pf                           |                              |                              |
| phenolic                     |                              |                              |
| phenolic acid                |                              |                              |
| phytoene synthase            |                              |                              |
| polyphenol                   |                              |                              |
| recessiveZEP                 |                              |                              |
| TP                           |                              |                              |
| tuberigen activation complex |                              |                              |
| violaxanthin                 |                              |                              |
| violaxanthin-like carotenoid |                              |                              |
| xanthophyll                  |                              |                              |
| zeaxanthin                   |                              |                              |
| zeaxanthin epoxidase         |                              |                              |

**Table 5.1:** Sets representing first order (direct) neighbors of flesh color nodes. Set A represents first-order neighbors of tuber flesh color nodes found in full-text articles. Set B represents first-order neighbors of tuber flesh color nodes found in abstracts of articles of the training set. The difference between these sets (SET A - SET B) represents all entities that are first-order neighbors of tuber flesh color in full-text articles, but not in abstracts alone.

This highlighted a quantitative difference between these two representations of a scientific article. We hypothesized that the abstract would concretely and concisely present the core outputs of a publication, whereas the introduction section would mainly recapitulate established theories and relevant biological connections but without contributing new knowledge. Finally, the results and discussion sections would combine, in greater detail, the significant contributions of the article, and make further suggestions for future experimentation. We found supporting evidence for this hypothesis, as the abstract-only network still includes the entities experimentally shown to be most important for tuber flesh color. In Sets A and B, **Table 5.1** lists the direct neighbors of tuber flesh color node in the KNs of full text representation (**Figure 5.1**) and abstracts only (**Figure 5.2**).

The difference between these two sets (**Table 5.1**; Set A - Set B) is also shown. These 20 entities occur as direct neighbors of flesh color in the full-text KN, but not in the abstract-only KN. Of these 20 entities, 6 (AN1, lutein, lutein-5,6-epoxide, polyphenol, phytoene synthase, violaxanthin) are still present in the KN of abstracts (**Figure 5.2**), even though they are not direct neighbors, but rather second-order neighbors of tuber flesh color and first-order neighbors of carotenoids, BCH, or ZEP. Furthermore, recessive ZEP is also represented in the abstract-only KN. Since the recessive allelic variant of ZEP is similar to the dominant one, these nodes are not represented as separate entities. The same applies to other aspects of gene/protein characteristics, such as chemical isomers and trait measures, which we grouped together with the main entity to reduce fragmentation in our KNs. The remaining 12 entities (nonepoxide, peonidin, anthocyanidin, petunidin, pelargonidin, cyanidin, pf, malvidin, epoxides, glycosides) are not represented in the abstract-only KN. These entities are associated with key metabolites causing changes in flesh color. However, they do not influence the trait directly. Hence, our results illustrate that the most important nodes in the full-text network are still present in the reduced abstract-only network.

## Case 2: Analysis of testing corpus (PubMed abstracts)

To assess how our NLP model performed on an unknown corpus, we deployed it on a testing corpus of 4023 abstracts from PubMed articles. Watson retrieved a KN with a total of 681 nodes and 976 unique edges (**Figure 5.3a**), more than in Case 1 (293 resp. 551), which means our model was able to identify new nodes and edges in this corpus. Carotenoid was again the primary centroid of this network, with 107 first-order neighbors. Our trait under study, tuber flesh color, has 21 first-order neighbors, comprising 9 genes / proteins and 12 metabolites (see Cytoscape network at (G. Singh and Papoutsoglou, 2019a)).

While our model is tailored toward potato tuber flesh color (ranging between white and orange), additional traits and their respective biological associations were detected as well. For example, the KN from the test set also detected genes/proteins and metabolites which influence other traits, such as enzymatic discoloration, tuber initiation, tuber development, tuber maturation, cooking types, stolon swelling, flower development etc. (**Figure 5.3b**). This illustrates that the information content extends beyond the specific use case. Moreover, our NLP model can extract information related to tuber flesh color in a wider context than the use case only, without requiring further specific training.





5

## Identifying emerging candidates with time analysis

To assess the accumulation of knowledge over time, the abstracts of the test set were organized in subsets ordered chronologically (i.e. by the date of their publication). Starting from the year 2000 and incrementing yearly (i.e. all publications up to 2000, all publications up to 2001, . . . , all publications up to 2016), subsets were formed. Each of these subsets was used to construct a separate KN. A network of a given year is always a subset of a KN from the following years and a superset of the previous years.

To study the development of entity connections with regard to our trait of interest (tuber flesh color), we worked backwards. The most recent collection was the most complete, so the nodes widely concerning tuber flesh color were chosen (color, flesh, flesh color, flesh trait, orange flesh color, tuber color, tuber flesh, tuber flesh color, white flesh color, yellow-orange flesh color) and are henceforth referred to as flesh color nodes. We focused our attention on the nodes that eventually ended up directly connected to a flesh color node. Then, we tracked the distance of these selected nodes to each individual flesh color node, and the changes over time. **Supplementary table S5.2** shows an example of such a table for changes occurring between 2009 and 2010. Scripts were finally written to parse the collections for all years in the corpus. Based on these year-by-year summaries, a master summary table was made (**Table 5.2**).

**Table 5.2** shows that the literature already contained significant indications as to the relevance of specific genes that were found to be important for potato flesh color (Acharjee *et al.*, 2011). Most prominently, both beta-carotene hydroxylase (BCH) and zeaxanthin epoxidase (ZEP) were in close proximity (2nd order neighbors) from 2007 onwards and made the transition to direct neighbors of flesh color nodes in 2010. While investigating the sentence that contributes to the transitions of ZEP in the time ranges from 2006 to 2010, we found that this gene was hypothesized to be associated with flesh color Dretto *et al.*, 2007b; Wolters *et al.*, 2010 before experimental evidence was published in 2011. The details about the literature (publication and exact sentences) providing these connections can be found in **Supplementary notes S5.1**.

Similarly, false positives such as lycopene, a metabolite not found in potato tubers, arise in the KN as first-order neighbors. While for most domain experts it is clear that lycopene is the compound responsible for flesh color in tomato, and therefore trivial to eliminate from the knowledge network as a significant player, it does reinforce the requirement for domain specialists to apply their knowledge to these results.

## Discussion

This work served as a pilot to study the benefits of using NLP platforms, such as Watson, for performing knowledge discovery in plant science literature. With the exponential increase in the number of scholarly publications and the sheer volume of available biological literature, researchers are finding it increasingly difficult to keep up-to-date with all information relevant to their field. Assembling knowledge from available literature in a single network is useful to generate new hypotheses or aid researchers in assembling a better overall picture of the components surrounding their area of interest. However, unlike for a human research expert, it is more challenging for a machine to comprehend biological insights from complex sentences and text structures

| node / year                         | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 |
|-------------------------------------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|
| CCD                                 | x    | x    | x    | x    | 3    | 3    | 3    | 3    | 3    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    |
| CHY                                 | x    | x    | x    | x    | x    | x    | 2    | 2    | 2    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    |
| DXS                                 | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    |
| PSY                                 | 2    | 2    | 2    | 2    | 2    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    |
| TP                                  | x    | x    | x    | x    | x    | x    | x    | x    | x    | 3    | 3    | 2    | 2    | 2    | 2    | 2    | 1    |
| abscisic acid                       | x    | x    | 4    | 2    | 2    | 2    | 2    | 2    | 2    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    |
| aminocyclopropane-1-carboxylic acid | x    | x    | x    | x    | x    | 5    | 5    | 5    | 5    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    |
| anthocyanin                         | x    | x    | x    | x    | x    | x    | x    | x    | x    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    |
| beta-carotene                       | x    | x    | 4    | 4    | 4    | 4    | 4    | 2    | 2    | 2    | 1    | 1    | 1    | 1    | 1    | 1    | 1    |
| hydroxylase                         | x    | x    | x    | x    | x    | x    | x    | x    | x    | x    | x    | x    | x    | x    | x    | x    | x    |
| bHLH                                | x    | x    | x    | x    | x    | x    | x    | x    | x    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    |
| carotenoid                          | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    |
| chlorophyll                         | x    | x    | x    | x    | x    | 3    | 3    | 3    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    |
| ethylene                            | x    | x    | x    | x    | x    | 5    | 5    | 5    | 5    | 3    | 1    | 1    | 1    | 1    | 1    | 1    | 1    |
| flavonoid                           | x    | x    | x    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    |
| flavonol                            | x    | x    | x    | x    | x    | x    | x    | x    | x    | x    | x    | x    | x    | x    | 1    | 1    | 1    |
| hydroxycinnamic acid                | x    | x    | x    | x    | x    | x    | x    | x    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    |
| lycopene                            | 3    | 2    | 2    | 2    | 2    | 2    | 2    | 2    | 2    | 2    | 1    | 1    | 1    | 1    | 1    | 1    | 1    |
| lycopene e-cyclase                  | x    | x    | x    | x    | x    | 2    | 2    | 2    | 2    | 2    | 1    | 1    | 1    | 1    | 1    | 1    | 1    |
| phenolic                            | x    | x    | x    | x    | 3    | 3    | 2    | 2    | 2    | 2    | 2    | 2    | 2    | 2    | 2    | 2    | 1    |
| phenylalanine                       | x    | x    | x    | x    | x    | x    | x    | x    | x    | x    | x    | x    | 3    | 3    | 3    | 3    | 1    |
| ammonia lyase                       | x    | x    | 3    | 3    | 3    | 3    | 3    | 2    | 2    | 2    | 1    | 1    | 1    | 1    | 1    | 1    | 1    |
| zeaxanthin epoxidase                | x    | x    | 3    | 3    | 3    | 3    | 3    | 2    | 2    | 2    | 1    | 1    | 1    | 1    | 1    | 1    | 1    |

**Table 5.2:** Overview of yearly changes in the network, based on the individual year summaries (for example **Supplementary table S5.2**). Each column represents a year, with an eventual neighbor of flesh color nodes listed in each row. The distances are the shortest path, at the time, from the node indicated to any flesh color node.



of scientific literature. Each NLP model has a limited scope of research questions it can address. The developed type system of our NLP model cannot capture and reflect all biological complexities in knowledge networks (KNs). However, our developed NLP model is intended to only mine genotypic-phenotypic information from scientific literature into KNs, so that this knowledge can be structured data, easily readable by both machines and humans.

Further, only generic relationships (“is related to”) of association between these entities were captured. The degree of association between two entities (positive, negative, inexplicit) was ignored in our model. The performance of our model, nevertheless, is satisfactory for the pilot study and addresses the above stated research objective. In order to optimize the efficiency of the process of manual annotation of the training set, we restricted ourselves to a limited training corpus of 34 full-text articles. Although training was thus limited, it was still sufficient to enable our model to extract similar knowledge from the test set, a collection of documents referring to different crops, traits and processes.

While making the testing corpus for our NLP model, we included literature from other *Solanaceae* crop species (tomato, capsicum, eggplant) as well. Mining and assembling information from all of these different literature resources into a single KN was somewhat controversial. Many genes and metabolites are involved in a similar bio-mechanism across these crop species. However, in some cases literature on other species may introduce noise, whereas in other cases it may be a source of ideas. There is a certain tradeoff to be observed here: the wider the scope of the processed documents, the higher the margin for noise, but also the potential. The premise for this trial, after all, was that newly published research in a broad domain of science would indiscriminately be funneled into an NLP model, to produce networks that can assist humans.

A similar balance exists when it comes to the parts of documents that are used for text analysis. Abstracts are an easily accessible and summarized form of significant information from an article. However, different journals prescribe different formats for their abstracts and other sections of scientific articles they publish. Therefore, the quality of minable information mentioned in an abstract depends on the journal as well as the type of article. Abstracts of articles such as reviews, scientific methods, or articles that cover a wide range of topics, might not provide comprehensive minable scientific leads. For example, in the journal *Nature*, contributions may not always formally describe all scientific leads in their abstract, and results are more frequently mentioned in the main text.

It is worth mentioning that there were instances where the NLP approach failed to meet expectations. In cases where biological entities were abbreviated, or associations between two entities were mentioned in more than one sentences, our NLP model could not predict these entities and relationships. Watson's type system includes facilities to co-refer abbreviated entries or pronouns to their original forms. However, due to the relatively small number of instances in our training corpus, Watson's NLP model was not able to capture these entities and relations. However, Watson is not unique in this respect. In fact, most NLP tools suffer from the same flaw. Biological abbreviations are haphazard. Frequently, two biological concepts have the same abbreviation. For example, an abbreviation MIC might mean Minimal Inhibitory Concentration, or refer to a Major Histocompatibility Complex (MHC) class I chain related (MIC) gene. Training

on a larger corpus might increase accuracy in predicting the correct entities.

Overall, our work produced a model that powered the construction and time analysis of meaningful KNs under restricted-effort conditions. We conclude that having the information we describe above available can provide key indications of scientifically relevant links, before such links are experimentally substantiated or published. Therefore, we believe that a more intensive effort would yield improved results and could play an important role in bringing together diverse information from large literature corpora and in hypothesis generation. Presently our KNs contain unweighted edges. In the future, we would like to assign edges a weighted score, based on experimental knowledge from databases, and the number of times a particular relationship occurred in text. In this way, text mining can be used to compare established and emerging knowledge.

## Conclusions

Our work strongly indicates that the computer-assisted extraction of knowledge from plant science literature can facilitate research. The results of our time analysis suggest that the individual components necessary for the formulation of new hypotheses may be published but remain unassociated for longer periods. Therefore, integrating these components into comprehensive knowledge networks can accelerate the generation of new hypotheses.

## Methods

### Experimental corpora

To make a supervised NLP model, we assembled scientific articles into 2 corpora, comprising a training set and a test set. The training set consisted of open source full-text articles, while the test set was built from PubMed abstracts.

The training corpus is a collection of 34 full-text scientific articles (see **Supplementary material S5.3**) which focus on tuber flesh color and known biological entities like metabolites and proteins involved in the carotenoid pathway, for example, beta-carotene hydroxylase and zeaxanthin epoxidase (Acharjee *et al.*, 2016). The training set was manually annotated with Watson Knowledge Studio (WKS). WKS uses these manual annotations to generate a supervised NLP model that can capture phenotypic tuber traits and the associated genes, proteins and metabolites. Later, we assessed the capabilities of this supervised NLP model to construct a knowledge network (KN) on this training set as well as on a larger test set.

The test set consists of 4023 abstracts from PubMed from the years 2000 to 2016 (which can be found at (Papoutsoglou and G. Singh, 2020a)). These abstracts are plant genetics-based articles which focus on 4 major Solanaceous crops (tomato, potato, eggplant and capsicum). To limit the scope of the NLP model to find direct genomic associations related to tuber flesh color, no pathogen related articles were included in the test set. Our developed NLP model is capable of extracting KNs for the tuber flesh color trait. However, the articles in the test sets deal with a variety of different topics in plant genetics and are not limited only to the tuber flesh color trait. This test set

challenges the NLP model to a more real-world application, as opposed to a restricted use case in our training set.

In addition, to analyze the difference between information contained in abstracts and full text representations of an article, we divided the training set into section-based subsets. We also divided the test set of abstracts into subsets based on their year of publication, to study the evolution of knowledge over time.

## Watson Knowledge Studio and Watson Explorer

IBM's Watson Knowledge Studio (WKS) is a proprietary text mining solution. It can be used to build machine learning models that perform named entity recognition (NER) and relationship extraction, using state of the art methods (Florian *et al.*, 2003; Kambhatla, 2004; M. C. McCord *et al.*, 2012; C. Wang *et al.*, 2012). The models can be tailored to different kinds of text (e.g. marketing, legal, scientific), and customized as to the type of annotations they produce.

To build a machine learning annotator in WKS, users must first define a type system to establish the "entities" (i.e. categories/classes of things that they wish for it to capture) and the "relations" between them. With the type system in place, they mark all occurrences ("mentions") of these entities and relations in collections of representative texts, producing a ground truth. Part of these collections, the training set, is then analyzed by WKS for linguistic structures, patterns and nuances specific to the domain, to produce the machine learning model. The other part, the test set, is only used to quantify the performance of the model (precision, recall). The type system and the annotations can be changed iteratively until the model performs satisfactorily.

Our final type system comprised three entities (Gene/Protein, Metabolite, Trait) and seven relations between them, as seen in **Figure 5.4**. We attained the best results with relations of a simple and all-encompassing nature, which is why many of the relations are only labeled as "related to". The exceptions ("encodes", "part of") were included since the high number of instances in the corpus allowed WKS to produce models that could successfully identify them in the text.

Each entity can be supported by an entity-specific dictionary. Dictionaries are used in a pre-annotation step of NER, before the corpus is annotated manually. To minimize noise (undesirable annotation of entities and relations), all dictionaries were made small and are limited to molecular entities known to be associated with tuber flesh color or with the carotenoid pathway. We selected our preferred labels from known molecular databases or ontologies. The Gene/Protein and the Metabolite dictionaries contain 183 genes/proteins and 85 metabolites, respectively. 56 potato-related traits taken from the Solanaceae Phenotype Ontology (*SPTO: Solanaceae Phenotype Ontology* 2018) comprise the Trait dictionary.

Watson Explorer (WEx) can use the model to annotate new documents. A schematic of its pipeline can be seen in **Supplementary figure S5.2**. Its outputs are text documents in XML/CAS files, containing annotations of the entities and their relations that have been extracted, and their documents (and document position) of origin. We use these XML/CAS files to build our KNs.

## Modeling decisions

To train our NLP model to capture KNs of only genotypic-phenotypic entities and their relationships, the type system underwent a number of major changes and revisions in an iterative process. With trial-and-error optimization, entities and relationships were introduced as well as discarded, based on how well the knowledge is captured and presented in the KN. In our analysis, a knowledge triple is defined as a data structure consisting of two entities and a label for their underlying relationship.

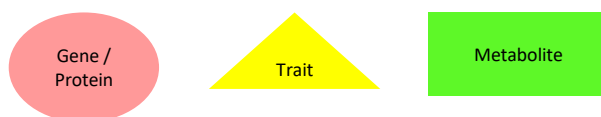
Some modeling decisions important to be mentioned are presented below.

- Biological entities that were tested but not included in the final model:
  - biochemical processes
  - metabolic pathways
  - trait values
  - organism names, species names and genotypes

While these biological entities occur in text and contain sources of knowledge to understand the biological mechanisms involved in the phenotypes, the numbers of mentions in the text were insufficient for WKS to adequately train a model. We therefore chose not to include these entities in the type system of our NLP model. Furthermore, including these entities in our model would have shifted the focus away from the research question of mining genotypic-phenotypic relationships in text.

- Combination of genes and proteins to a single entity:

### A. Entities



### B. Relationships

1. Gene / Protein *encodes* Gene / Protein
2. Gene / Protein *(is) related to* Gene / Protein
3. Gene / Protein *(is) related to* Metabolite
4. Gene / Protein *(is) related to* Trait
5. Metabolite *(is) part of* Metabolite
6. Metabolite *(is) related to* Metabolite
7. Metabolite *(is) related to* Trait

**Figure 5.4:** Watson Knowledge Studio (WKS) configurations of the type system for a customized NLP annotator. **(A)** 3 types of entities in the type system. **(B)** 7 types of relationships defined in the type system of an annotator.

Initially, we kept genes and proteins as two separate entities. However, during manual annotation, difficulties were encountered in distinguishing between the two, as they are frequently used interchangeably in the text. Furthermore, for subject matter experts, there is little information lost by combining them, and separating them introduced many misclassifications. Hence, in our type system genes and proteins are a single entity.

- Annotation rule for metabolites (specific metabolite mentions vs generic mentions):

Metabolites are included in scientific literature in different forms. Mentions may consist of specific composite terms (e.g. petunidin-3-p-coumaroyl-rutinoside-5-glucoside) or more generic ones (e.g. carotenoids). According to our type system, we annotated all forms of metabolite mentions as in this way we can capture both knowledge triples with specific entities and knowledge triples with generic entities.

- Annotation rules for genes:

As is the case with metabolites, genes may be introduced in different formats. Sometimes the full name is presented (zeaxanthin epoxidase), sometimes the short form (ZEP), and other times there is a species indicator as a prefix (LeZEP). We chose to annotate all these cases to train the model.

## Building and visualization of knowledge networks

For the construction of a KN, only entities with relationships were used. The mention of an entity by itself, with no connections, was not included in the KN. With help of Python scripts, we filtered out data of entities and relationships data from XML/CAS files (Papoutsoglou and G. Singh, 2020b). This script captured relationships as knowledge triples in easily parsable CSV files containing the relationship ID, relationship type, original mention of each entity, entity label, entity type, document in which this sentence occurred, sentence position and position of the source and target nodes.

As various entities appear in a variety of spellings in the corpus (e.g.  $\beta$ -carotene, b-carotene, beta-carotene), we also included a normalization step, attributing an additional preferred label to each entity. This was done manually on the list of individual entities that had been extracted. In the normalization process we first converted all spellings of entities and relationships to American English uppercase characters. Additionally, prefixes relating to species were removed from gene names. For example, the term StAN1, referring to anthocyanin 1 in *Solanum tuberosum* (potato), was converted to AN1. Similarly, suffixes indicating individual members of gene families were also removed, for example BCH1 and BCH2 (both referring to forms of beta-carotene hydroxylase), were converted to beta-carotene hydroxylase.

For metabolites, EC number references were converted to full names of enzymes. Further, apostrophes and # notations were removed, e.g. flavonoid-3',5'-hydroxylase becomes flavonoid-3,5-hydroxylase, 9#-cis-neoxanthin becomes 9-cis-neoxanthin. Lastly, all abbreviations were expanded to the long form, for example, NCED2 into 9-cis-epoxycarotenoid dioxygenase. These preferred labels were based on Uniprot (Pundir *et al.*, 2017) for genes/proteins, KEGG (Kanehisa *et al.*, 2016) for metabolites, and the Solanaceae Phenotype Trait Ontology (R. Shrestha *et al.*, 2012) for traits.

While the above steps reduce the specificity of a particular entity (for example we labeled BCH1 and BCH2 as BCH), as is always the case with tokenization, this simplification boosts network connectivity, despite the loss of information.

Finally, Cytoscape version 3.7.1 was used to visualize these KNs (Shannon *et al.*, 2003). Cytoscape can plot KNs using CSV files as input.

## Availability of data and materials

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Additionally, a supervised NLP model made on Watson Knowledge Studio (WKS) to extract genotypic-phenotypic relations in scientific articles of potato is archived here (G. Singh and Papoutsoglou, 2019b).

## Acknowledgements

We would like to thank Dick de Ridder (Bioinformatics, Wageningen University & Research), Willem Jan Knibbe (Data Competence Center, Wageningen University & Research) and Matthijs Brouwer (Plant Breeding, Wageningen University & Research) for critically reading the manuscript and for their valuable feedback. Additionally, we are thankful to IBM Technical Support for their continued support.

This article is based on the results of one of the chapters of the Ph.D thesis of one of the first authors Gurnoor Singh entitled " Genomics data integration for knowledge discovery using genome annotations from molecular databases and scientific literature" delivered at Wageningen University (G. Singh, 2019).







## Chapter 6

# General discussion



In this chapter I present an overview of the progress made in this thesis. First, I dive into an overview of the data-related challenges in research, and in plant phenotyping in particular, and outline how the FAIR data principles can resolve them. Then I detail the contributions of each chapter and reflect on their implications for the broader data landscape. With those insights in mind, I consider future prospects for 1) the general state of FAIR, and the specific needs for future implementation that became evident in this thesis, and 2) the road ahead for FAIR in plant phenotyping.

## Data-related challenges

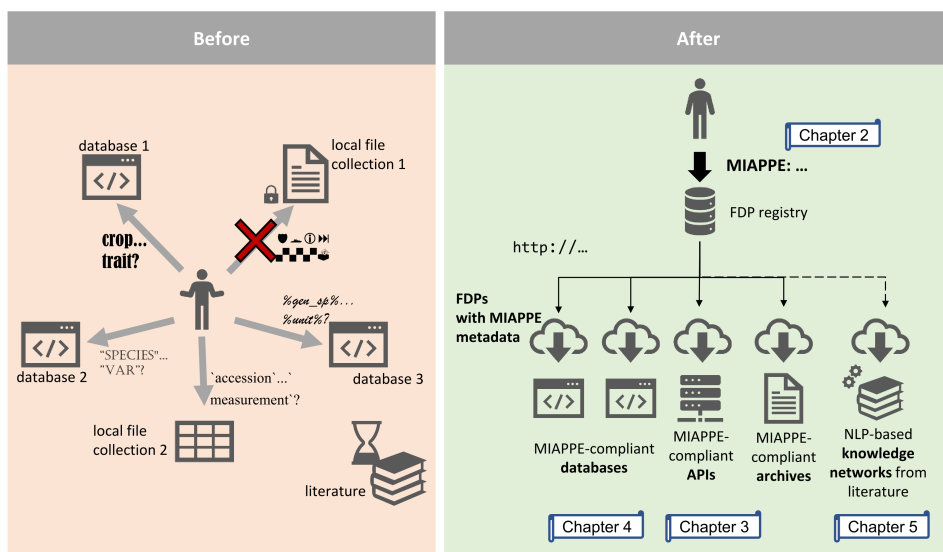
Over the last decades, scientists have grown increasingly aware of flaws in the commonly adopted attitudes toward data which hamper not only reusability but also experiment reproducibility, one of the cornerstones of the scientific method. The gravity of the situation as pertains to the latter is unquestionable; in fact, a 2012 review of biomedical and life-science research articles showed that 43.4% of retractions were due to “fraud or suspected fraud” (F. C. Fang *et al.*, 2012), with neither fraud itself nor unverifiable suspicions thereof boding well for science. There can be no reproducibility without reuse, depending on the definition of the latter, but the benefits of reuse extend beyond reproducibility, to meta-analyses (e.g. Chivenge *et al.*, 2011; Poorter *et al.*, 2012; J. Wang *et al.*, 2021), novel insight generation through data integration (e.g. Lee *et al.*, 2005; H. Tong and Nikoloski, 2021) and model-building pursuits (e.g. Bernal-Vasquez *et al.*, 2017; Heslot, 2014).

The FAIR data principles introduced guidelines to improve data and metadata practices not only to enable reuse but also to improve the conditions surrounding it: from licensing and documentation to rich semantics boosting machine readability. However, the principles are generic, and suggest practices that are heavily domain-dependent (for data and metadata descriptors). It is up to the respective communities to undertake the development of domain-specific data and metadata standards. This work focuses on plant phenotyping, a domain where the data landscape is challenged because of factors such as data fragmentation (multitude of institutional databases as opposed to a central repository), content heterogeneity (different kinds of phenotyping data, e.g. fruit weight, flowering times), variation in experimental settings (lab, field), designs (e.g. replications, layouts) and management practices (e.g. watering, fertilizer application), inconsistent or lacking documentation practices, and data file format and syntax variations.

This work explores different aspects of the FAIR principles for plant phenotyping, investigates the associated challenges and potential, and takes steps to further the state of the art for all of those aspects. **Figure 6.1** shows how the components presented in this work might come together to improve the data landscape for plant phenotyping, and **Figure 6.2** presents the potential for data integration, based on advances in this thesis, between phenotyping and domains related to it (genotyping, environmental data).

## The improved MIAPPE metadata standard

MIAPPE was originally conceived as a flat checklist to help plant scientists document their phenotyping experiments (Krajewski *et al.*, 2015). The developments on this

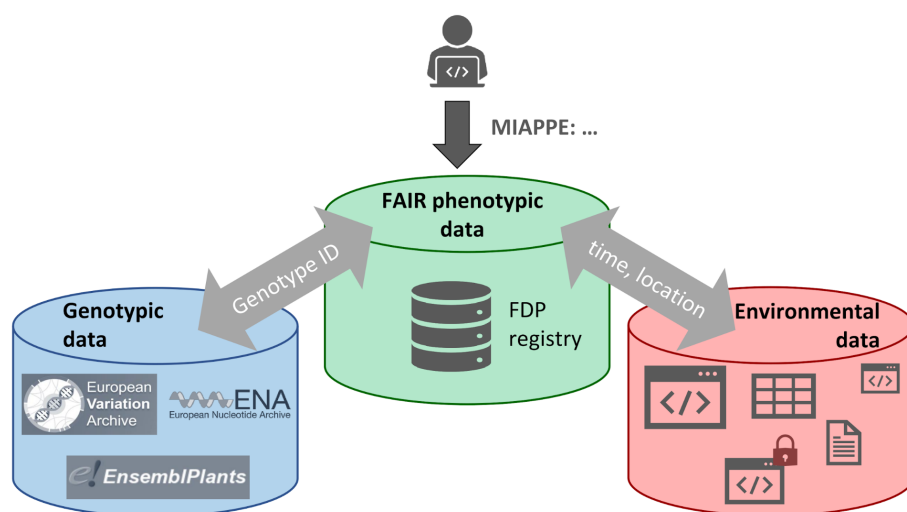


**Figure 6.1:** This diagram reflects the progress made in this thesis for the phenotyping domain. The left side (“Before”) shows the generally adopted approach in the domain when it comes to data collection: researchers have to look in a variety of repositories as there are no central hubs in the community. The data is served from institutional databases and local files, with heterogeneity in respect to content, way of querying, format and accessibility. The right side (“After”) shows the proposed architecture for FAIR plant phenotypic data and the components presented in this thesis. A scientist can direct their question to a central FAIR Data Point registry, which is responsible for forwarding it to individual FAIR Data Points. FAIR Data Points can be responsible for covering database, API and file archive metadata. Communication between all components must be conducted based on any MIAPPE, and the FDPs can point to any MIAPPE implementation. BrAPI is particularly suited to this kind of communication. This process obviates the need, on the scientist’s part, to look into individual institutional databases or use different APIs to discover relevant datasets (F), to find ways to request acquisition of those datasets (A), and to integrate them based on source-specific transformations and considerations (I). Compliance to community standards, in addition to the above, promotes data reusability (R). Methods proposed in this thesis can also be used to extract information from scientific literature with NLP, with the future prospect of making it FAIR and part of this ecosystem. Finally, the parts of the diagram that each chapter in this thesis has contributed the most to are indicated.

front are described in Chapter 2. Its new scope extension means that it can be more inclusive and provide a means for a greater part of the community to describe datasets in an organized manner. Furthermore, the documentation improvements render it more accessible to plant scientists and clarify ambiguities that were identified in its previous version. However, it is the addition of an explicit structure that truly elevates it beyond its previous status as a checklist into a formal data model with rich domain-specific semantics that can truly be used for FAIR.

This work introduces a data model for MIAPPE. It clarifies the connections between entities and attributes in it, and provides concrete MIAPPE implementations. This allows for metadata to be compared and contrasted with the structures used by institutional databases, other repositories or storage media. These steps are necessary for the harmonization of experimental datasets, and alleviate the burden associated with making assumptions about the structure and meaning of data acquired from external sources. More than that, the data model means that automated agents can explore the metadata reliably (since all possible attributes and connections are specified).

The formalized data model that was expressed as an OWL ontology in this work also fits in with the vision of the semantic web, in which the web is enriched with machine-readable, interlinked descriptors. Although all MIAPPE implementations unequivocally organize its content, the MIAPPE ontology is necessary to make it computer-readable



**Figure 6.2:** The three cylinders on the diagram represent three different domains: phenotyping, genotyping, and environmental. The arrows connecting them show the concepts based on which inter-domain data integration can be achieved. Both types of integration (between phenotypic and genotypic data; between phenotypic and environmental data) are currently achievable from the perspective of phenotyping datasets thanks to MIAPPE. Advances for standardization and semantic representation of datasets are necessary for genotypic and environmental data, but the groundwork laid by MIAPPE represents a significant first step.

in a global context. By providing unique identifiers (more specifically, URIs) for all MIAPPE concepts, the ontology can ensure that connections can be made to it even from other fields, including but not limited to mappings. For example, the concept of a genotype is important for sequencing, a domain that is served by other standards.

A significant advantage to having a computer-readable metadata standard is the explicit decoupling of metadata from data. The FAIR principles themselves make mention of this separation in Principle A2 (“Metadata should be accessible even when the data is no longer available”), which can be counterintuitive. MIAPPE, by presenting itself as a metadata standard, becomes a reminder that metadata should be available in a structured format and not only hidden away in the free text of scientific publications and documentation.

## The Breeding API (BrAPI)

Chapter 3 describes the development of BrAPI, a standardized REST API specification for the communication of plant breeding data. BrAPI has found wide community adoption in the years following its emergence. Its rapid development rate is evidence of the strong interest that plant breeding database users and developers have in it, and the biannual hackathons confirm that the adopters are committed and increasing in numbers. Another indicator for its success is the growing library of “BrAPPs” (<https://brapi.org/brapps>), i.e. tools that rely on BrAPI, developed by the community.

Although BrAPI boasts a scope that is not limited to the plant phenotyping side, this side is one of its focal points. Its development has been influenced by the progress of MIAPPE, and the contributions of this thesis ensure that BrAPI version 2 has a complete mapping to the standard. While MIAPPE ensures that BrAPI metadata adheres (content- and structure-wise) to community recommendations for better data reusability, BrAPI contributes majorly toward all other aspects of FAIR. Projects like FAIDARE <sup>(1)</sup> can build upon it by harvesting information from BrAPI endpoints and serving as a central aggregator portal for all of them. This synergy is perhaps also the biggest achievement of the plant phenotyping community to date with respect to FAIR.

The scope of BrAPI is however not limited to metadata. In fact, most BrAPPs make use of the data exposed through endpoints for visualization purposes, which can be convenient for initial explorations of unfamiliar datasets before one delves further into their attributes. The BrAPI community acknowledges also the potential of the API to serve as a linked data provider. With the MIAPPE ontology, this possibility could materialize in the form of JSON-LD, providing semantic contexts for BrAPI content for unique characterization of resources, as discussed above. This semantic enhancement would also facilitate the integration of BrAPI outputs with data from different sources, bridging together fragments of distributed data for better reusability.

## Lessons learned from the implementation of MIAPPE in a FDP to improve the reusability of plant phenotyping data

Chapter 2 presents the improved MIAPPE 1.1. The same chapter draws upon examples for the use of the standard to serve its evaluation, as well as the evaluation of its

<sup>(1)</sup><https://urgi.versailles.inrae.fr/faidare/>

implementations. Although the standard itself declares that it supports FAIR and is necessary for it, it is only a part of the picture behind the principles which should also feature tools and infrastructure to support FAIR.

In Chapter 4, we use MIAPPE to compose a broader picture of FAIR (meta)data and highlight features of the process as well as challenges. We use a dataset composed of experiments conducted in previous years by different people and express all metadata as dictated by the standard. We uniformize part of the data as well, transform everything to linked data and expose it on a triple store. FAIR data points (FDPs) have been proposed as part of the infrastructure that can support FAIR, organizing datasets into catalogs and exposing their metadata, in addition to various formats of the datasets. In this case we use an example FDP to present the dataset and point to the triple store where it resides.

The main contributions of this chapter are related to the process of making data FAIR and the human factor. There are three main actors involved in the process: 1) the original data producers; 2) the person who assembled different datasets for an integrated analysis; 3) the person who has made the data FAIR for further reuse. The transitions between actors 1-2 and 2-3 were charged with challenges related to documentation and ambiguities. Some of these challenges were resolvable with communication (a luxury that, unfortunately, should not be taken for granted when contacting former researchers or unfamiliar parties), but there was some information loss involved. As much as this process may resemble the telephone game, where messages get distorted upon each transmission, it is unfortunately unknown whether attributes were similarly distorted in the process of making data FAIR. From actor 3 onwards, information loss should be entirely avoidable.

The core message from this chapter is therefore that, indeed, making data FAIR should not be an afterthought that only occurs upon the completion of a project. Data should be shared following the FAIR principles by its original producers, and the planning for this process should be developed along with the experimental setup itself. The process of making existing datasets in general is resource intensive, as databases or local files kept by researchers are likely to be disorganized and incomplete with the information that domain standards will demand in the future. On the flipside, the message is also that, if data is FAIR, the steps required to reuse and integrate it are trivialized. The time investment for finding and acquiring a dataset through an FDP, based on specific search terms, decreases. The same applies to getting an overview of a dataset with respect to the experimental attributes deemed important by the community (e.g. biological materials, observed variables, experimental factors and designs). Finally, structured, standardized data can be readily processed and combined regardless of its source. Provided an initial investment, the potential for easy reuse is high.

## **The IBM Watson use case: Tuber flesh color**

The FAIR principles aim to encourage good data practices across the board and are usually only considered in the context of structured datasets. In this work, we also investigated another approach when it comes to locating, integrating and reusing information: one that focuses on scientific literature. The free text in articles is accessible to humans, but limited by the rate of human information consumption - as opposed to that of machine agents.

With our pilot on the IBM Watson use case, we explored the possibility of genotypic-phenotypic association mining with an NLP model for the construction of knowledge networks, and ultimately integration of knowledge in a way that can best support researchers (Chapter 5). This is a complementary path to the structured, well-described metadata requirements behind FAIR, but at the same time it is a step in the same direction, as it aims to make existing information easier to discover and make use of. The fact remains that we are making steps toward better-structured data which is more readily exploitable.

The Watson Potato experiment highlights the need for data integration as, even with a simplistic type system, it was possible to analyze connections in the network that hinted at biological associations before they manifested in an actual article. A more powerful model with a plurality of entities and more detailed relationships between them could leverage the full potential of Watson and offer further benefits to researchers.

The focus of this chapter (Chapter 5) on the potential of NLP for timely hypothesis generation is specific. The time analysis conducted is evidence that it does exist, but there are still several challenges to be tackled. A “guided” tour in a knowledge network, i.e. starting with the entities related to the researcher-user’s current objective, is indeed a feasible way to make use of integrated knowledge. We envision that researchers would look at 2nd/3rd degree neighbourhoods around their entities of interest, but on a larger scale, as the entities of interest increase, this may become more chaotic purely because of the number of nodes involved. A smarter approach would involve a computer-aided exploration. Such scenarios are explored by KNetMiner in the shape of graph pattern mining, graph interestingness and gene ranking (Hassani-Pak *et al.*, 2020) and would be applicable to biological networks relevant to our use case.

## Future prospects

### Needs for FAIR data in general

The FAIR data principles need to be implemented in different ways to address the needs of each community. Concerns such as confidentiality and scalability are paramount, but remain outside the scope of this work. With respect to the former, the recommendation remains that, even when data itself is not to be public, the existence of the dataset can be ascertained solely through metadata and ideally outlining a path to access for external parties. After all, data does not necessarily have to be open in order to ensure that it is FAIR. As far as scalability is concerned, this work makes note of a single point, which is independent of the specific technologies that can be chosen for FDP implementations. Namely, that the current FDP specification is lacking any indication that metadata describing the content of each dataset in detail (for plant phenotyping, e.g. biological materials, observed variables, experimental factors) should be attached to the dataset layer. As such metadata is plentiful, especially compared to the resource-descriptive metadata that FDP layers hold, it may impact scalability. Therefore, consideration should be given to the metadata that is necessary for the task.

Overall, this work contributed to the design, test and implementation of a (meta)data sharing standard for the plant phenotyping community that can address the needs for FAIR, at least with respect to elementary attributes of experiments. Plant sciences

however encompass more types of data beyond phenotyping, only some of which are currently subject to standardization. High-throughput data produced by automated phenotyping systems comprises ever more and bigger datasets, which carry metadata attributes, mainly from equipment manufacturers. Again, this data is highly heterogeneous and has to be connected to each specific use of those systems. Also, environmental data is central to field experiments, and indeed efforts are being made to improve its documentation through initiatives such as EMPHASIS <sup>(2)</sup>. The broader landscape, beyond plant sciences, holds promise as well as challenges for standardization and FAIR. The geo-information community, for example, has long been committed to producing standards in specific formats and widely adhering to them (Tom, 1994). On the other hand, as evidenced by this work (Chapter 5), the standardization of weather data remains unclear as a number of standards exist, though none are widely used. For standardization in general, and consequently FAIR, to succeed, not only does there need to be a respected consensus in each domain, but that consensus also has to be well-documented enough for members outside the given communities to understand and use.

In addition to scientific content, the structure of FDPs also needs to adhere to standards. An implementation specification does exist, but does not appear to be widely adopted at this moment. This is necessary as different domains need to be consistent in their implementation of FDPs, if FAIR interdisciplinary research is to become a possibility. Although the different FDPs can be fully independent in their implementations (as long as they uphold the domain-relevant formats and information), something akin to FDP registries need to exist: without those, information will remain hard to be publicly discovered and thus unusable. A FDP registry would conceivably list the base address of each FDP and possibly index the top level(s) of each, though the extent of that depth and the overall capabilities of such a service should be a point for future research. The same applies to the identification of providers, for each domain, that could feasibly host such a registry. To further conceptualize this on a global scale, a “registry of registries” would also have to be considered, as scientists may not be necessarily aware of central data service providers in different disciplines.

An open question remains with respect to FAIR and linked data. Although the principles themselves make no explicit mention of linked data technologies, they are a commonly interpreted corollary. Rich annotations and metadata as well as unique identifiers fit into this vision, and enable linking data originating from different sources. In particular, establishing controlled vocabularies or ontologies, on top of semantic data models, is necessary for identifying objects in an open semantic web. The more specific these vocabularies are, the more accurate the descriptions that they are able to provide to allow computers to act on the data. On the other hand, humans are challenged to use extensive vocabularies and over-specific and complicated models, especially potential re-users outside of their native domain (e.g. a plant scientist attempting to reuse weather data). There needs to be a balance between these two extremes. Unfortunately, only part of this challenge can be alleviated through tools such as ontology lookup services. Therefore, it may turn out important for data management guidelines to evolve along with the willingness of users to comprehend the more intricate sides of linked data.

---

<sup>(2)</sup><https://emphasis.plant-phenotyping.eu/>



Secondary to the above comes the issue of trustability. Having an ocean of data at one's disposal, all of which is in theory ready to be exploited, comes with the requirement of responsible reuse. High quality data can lead to high quality outputs, but when a myriad of sources are readily accessible, verifying the validity of each can be time-consuming. Consequently there is a need for these datasets to be curated, but no clear answer as to the authority that might be responsible for such actions, or the way in which such a task may be undertaken. A starting point could be data validation with tools such as shape expressions or the shapes constraint language. Even with good quality datasets (meaning datasets that correctly and comprehensively follow annotation guidelines), a re-user may be led to poor decisions and conclusions, only because interdisciplinary research is inherently delicate, and effortless access to datasets adjacent to one's own domain may be too easy to warrant a deeper examination of those datasets. Therefore, the role of more authoritative sources and peer review for data will become even more important.

A solution to the above conundrum would be to eliminate the human factor, and rely on well-defined computer algorithms to identify compatible datasets across domains, or even within the same domain. For the latter, it is conceivable that smart applications could be produced to identify at least a subset of datasets that could be used in an analysis alongside one's own, or for a given purpose. The former scenario would be significantly more complicated to envision. Overall, FAIR data holds promise for reproducibility. In this work, we ensure reproducibility with the provision of Docker containers holding the data and the databases, and with Jupyter notebooks which can run on those (Kluyver *et al.*, 2016; Merkel, 2014). This means that there are no ambiguities about the data or any step of the process that was followed. More progress has been made on the FAIR data principles as they should apply to software by Lamprecht *et al.*, 2020.

All in all, it is important to remember that FAIR is not a binary state, and that the benefits that accompany it will increase in proportion with the adoption across and within communities. There is ongoing work for recommendations for the quantification of resource FAIRness (Wilkinson *et al.*, 2018), which can act as motivator to data holders (i.e. better numerical scores would be more desirable) since the principles are being promoted by official bodies (for example, the European Union (Collins *et al.*, 2018)). Eventually, a critical mass of users and FAIR datasets should be reached that can sufficiently demonstrate that FAIR data is really worth the investment, and that it can indeed move from being a liability to something that researchers benefit from on a daily basis.

## FAIR data in plant phenotyping

This work takes steps toward FAIR plant data by contributing to metadata standards (Chapter 2), data exchange interfaces (Chapter 3), case studies (Chapter 4) and data mining (Chapter 5). It is clear that, while these steps provide real-world value, they are not the final ones in the process.

First and foremost, data curation following the FAIR principles needs to be promoted to the researchers and data holders - not as a general concept which is difficult to disagree with - but as a process that individual users need to follow. Currently, the return on investment for efforts aiming to make data FAIR may be perceived as low, which

can be tackled from either perspective, that of the return or that of the investment. At these initial stages, the more realistic approach may be to decrease the initial investment. This can be done by providing clear instructions to scientists, minimizing the time required to make a dataset FAIR, and ensuring that data stewards are available. Part of the instructions should be about the standards that are adopted, and another part should be about the process surrounding the management of the (meta)data produced. The required time can be reduced with the creation of graphical interfaces that guide scientists and behave “smartly”, pre-filling fields and making automatic suggestions wherever possible. Data stewards should be able to answer questions, address concerns and provide technical support. To support these efforts, central FAIR data hubs/registries should be created, pointing to institutional and other repositories that expose FAIR datasets. On the other side, increasing the given return may be more difficult until that critical mass of FAIR data has been reached in a domain. Until then, researchers should be made more aware of the attribution advantages that may come to them (in the form of citations) (Piwowar *et al.*, 2007). Finally, scientific journals should push for standardized metadata more strongly, when a publication comes with an accompanying dataset. This should not remain on the high level of descriptive resource metadata, but dive into domain-specific descriptors.

The plant phenotyping community has invested in the development of ontologies and controlled vocabularies, as seen for example in the Crop Ontology, the Plant Trait Ontology, the Plant Ontology. They are most commonly used in databases that have been explicitly designed to support them, and now they are promoted by the MIAPPE standard. Their limited use in literature however tells another story that demonstrates the lack of awareness, as far as most plant scientists are concerned, insufficiency due to their scope, or both. In Chapter 4, an effort was made to assign Crop Ontology identifiers (specifically, from the Potato Ontology (Research Informatics Unit (RIU), CIP, 2020)) to the variable metadata and, in most cases, this effort was fruitless: even when a trait does exist in the ontology, the current trait-method-scale combinations were rarely suitable for reuse. This indicates that, indeed, further development of these vocabularies is necessary, or perhaps additionally a change in the organizational scheme followed. For example, **Figure 6.3** shows that for the trait “Average of tuber weight”, the only given method is associated with a per-plot average value. A modular approach is necessary, able to combine information across scales and offer conceptual foundation for the phenomenology underlying scientific observations (Villa *et al.*, 2017).

NLP could give a means to rapidly expand controlled vocabularies revolving around plant traits, anatomy, environments, or other experimental documentation details. In Chapter 5, we saw that IBM Watson was able to detect relationships between entities reliably, based on the domain-specific training it received. This pilot was restricted to training on data about potato and the flesh color of its tubers, and it is true that scientists may use different language to describe other plants and traits of other categories. A more general training, on a broader scope, could contribute to a model that is able to identify entities that should be included in controlled vocabularies.

Taken a step further, there is a path for NLP to truly contribute to FAIR in a more direct way. Many phenotyping experiments are described in unstructured text, published or otherwise. For a human, examining those texts in detail to locate attributes that belong to the MIAPPE metadata is a time consuming task. A model trained to identify

these attributes in the text and use them to present at least part of the metadata that pertains to an experiment would be an important facilitator for data reusability. Even if this last step is skipped and the attributes are only annotated in the text, it could be a great help - for example, that would have been the case in the metadata collection stage of Chapter 4.

FAIR plant (meta)data hold great potential for science, but there are parties other than scientists that are interested. The motivation behind the Farm Data Train (FDT) includes farmers themselves in the group of primarily interested stakeholders. Precision agriculture has become a driver for agricultural decisions, integrating heterogeneous data from on-situ sensors, drones and genome analyses. The distributed nature of the data and the variety of sources necessitate some means for easy sharing and interoperability, for which the FAIR principles are highlighted. Chapter 4 includes the building blocks for a FDT infrastructure in their simplest form: A data station (FAIR data point - FDP) is contacted by a train (Jupyter notebook) and serves its data to it. The connection between the train and the station in this case is direct - there is no need for matchmaking between queries and resources, therefore no track for the FDT. For a more realistic scenario, this track is crucial: there would be many more trains and stations, each with different data types and restrictions, so the track would have to direct each train (query) to the appropriate station(s) and deliver the response back. The FDP can use the MIAPPE standard for phenotypic metadata as in Chapter 4, and based on that metadata, the track can determine which datasets on which FDPs can address which train's response. BrAPI can be the implementation used to communicate data as well as metadata, as privacy is also covered in it and it can filter out unauthorized requests.

In this thesis, decisive steps were made toward establishing FAIR plant data for the FDT to use. MIAPPE itself can provide identifiers for biological materials (accessions, genotypes) that can be used for integration with genotypic data, a lot of which is already available on platforms such as EVA (EBI, 2020) and ENA (Leinonen *et al.*, 2010). We

The screenshot displays the CIP Potato Ontology web interface. At the top, there is a search bar containing 'CO\_330' and navigation links: 'Add New Terms', 'API', 'Help', 'Agtrials', 'Annotation Tool', 'Register', and 'Login'. Below the search bar, the 'Traits, methods and scales' section is visible, showing a hierarchical tree of terms. The 'Average of tuber weight - method' term is highlighted. To the right, the 'Term information' section provides details for the selected term:

| Term information   |   |
|--|---|
| Average of tuber weight - method <span>Permalink</span> <span>General</span> <span>0 Comments</span> |   |
| Identifier   | CO_330.0000332  |
| Formula  | $\frac{[(\text{Total tuber weight}/\text{plot})/(\text{Total number of tubers}/\text{plot})]*1000}{}$ |
| Method class   | Computation   |
| Method description   | Compute the average tuber weight in grams using the formula   |

**Figure 6.3:** The only method associated with the “Average of tuber weight” trait currently (as of February 2021) in the CIP Potato Ontology (Research Informatics Unit (RIU), CIP, 2020).

show that environmental data can be integrated as well, e.g. when it comes to field trials or farms, thanks to attributes pertaining to the time and location of an experiment. This would necessitate the provision of FAIR weather data as well, which merits further investigation in the environmental domain. However, once those attributes are not only present but also FAIR for these plant-related domains (genotyping and environmental), integration and better exploitation of combined datasets should become trivial. Thus they fit into the scenarios that the FDT envisions and into the analyses it is expected to power.

In a recent review of machine learning applications in plant sciences and breeding, integration is presented as essential for the composition of datasets that can support analyses intended to disentangle the relationships between genotype, phenotype and environment (A. D. J. v. Dijk *et al.*, 2020). An overview of articles (**Table 6.1**) describing multi-environment analyses (ones published in 2020) shows that the description of the datasets used is not always clear, in particular when it comes to whether they were generated for previous analyses and reused, or generated specifically for the purpose of the analysis presented in the publication. For the most part, they are performed with data likely collected by a single stakeholder - so data is likely not being reused for such scenarios. Most of these multi-environment trials rely on data from relatively few (12 or fewer) locations. Usually, when information is there, it is provided on a very high level (e.g. for environment measurements, experimental designs, treatments, cultural practices). None of the examined publications present metadata in a structured way, therefore reuse would be time consuming, if even possible at all (since some elements are not documented at all). Importantly, FAIR datasets and metadata would eliminate such doubts and make the generation of such an overview easier. This underlines the fact that FAIR data will be highly advantageous in the years to come as crop stability (e.g., yield) studies (which must use multiple environments by definition) are becoming increasingly common. The scientists conducting them would not only gain access to more datasets, but they would also find them readily interpretable and with a low integration cost.

The current landscape of data practices in plant phenotyping is broad, heterogeneous and messy. It is common for departments using such data to come up with their own guidelines for better data management. However, data management plans usually only cover top-level attributes about the resources and have no specification about the description of the dataset contents, as MIAPPE does. The first and most important step to be taken to start transitioning toward a more FAIR state of data is simple: it is about accepting the idea that metadata/documentation according to community guidelines is not optional, and neither is the official allocation of time for such tasks. This can be facilitated with the provision of adequate examples, training workshops and support personnel. Furthermore, the establishment of databases to host this metadata would encourage reuse and collaborations - though of course privacy concerns need to be addressed. These steps are not technically complicated, but generally face resistance due to time or willingness constraints on the human side.

The human side - researchers - often don't perceive the benefits of FAIR data. It is equated with mountains of paperwork that exists only to satisfy management criteria; it is only standing in the way of science instead of supporting it. As discussed, the benefits will become clearer when a critical mass of adopters is reached. However, until then,

relying on the efforts of the few enthusiastic members of the scientific community who see the “carrot” of the situation, may not be enough, however promising the results of Chapter 4 may be. We are given a limited peek into the benefits: metadata and established (meta)data structures trivialize integration between different phenotypic datasets and between phenotypic and weather data precisely because data is FAIR. A supplementary “stick” will probably be necessary in some form, as indicated earlier, to supplement the numbers of those critical early adopters.

With the concluding remark of this thesis, I would like to emphasize once again that better data management based on rich metadata (MIAPPE on the phenotyping side, other standards for other domains), structured data and repositories that collect them can pave the way toward more, better, small or large scale analyses. Plant research in particular would benefit from investigations into different aspects of plant biology and biotic/abiotic interactions on an unprecedented scale.

Table 6.1 (multiple pages)

| # | Title   | reuse?   | T/L/E                  | S/Y | L   | C   | ED      | T   | OV  | Soil | Rain | Temp | CV  |
|---|---|--|------------------------|-----|-----|-----|---------|-----|-----|------|------|------|-----|
| 1 | Adaptability and stability analyses of plants using random regression models<br>(Souza <i>et al.</i> , 2020)  | unclear  | 13 trials /<br>2 loc.  | yes | yes | yes | general | yes | yes |      |      |      |     |
| 2 | Broomrape as a major constraint for grass pea ( <i>Lathyrus sativus</i> ) production in mediterranean rain-fed environments<br>(Rubiales <i>et al.</i> , 2020)                                    | reuse of experiments where the authors were involved | 17 trials /<br>3 loc.  | yes | yes | yes | general |     | yes | yes  | yes  | yes  | yes |
| 3 | Multiple-trait, random regression, and compound symmetry models for analyzing multi-environment trials in maize breeding<br>(Ferreira Coelho <i>et al.</i> , 2020)                                | original datasets                                    | 4 trials /<br>1 loc.   | yes | yes | yes | general |     | yes | yes  | yes  | yes  | yes |
| 4 | Genetic dissection of component traits for salinity tolerance at reproductive stage in rice<br>(Chattopadhyay <i>et al.</i> , 2020)   | original datasets                                    | 2 trials /<br>1 loc.   | yes | yes | yes | general | yes |     | yes  |      |      |     |
| 5 | Genotype-by-environment interaction and yield stability analysis of biomass sorghum hybrids using factor analytic models and environmental covariates<br>(I. C. M. Oliveira <i>et al.</i> , 2020) | unclear; reuse of weather data                       | 29 trials /<br>13 loc. | yes | yes | yes | general |     | yes |      | yes  | yes  |     |
| 6 | Integrating univariate and multivariate statistical models to investigate genotype x environment interaction in durum wheat<br>(Mohammadi <i>et al.</i> , 2020)                                   | unclear  | 4 trials /<br>4 loc.   | yes | yes | yes | general |     | yes |      | yes  | yes  | yes |

Table 6.1 (multiple pages)

| #  | Title  | reuse?            | T/L/E                          | S/Y | L   | C   | ED      | T       | OV  | Soil | Rain | Temp | CV      |
|----|--|-------------------|--------------------------------|-----|-----|-----|---------|---------|-----|------|------|------|---------|
| 7  | Achievements and challenges towards a sustainable conservation and use of 'Galega vulgar' <i>Olea europaea</i> variety (Sales <i>et al.</i> , 2020)                                      | N/A: review       |                                |     |     |     |         |         |     |      |      |      |         |
| 8  | Genetic basis of phenotypic plasticity and genotype x environment interactions in a multi-parental tomato population (Diouf <i>et al.</i> , 2020)  | original datasets | 12 trials / 3 loc.             | yes | yes | yes |         | yes     | yes | yes  |      | yes  | yes     |
| 9  | Genotype-by-environment interaction analysis across three crop cycles in sugarcane (Momotaz <i>et al.</i> , 2020)  | unclear           | 5 trials / 5 loc.              | yes | yes | yes | general |         | yes | yes  | yes  | yes  | yes     |
| 10 | Genotype by environment interaction on resistance to cassava green mite associated traits and effects on yield performance of cassava genotypes in Nigeria (Jiwuba <i>et al.</i> , 2020) | unclear           | 6 trials / 3 loc.              | yes | yes | yes | general |         | yes | yes  | yes  | yes  | yes     |
| 11 | Genotype by environment interaction for oil quality components in olive tree (Navas-Lopez <i>et al.</i> , 2020)  | unclear           | 5 trials / 5 loc.              | yes | yes | yes | general |         |     | yes  |      |      |         |
| 12 | Retrospective quantitative genetic analysis and genomic prediction of global wheat yields (Juliana <i>et al.</i> , 2020)   | reuse             | 534(519) + 36 trials / 60 loc. | yes | yes |     | partial | partial |     |      |      |      | partial |

Table 6.1 (multiple pages)

| #  | Title   | reuse?  | T/L/E             | S/Y | L     | C   | ED      | T | OV  | Soil | Rain | Temp | CV  |
|----|---|---|-------------------|-----|-------|-----|---------|---|-----|------|------|------|-----|
| 13 | Nonlinear kernels, dominance, and envirotyping data increase the accuracy of genome-based prediction in multi-environment trials (Costa-Neto <i>et al.</i> , 2021)  | reuse of institutional datasets; some authors were involved in producing them | 2 trials / 2 loc. |     | yes   |     |         |   | yes |      | yes  | yes  | yes |
| 14 | Multi-trait multi-environment diallel analyses for maize breeding (Coelho <i>et al.</i> , 2020)   | original datasets   | 4 trials / 4 loc. | yes | yes   |     | general |   | yes |      | yes  | yes  | yes |
| 15 | Genome-based trait prediction in multi-environment breeding trials in groundnut (Pandey <i>et al.</i> , 2020)   | unclear   | 6 trials / 3 loc. | yes | yes   | yes | general |   | yes |      |      |      |     |
| 16 | CERES-Maize model for simulating genotype-by-environment interaction of maize and its stability in the dry and wet savannas of Nigeria (Adnan <i>et al.</i> , 2020) | original datasets; reuse of weather data                                      | 8 trials / 4 loc. | yes | yes   | yes | general |   | yes | yes  |      |      | yes |
| 17 | Genomic prediction enhanced sparse testing for multi-environment trials (Jarquin <i>et al.</i> , 2020a)   | unclear   | unclear / 3 loc.  |     | vague | yes | unclear |   |     |      |      |      |     |
| 18 | Genomic prediction applied to multiple traits and environments in second season maize hybrids (A. A. d. Oliveira <i>et al.</i> , 2020)                              | unclear   | 6 trials / 1 loc. | yes | yes   |     | general |   | yes |      |      |      |     |



Table 6.1 (multiple pages)

| #  | Title   | reuse?  | T/L/E                 | S/Y | L   | C       | ED      | T   | OV      | Soil | Rain | Temp | CV  |
|----|---|---------|-----------------------|-----|-----|---------|---------|-----|---------|------|------|------|-----|
| 19 | Combining crop growth modeling with trait-assisted prediction improved the prediction of genotype by environment interactions (Robert <i>et al.</i> , 2020)                 | reuse   | 42 env. /<br>18 loc.  | yes | yes |         | general | yes |         | yes  |      | yes  |     |
|    |   |         |                       |     |     |         |         |     |         |      |      |      |     |
| 20 | Additive main effect and multiplicative interaction analysis for grain yield in bread wheat (Khan <i>et al.</i> , 2020)   | unclear | 9 env.                | yes | yes | yes     | general |     |         |      | yes  | yes  | yes |
| 21 | G x E interactions in QTL introgression lines of Spanish-type groundnut ( <i>Arachis hypogaea</i> L.) (Rathnakumar <i>et al.</i> , 2020)                                    | unclear | 5 trials /<br>5 loc.  | yes | yes | yes     | general |     | yes     | yes  | yes  |      |     |
| 22 | Multi-environmental evaluation of maize hybrids developed from tropical and temperate lines (Mushayi <i>et al.</i> , 2020)  | unclear | 5 trials /<br>5 loc.  | yes | yes | partial | general |     | partial | yes  | yes  | yes  | yes |
| 23 | Strengths and weaknesses of national variety trial data for multi-environment analysis: A case study on grain yield and protein content (Rahimi-Eichi <i>et al.</i> , 2020) | reuse   | unclear /<br>206 loc. |     |     | yes     |         |     |         |      |      |      |     |
| 24 | Genotype x Environment interaction patterns in rangeland variety trials of cool-season grasses in the western United States (Robins <i>et al.</i> , 2020)                   | unclear | 5 trials /<br>5 loc.  | yes | yes | yes     | general |     | yes     | yes  | yes  | yes  | yes |

Table 6.1 (multiple pages)

| #  | Title  | reuse?   | T/L/E               | S/Y | L   | C   | ED      | T | OV  | Soil | Rain | Temp | CV      |
|----|--|--|---------------------|-----|-----|-----|---------|---|-----|------|------|------|---------|
| 25 | Key locations for soybean genotype assessment in South Brazil region (Dalló <i>et al.</i> , 2020)  | reuse  | 132 env. / 43 loc.  | yes | *   | *   | general |   | yes |      |      |      |         |
| 26 | Increased prediction accuracy using combined genomic information and physiological traits in a soft wheat panel evaluated in multi-environments (Guo <i>et al.</i> , 2020) | original dataset; reuse of weather data            | 4 trials / 2 loc.   | yes | yes |     | general |   | yes |      |      |      | yes     |
| 27 | Optimization of Eucalyptus breeding through random regression models allowing for reaction norms in response to environmental gradients (Alves <i>et al.</i> , 2020)       | unclear  | 4 trials / 4 loc.   |     | yes | yes | general |   | yes |      | yes  | yes  |         |
| 28 | Resistance to legume pod borer ( <i>Maruca vitrata Fabricius</i> ) in cowpea: genetic advances, challenges, and future prospects (Sodedji <i>et al.</i> , 2020)            | N/A  | N/A: review         |     |     |     |         |   |     |      |      |      |         |
| 29 | Adaptation of one-flowered Vetch ( <i>Vicia articulata Hornem.</i> ) to mediterranean rain fed conditions (Rubiales and F. Flores, 2020)                                   | unclear; reuse of weather data                     | 9 trials / 3 loc.   | yes | yes | yes | general |   | yes | yes  | yes  | yes  | yes     |
| 30 | Multi-environment analysis of sorghum breeding trials using additive and dominance genomic relationships (Hunt <i>et al.</i> , 2020)                                       | reuse; datasets generated by authors' institutions | 16 trials / 12 loc. | yes | yes |     | general |   | yes |      |      |      | partial |

Table 6.1 (multiple pages)

| #  | Title   | reuse?   | T/L/E                | S/Y | L   | C   | ED      | T | OV      | Soil | Rain | Temp | CV |
|----|---|--|----------------------|-----|-----|-----|---------|---|---------|------|------|------|----|
| 31 | One compound approach combining factor-analytic model with AMMI and GGE biplot to improve multi-environment trials analysis (W. Zhang <i>et al.</i> , 2020) | unclear  | 6 trials /<br>6 loc. | yes | yes |     | yes     |   | yes     |      | yes  |      |    |
|    |   | reuse;<br>dataset was<br>generated by<br>one of the<br>authors |                      |     |     |     |         |   |         |      |      |      |    |
| 32 | Enhancing hybrid prediction in pearl millet using genomic and/or multi-environment phenotypic information of inbreds (Jarquin <i>et al.</i> , 2020b)        |  | 4 env.               | yes | yes |     |         |   |         |      |      |      |    |
| 33 | Assessment of ecological stability in yield for breeding of spring barley cultivars with increased adaptive potential (Hudzenko <i>et al.</i> , 2020)       | unclear  | 7 trials /<br>3 env. | yes | yes | yes | general |   | general |      | yes  | yes  |    |
|    |   |  |                      |     |     |     |         |   |         |      |      |      |    |
| 34 | Performance evaluation and yield stability of upland rice ( <i>Oryza sativa</i> L.) varieties in Ethiopia (Zewdu <i>et al.</i> , 2020)                      | unclear  | 4 trials /<br>4 loc. | yes | yes | yes | general |   | general | yes  | yes  | yes  |    |
| 35 | Environmental modeling of interaction variance for grain yield for medium early maturity maize hybrids (Mitrović <i>et al.</i> , 2020)                      | unclear;<br>reuse of<br>weather data                           | 8 trials /<br>8 loc. | yes | yes |     | general |   | yes     | yes  |      |      |    |

Table 6.1 (multiple pages)

| #  | Title   | reuse?            | T/L/E                  | S/Y | L   | C   | ED       | T   | OV  | Soil | Rain | Temp | CV  |
|----|---|-------------------|------------------------|-----|-----|-----|----------|-----|-----|------|------|------|-----|
| 36 | Variance component estimations and mega-environments for sweetpotato breeding in West Africa (Swanckaert et al., 2020)  | unclear           | 16 trials /            | yes | yes | yes | general  |     |     | yes  | yes  |      | yes |
|    |   |                   | 9 loc. /<br>3 datasets |     |     |     |          |     |     |      |      |      |     |
| 37 | Genotype x environment interaction of some traits in sunflower ( <i>Helianthus Annuus</i> L.) lines (Radić et al., 2020)  | original datasets | 6 trials /             | yes | yes | yes | general  |     | yes | yes  | yes  |      | yes |
|    |   |                   | 2 loc.                 |     |     |     |          |     |     |      |      |      |     |
| 38 | Analysis of genotype-environment interaction in fennel using Sudoku design (Al-Mehemdi et al., 2020)  | original datasets | 2 trials /             | yes | yes | yes | detailed |     |     |      |      |      |     |
|    |   |                   | 2 loc.                 |     |     |     |          |     |     |      |      |      |     |
| 39 | Delineating Genotype x Environment interactions towards durable resistance in mungbean against Cercospora leaf spot ( <i>Cercospora canescens</i> ) using GGE biplot (Das et al., 2020) | original datasets | 12 trials /            | yes | yes | yes | general  | yes | yes | yes  | yes  |      | yes |
|    |   |                   | 6 loc.                 |     |     |     |          |     |     |      |      |      |     |

**Table 6.1:** A review of 2020 publications presenting analyses of multi-environment trials investigating GxE interactions, as retrieved from the Web of Science with the query: TS=((("multi-environment" OR "multi-local") AND ("genotype by environment" OR "genotype x environment" OR "GxE" ) AND PY= ("2020" ). Column header abbreviations: reuse? - indicates whether it is explicitly stated that the datasets were reused; T/L/E - number of trials / locations / environments; S/Y - indicates whether season or year information is mentioned; L - location details; C - indicates whether the cultivars / accessions / genotypes used are described; ED - indicates whether the experimental design is described; T - indicates whether treatments are described; OV - indicates whether the observed variables are described; soil - indicates whether there is soil information; rain - indicates whether there is rain information; Temp - indicates whether temperature information is given. A \* indicates that the information may be in a supplementary material which is unavailable through the publication's DOI page. A "yes" in a cell means that some information about that aspect of the experiment is provided, though there is no requirement as to the level of detail or completeness.





# Supplementary Materials



## Chapter 2: Enabling reusability of plant phenomic datasets with MIAPPE 1.1

**Table S2.1:** Detailed mapping between MIAPPE, ISA-Tab and BrAPI fields.

| Mapping MIAPPE-BrAPI |                           |                      |                    |                    |
|----------------------|---------------------------|----------------------|--------------------|--------------------|
| MIAPPE               |                           | BrAPI                |                    |                    |
| line #               | MIAPPE Check list         | BrAPI Call           | BrAPI Object(s)    | BrAPI Field(s)     |
| DM-1                 | Investigation             |                      |                    |                    |
| DM-2                 | Investigation unique ID   | /trials/{trialDbld}  | None               | trialDbld          |
| DM-3                 | Investigation title       | /trials/{trialDbld}  | None               | trialName          |
| DM-4                 | Investigation description | /trials/{trialDbld}  | None               | trialDescription   |
| DM-5                 | Submission date           | /trials/{trialDbld}  | datasetAuthorships | submission-Date    |
| DM-6                 | Public release date       | /trials/{trialDbld}  | datasetAuthorships | publicRelease-Date |
| DM-7                 | License                   | /trials/{trialDbld}  | datasetAuthorships | license            |
| DM-8                 | MIAPPE version            |                      | out of scope       |                    |
| DM-9                 | Associated publication    | /trials/{trialDbld}  | publications       | publicationPUI     |
| DM-10                | Study                     |                      |                    |                    |
| DM-11                | Study unique ID           | /studies/{studyDbld} | None               | studyDbld          |
| DM-12                | Study title               | /studies/{studyDbld} | None               | studyName          |



| line # | MIAPPE                                 | BrAPI                |                    |                               |
|--------|--|----------------------|--------------------|-------------------------------|
|        | MIAPPE Check list                      | BrAPI Call           | BrAPI Object(s)    | BrAPI Field(s)                |
| DM-13  | Study description                      | /studies/{studyDbld} | None               | studyDescription              |
| DM-14  | Start date of study                    | /studies/{studyDbld} | None               | startDate                     |
| DM-15  | End date of study                      | /studies/{studyDbld} | None               | endDate                       |
| DM-16  | Contact institution                    | /studies/{studyDbld} | None               | instituteName                 |
| DM-17  | Geographic location (country)          | /studies/{studyDbld} | location           | countryName / countryCode     |
| DM-18  | Experimental site name                 | /studies/{studyDbld} | location           | name                          |
| DM-19  | Geographic location (latitude)         | /studies/{studyDbld} | location           | latitude                      |
| DM-20  | Geographic location (longitude)        | /studies/{studyDbld} | location           | longitude                     |
| DM-21  | Geographic location (altitude)         | /studies/{studyDbld} | location           | altitude                      |
| DM-22  | Description of the experimental design | /studies/{studyDbld} | experimentalDesign | description                   |
| DM-23  | Type of experimental design            | /studies/{studyDbld} | experimentalDesign | PUI                           |
| DM-24  | Observation unit level hierarchy       | /studies/{studyDbld} | additionalInfo     | observationUnitLevelHierarchy |

| line # | MIAPPE                         | BrAPI                |                 |                             |
|--------|--------------------------------|----------------------|-----------------|-----------------------------|
|        | MIAPPE Check list              | BrAPI Call           | BrAPI Object(s) | BrAPI Field(s)              |
| DM-25  | Observation unit description   | /studies/{studyDbId} | None            | observationUnitsDescription |
| DM-26  | Description of growth facility | /studies/{studyDbId} | growthFacility  | description                 |
| DM-27  | Type of growth facility        | /studies/{studyDbId} | growthFacility  | PUI                         |
| DM-28  | Cultural practices             | /studies/{studyDbId} | None            | culturalPractices           |
| DM-29  | Map of experimental design     | /studies/{studyDbId} | additionalInfo  | mapOfExperimentalDesign     |
| DM-30  | Person                         |                      |                 |                             |
| DM-31  | Person name                    | /studies/{studyDbId} | contacts        | name                        |
| DM-32  | Person email                   | /studies/{studyDbId} | contacts        | email                       |
| DM-33  | Person ID                      | /studies/{studyDbId} | contacts        | orcid / contactDbId         |
| DM-34  | Person role                    | /studies/{studyDbId} | contacts        | type                        |
| DM-35  | Person affiliation             | /studies/{studyDbId} | contacts        | instituteName               |
| DM-36  | Data File                      |                      |                 |                             |
| DM-37  | Data file link                 | /studies/{studyDbId} | dataLinks       | type                        |
| DM-38  | Data file description          | /studies/{studyDbId} | dataLinks       | name / url                  |
| DM-39  | Data file version              | /studies/{studyDbId} | dataLinks       | version                     |

| line # | MIAPPE   | BrAPI                           |                 |                         |
|--------|--|---------------------------------|-----------------|-------------------------|
|        | MIAPPE Check list  | BrAPI Call                      | BrAPI Object(s) | BrAPI Field(s)          |
| DM-40  | Biological Material  |                                 |                 |                         |
| DM-41  | Biological material ID   | /germplasm/{germplasmDbld}      | None            | accessionNumber         |
| DM-42  | Organism   | /germplasm/{germplasmDbld}      | taxonIds        | sourceName, taxonId     |
| DM-43  | Genus  | /germplasm/{germplasmDbld}      | None            | germplasm-Genus         |
| DM-44  | Species  | /germplasm/{germplasmDbld}      | None            | germplasm-Species       |
| DM-44' | Infraspecific name   | /germplasm/{germplasmDbld}      | None            | subtaxa                 |
| DM-45  | Biological material latitude                                   | /germplasm/{germplasmDbld}      | germplasmOrigin | latitudeDecimal         |
| DM-46  | Biological material longitude                                  | /germplasm/{germplasmDbld}      | germplasmOrigin | longitudeDecimal        |
| DM-47  | Biological material altitude                                   | /germplasm/{germplasmDbld}      | germplasmOrigin | altitude                |
| DM-48  | Biological material coordinates uncertainty                    | /germplasm/{germplasmDbld}      | germplasmOrigin | coordinate Uncertainty  |
| DM-49  | Biological material preprocessing                              | /germplasm/{germplasmDbld}      | None            | germplasm Preprocessing |
| DM-50  | Material source ID (Holding institute/stock centre, accession) | /germplasm/{germplasmDbld}/mcpd | donorInfo       | donorAccession Number   |

| line # | MIAPPE                                  | BrAPI                                   |                                |                        |
|--------|---|---|--------------------------------|------------------------|
|        | MIAPPE Check list                       | BrAPI Call                              | BrAPI Object(s)                | BrAPI Field(s)         |
| DM-51  | Material source DOI                     | /germplasm/<br>{germplasmDbld}<br>/mcpd | donorInfo                      | donorAccessionPui      |
| DM-52  | Material source latitude                | /germplasm/<br>{germplasmDbld}<br>/mcpd | collecting-Info.collectingSite | latitudeDecimal        |
| DM-53  | Material source longitude               | /germplasm/<br>{germplasmDbld}<br>/mcpd | collecting-Info.collectingSite | longitudeDecimal       |
| DM-54  | Material source altitude                | /germplasm/<br>{germplasmDbld}<br>/mcpd | collecting-Info.collectingSite | elevation              |
| DM-55  | Material source coordinates uncertainty | /germplasm/<br>{germplasmDbld}<br>/mcpd | collecting-Info.collectingSite | coordinate Uncertainty |
| DM-56  | Material source description             | /germplasm/<br>{germplasmDbld}          | None                           | seedSource Description |
| DM-57  | Environment                             |   |                                |                        |
| DM-58  | Environment parameter                   | /studies/{studyDbld}                    | environmentParameters          | parameter-Name         |
| DM-59  | Environment parameter value             | /studies/{studyDbld}                    | environmentParameters          | description            |
| DM-60  | Experimental Factor                     |   |                                |                        |
| DM-61  | Experimental Factor type                | /observationunits                       | Treatment                      | Factor                 |
| DM-62  | Experimental Factor description         | None                                    | None                           | None                   |

| line # | MIAPPE                            | BrAPI             |                      |                                 |
|--------|-----------------------------------|-------------------|----------------------|---------------------------------|
|        | MIAPPE Check list                 | BrAPI Call        | BrAPI Object(s)      | BrAPI Field(s)                  |
| DM-63  | Experimental Factor values        |                   | out of scope         |                                 |
| DM-64  | Event                             |                   |                      |                                 |
| DM-65  | Event type                        | /events           | None                 | eventType-Name                  |
| DM-66  | Event accession number            | /events           | None                 | eventTypeDbld                   |
| DM-67  | Event description                 | /events           | None                 | description                     |
| DM-68  | Event date                        | /events           | None                 | date                            |
| DM-69  | Observation Unit                  |                   |                      |                                 |
| DM-70  | Observation unit ID               | /observationunits | None                 | observation-unitDbld            |
| DM-71  | Observation unit type             | /observationunits | None                 | observation-Level               |
| DM-72  | External ID                       | /observationunits | observation-unitXref | id/source                       |
| DM-73  | Spatial distribution              | /observationunits | None                 | observation-Levels              |
| DM-74  | Observation Unit factor value     | /observationunits | treatments           | factor/modality                 |
| DM-75  | Sample                            |                   |                      |                                 |
| DM-76  | Sample ID                         | /samples          | None                 | sampleDbld                      |
| DM-77  | Plant structure development stage | /samples          | additionalInfo       | plantStructure DevelopmentStage |

| line # | MIAPPE                             | BrAPI      |                 |  |
|--------|------------------------------------|------------|-----------------|--|
|        | MIAPPE Check list                  | BrAPI Call | BrAPI Object(s) | BrAPI Field(s)                                     |
| DM-78  | Plant anatomical entity            | /samples   | None            | tissueType   |
| DM-79  | Sample description                 | /samples   | additionalInfo  | samplingDescription                                |
| DM-80  | Collection date                    | /samples   | None            | sampleTimestamp                                    |
| DM-81  | External ID                        | /samples   | additionalInfo  | externalId   |
| DM-82  | Observed Variable                  |            |                 |  |
| DM-83  | Variable ID                        | /variables | None            | observationVariable Name                           |
| DM-84  | Variable name                      | /variables | None            | observationVariable Name, observationVariable DbId |
| DM-85  | Variable accession number          | /variables | None            | xref, (observationVariableDbId)                    |
| DM-86  | Trait                              | /variables | trait           | traitName, description                             |
| DM-87  | Trait accession number             | /variables | trait           | (traitDbId)  |
| DM-88  | Method                             | /variables | method          | methodName   |
| DM-89  | Method accession number            | /variables | method          | (methodDbId)                                       |
| DM-90  | Method description                 | /variables | method          | description  |
| DM-91  | Reference associated to the method | /variables | method          | reference  |

| MIAPPE |                        | BrAPI      |                 |                |
|--------|------------------------|------------|-----------------|----------------|
| line # | MIAPPE Check list      | BrAPI Call | BrAPI Object(s) | BrAPI Field(s) |
| DM-92  | Scale                  | /variables | scale           | scaleName      |
| DM-93  | Scale accession number | /variables | scale           | (scaleDbld)    |
| DM-94  | Time scale             | None       | None            | None           |

### Mapping MIAPPE-ISA-Tab

| MIAPPE |                           | ISA-Tab       |  |                                   |
|--------|---------------------------|---------------|--|-----------------------------------|
| line # | MIAPPE Check list         | ISA-Tab File  | ISA-Tab Section (for Investigation file) | ISA-Tab Field                     |
| DM-1   | Investigation             |               |  |                                   |
| DM-2   | Investigation unique ID   | Investigation | INVESTIGATION                            | Investigation Identifier          |
| DM-3   | Investigation title       | Investigation | INVESTIGATION                            | Investigation Title               |
| DM-4   | Investigation description | Investigation | INVESTIGATION                            | Investigation Description         |
| DM-5   | Submission date           | Investigation | INVESTIGATION                            | Investigation Submission Date     |
| DM-6   | Public release date       | Investigation | INVESTIGATION                            | Investigation Public Release Date |
| DM-7   | License                   | Investigation | INVESTIGATION                            | Comment[License]                  |
| DM-8   | MIAPPE version            | Investigation | INVESTIGATION                            | Comment[MIAPPE version]           |

| MIAPPE |                                 | ISA-Tab       |  |                                    |
|--------|---------------------------------|---------------|--|------------------------------------|
| line # | MIAPPE Check list               | ISA-Tab File  | ISA-Tab Section (for Investigation file) | ISA-Tab Field                      |
| DM-9   | Associated publication          | Investigation | INVESTIGATION PUBLICATIONS               | Investigation Publication DOI      |
| DM-10  | Study                           |               |  |                                    |
| DM-11  | Study unique ID                 | Investigation | STUDY                                    | Study Identifier                   |
| DM-12  | Study title                     | Investigation | STUDY                                    | Study Title                        |
| DM-13  | Study description               | Investigation | STUDY                                    | Study Description                  |
| DM-14  | Start date of study             | Investigation | STUDY                                    | Comment[Study Start Date]          |
| DM-15  | End date of study               | Investigation | STUDY                                    | Comment[Study End Date]            |
| DM-16  | Contact institution             | Investigation | STUDY                                    | Comment[Study Contact Institution] |
| DM-17  | Geographic location (country)   | Investigation | STUDY                                    | Comment[Study Country]             |
| DM-18  | Experimental site name          | Investigation | STUDY                                    | Comment[Study Experimental Site]   |
| DM-19  | Geographic location (latitude)  | Investigation | STUDY                                    | Comment[Study Latitude]            |
| DM-20  | Geographic location (longitude) | Investigation | STUDY                                    | Comment[Study Longitude]           |
| DM-21  | Geographic location (altitude)  | Investigation | STUDY                                    | Comment[Study Altitude]            |



| MIAPPE |  | ISA-Tab       |  |   |
|--------|--|---------------|--|---|
| line # | MIAPPE Check list                      | ISA-Tab File  | ISA-Tab Section (for Investigation file) | ISA-Tab Field   |
| DM-22  | Description of the experimental design | Investigation | STUDY DESIGN DESCRIPTORS                 | Comment[Study Design Description]   |
| DM-23  | Type of experimental design            | Investigation | STUDY DESIGN DESCRIPTORS                 | Study Design Type   |
| DM-24  | Observation unit level hierarchy       | Investigation | STUDY DESIGN DESCRIPTORS                 | Comment[Observation Unit Level Hierarchy]   |
| DM-25  | Observation unit description           | Investigation | STUDY DESIGN DESCRIPTORS                 | Comment[Observation Unit Description]   |
| DM-26  | Description of growth facility         | Investigation | STUDY DESIGN DESCRIPTORS                 | Comment[Description of Growth Facility]   |
| DM-27  | Type of growth facility                | Investigation | STUDY DESIGN DESCRIPTORS                 | Comment[Type of Growth Facility]  |
| DM-28  | Cultural practices                     | Investigation | STUDY PROTOCOLS                          | Study Protocol Description (for Growth protocol)  |
| DM-29  | Map of experimental design             | Investigation | STUDY DESIGN DESCRIPTORS                 | Comment[Map of Experimental Design]   |
| DM-30  | Person                                 |               |  |   |
| DM-31  | Person name                            | Investigation | INVESTIGATION CONTACTS / STUDY CONTACTS  | Investigation Person Last Name - First Name - Mid Initials / Study Person Last Name - First Name - Mid Initials |

| MIAPPE |                        | ISA-Tab       |  |   |
|--------|------------------------|---------------|--|---|
| line # | MIAPPE Check list      | ISA-Tab File  | ISA-Tab Section (for Investigation file) | ISA-Tab Field   |
| DM-32  | Person email           | Investigation | INVESTIGATION CONTACTS / STUDY CONTACTS  | Investigation Person Email / Study Person Email             |
| DM-33  | Person ID              | Investigation | INVESTIGATION CONTACTS / STUDY CONTACTS  | Comment[Person ID]  |
| DM-34  | Person role            | Investigation | INVESTIGATION CONTACTS / STUDY CONTACTS  | Investigation Person Roles / Study Person Roles             |
| DM-35  | Person affiliation     | Investigation | INVESTIGATION CONTACTS / STUDY CONTACTS  | Investigation Person Affiliation / Study Person Affiliation |
| DM-36  | Data File              |               |  |   |
| DM-37  | Data file link         | Investigation | STUDY                                    | Comment[Study Data File Link]                               |
| DM-38  | Data file description  | Investigation | STUDY                                    | Comment[Study Data File Description]                        |
| DM-39  | Data file version      | Investigation | STUDY                                    | Comment[Study Data File Version]                            |
| DM-40  | Biological Material    |               |  |   |
| DM-41  | Biological material ID | Study         | None                                     | Source Name   |

| MIAPPE |  | ISA-Tab      |  |  |
|--------|--|--------------|--|--|
| line # | MIAPPE Check list  | ISA-Tab File | ISA-Tab Section (for Investigation file) | ISA-Tab Field  |
| DM-42  | Organism   | Study        | Source                                   | Characteristics[Organism]                                    |
| DM-43  | Genus  | Study        | Source                                   | Characteristics[Genus]                                       |
| DM-44  | Species  | Study        | Source                                   | Characteristics[Species]                                     |
| DM-44' | Infraspecific name   | Study        | Source                                   | Characteristics[Infraspecific Name]                          |
| DM-45  | Biological material latitude                                   | Study        | Source                                   | Characteristics[Biological Material Latitude]                |
| DM-46  | Biological material longitude                                  | Study        | Source                                   | Characteristics[Biological Material Longitude]               |
| DM-47  | Biological material altitude                                   | Study        | Source                                   | Characteristics[Biological Material Altitude]                |
| DM-48  | Biological material coordinates uncertainty                    | Study        | Source                                   | Characteristics[Biological Material Coordinates Uncertainty] |
| DM-49  | Biological material pre-processing                             | Study        | Source                                   | Characteristics[Biological Material Preprocessing]           |
| DM-50  | Material source ID (Holding institute/stock centre, accession) | Study        | Source                                   | Characteristics[Material Source ID]                          |

| MIAPPE |   | ISA-Tab       |  |  |
|--------|---|---------------|--|--|
| line # | MIAPPE Check list                       | ISA-Tab File  | ISA-Tab Section (for Investigation file) | ISA-Tab Field  |
| DM-51  | Material source DOI                     | Study         | Source                                   | Characteristics[Material Source DOI]                     |
| DM-52  | Material source latitude                | Study         | Source                                   | Characteristics[Material Source Latitude]                |
| DM-53  | Material source longitude               | Study         | Source                                   | Characteristics[Material Source Longitude]               |
| DM-54  | Material source altitude                | Study         | Source                                   | Characteristics[Material Source Altitude]                |
| DM-55  | Material source coordinates uncertainty | Study         | Source                                   | Characteristics[Material Source Coordinates Uncertainty] |
| DM-56  | Material source description             | Study         | Source                                   | Characteristics[Material Source Description]             |
| DM-57  | Environment                             |               |  |  |
| DM-58  | Environment parameter                   | Investigation | STUDY PROTOCOLS                          | Study Protocol Parameters Name (for Growth protocol)     |
| DM-59  | Environment parameter value             | Study         | Growth protocol                          | Parameter Value[ ]                                       |
| DM-60  | Experimental Factor                     |               |  |  |
| DM-61  | Experimental Factor type                | Investigation | STUDY FACTORS                            | Study Factor Name  |

| MIAPPE |                                 | ISA-Tab       |  |   |
|--------|---------------------------------|---------------|--|---|
| line # | MIAPPE Check list               | ISA-Tab File  | ISA-Tab Section (for Investigation file) | ISA-Tab Field   |
| DM-62  | Experimental Factor description | Investigation | STUDY FACTORS                            | Comment[Study Factor Description]                       |
| DM-63  | Experimental Factor values      | Investigation | STUDY FACTORS                            | Comment[Study Factor Values]                            |
| DM-64  | Event                           |               |  |   |
| DM-65  | Event type                      | Investigation | STUDY PROTOCOLS                          | Study Protocol Name (for protocol of type Event)        |
| DM-66  | Event accession number          | Investigation | STUDY PROTOCOLS                          | Study Protocol URI (for protocol of type Event)         |
| DM-67  | Event description               | Investigation | STUDY PROTOCOLS                          | Study Protocol Description (for protocol of type Event) |
| DM-68  | Event date                      | Event file    | None                                     | Event Date  |
| DM-69  | Observation Unit                |               |  |   |
| DM-70  | Observation unit ID             | Study / Assay | None                                     | Sample Name   |
| DM-71  | Observation unit type           | Study         | Sample                                   | Characteristics[Observation Unit Type]                  |
| DM-72  | External ID                     | Study         | Sample                                   | Characteristics[External ID]                            |
| DM-73  | Spatial distribution            | Study         | Sample                                   | Characteristics[Spatial distribution]                   |

| MIAPPE |                                   | ISA-Tab               |  |  |
|--------|-----------------------------------|-----------------------|--|--|
| line # | MIAPPE Check list                 | ISA-Tab File          | ISA-Tab Section (for Investigation file) | ISA-Tab Field                                      |
| DM-74  | Observation Unit factor value     | Study                 | Source / Sample                          | Factor Value[ ]                                    |
| DM-75  | Sample                            |                       |  |  |
| DM-76  | Sample ID                         | Assay                 | None                                     | Extract Name                                       |
| DM-77  | Plant structure development stage | Assay                 | Extract                                  | Characteristics[Plant Structure Development Stage] |
| DM-78  | Plant anatomical entity           | Assay                 | Extract                                  | Characteristics[Plant Anatomical Entity]           |
| DM-79  | Sample description                | Assay                 | Sampling protocol                        | Parameter Value[Sampling Description]              |
| DM-80  | Collection date                   | Assay                 | Sampling protocol                        | Parameter Value[Sampling Date]                     |
| DM-81  | External ID                       | Assay                 | Extract                                  | Characteristics[External ID]                       |
| DM-82  | Observed Variable                 |                       |  |  |
| DM-83  | Variable ID                       | Trait Definition File | None                                     | Variable ID  |
| DM-84  | Variable name                     | Trait Definition File | None                                     | Variable name                                      |
| DM-85  | Variable accession number         | Trait Definition File | None                                     | Variable accession number                          |
| DM-86  | Trait                             | Trait Definition File | None                                     | Trait  |

| MIAPPE |                                    | ISA-Tab               |  |                                    |
|--------|------------------------------------|-----------------------|--|------------------------------------|
| line # | MIAPPE Check list                  | ISA-Tab File          | ISA-Tab Section (for Investigation file) | ISA-Tab Field                      |
| DM-87  | Trait accession number             | Trait Definition File | None                                     | Trait accession number             |
| DM-88  | Method                             | Trait Definition File | None                                     | Method                             |
| DM-89  | Method accession number            | Trait Definition File | None                                     | Method accession number            |
| DM-90  | Method description                 | Trait Definition File | None                                     | Method description                 |
| DM-91  | Reference associated to the method | Trait Definition File | None                                     | Reference associated to the method |
| DM-92  | Scale                              | Trait Definition File | None                                     | Scale                              |
| DM-93  | Scale accession number             | Trait Definition File | None                                     | Scale accession number             |
| DM-94  | Time scale                         | Trait Definition File | None                                     | Time scale                         |

## Supplementary Notes S2.1: Summaries of the datasets used to evaluate MIAPPE 1.1

All datasets and files mentioned in this supporting information file are listed and accessible via the accompanying supplementary dataset (Papoutsoglou *et al.*, 2020b).

### Cork oak dataset (iBET)

The cork oak dataset derives from a report by Inácio *et al.*, 2017 and focuses on the evaluation of cork quality traits of cork oak (*Quercus suber*) trees and the putative correlation between those traits and DNA methylation in living cork cells. This *investigation* includes three *studies* corresponding to three cork oak stands in different locations in Portugal, characterized by Costa *et al.*, 2016. In each *study/stand*, 8 to 10 trees were randomly chosen, and in total 27 trees were assessed. For each tree, 20 cork quality traits were evaluated after debarking by manually phenotyping the cork plank. All the *traits* were described using the Woody Plant Ontology (Michotey and Chaves, 2020; Pommier *et al.*, 2019a).

As is typically the case, the cork oak trees in this dataset are identified only by means of their geographical coordinates. In this dataset, the *material source* identification is not described because cork harvesting for industrial applications starts on trees that are over 40 years old, and the life history of trees that old is often lost or unknown.

The dataset is available from the PHENO BrAPI endpoint Chaves *et al.*, 2020a, and in the spreadsheet template in Chaves *et al.*, 2020b.

### Arabidopsis dataset (IPK)

The *Arabidopsis* dataset is the result of an investigation of movement and soil cover effects on plant growth in a high throughput plant phenotyping system which combines a growth chamber for controlled environmental conditions and the imaging chambers for non-invasive trait assessment (Junker *et al.*, 2015). *Arabidopsis thaliana* plants were grown with a large number of replicates and their growth and development was evaluated with respect to two factors: *i*) “moving vs. stationary” to assess if the movement of plants on the conveyor belt influences plant growth and *ii*) “covered vs. uncovered” to assess if soil covers influence plant growth. These special soil covers are used for reducing transpiration and to facilitate segmentation of plant pixels from the background during image analysis. The dataset is an update of a previously published version (Junker *et al.*, 2020) based on MIAPPE 1.0 (Arend *et al.*, 2016b).

The dataset was encoded in ISA-Tab, uploaded to the Plant Genomics and Phenomics repository (Arend *et al.*, 2016a) and is available at (Junker, 2020).

### Barley dataset (IPK)

The barley dataset is the result of an investigation about the phenotypic assessment of growth and coloration dynamics as well as photosynthetic efficiency parameters in barley (*Hordeum vulgare*) HvASL (*Hordeum vulgare* alboatrians-like) mutants and wildtype plants (M. Li *et al.*, 2019). Barley HvASL mutants and wildtype plants were grown in a high throughput plant phenotyping facility for small plants. Seedlings of 9 different



genotypes (7 mutants and 2 wild types) were grown for 15 days in single-plant setups in the automated phytochamber and imaged daily using RGB and static fluorescence imaging. Automated image analysis routines were employed for the extraction of growth-related features and coloration dynamics. Additionally, three times per week seedlings were subjected to kinetic chlorophyll fluorescence imaging and photosynthetic efficiency and quenching parameters were assessed in the light-adapted state as well as during induction of photosynthesis after transition from dark to light. The dataset was encoded in ISA-Tab and is available as Junker and M. Li, 2020.

## Wheat dataset (GnpIS)

The wheat dataset is a subset of the Oury *et al.*, 2018 dataset, focusing on 38 measures related to the quality of the grain for bread making on 10 wheat (*Triticum aestivum*) varieties. It includes 80 studies conducted from 2000-2014 over 8 experimental field locations in France to study the impact of nitrogen nutrition on several traits of interest in wheat production. In this dataset, each study represents one location over one year, as the *biological material* changes each year. All the *variables* were measured from a single sample of grains harvested from each variety and described using the Wheat INRA Phenotyping Ontology (WIPO) (Pommier *et al.*, 2019a). An *experimental factor* (named “itk”) is used to discriminate between nitrogen “treated”, “low nitrogen” input or “none”. All *biological material* is identified using accession numbers generated with the French Small Grain Cereals Genbank (Small Grain Genetic Resource Centre, 2020).

The wheat dataset is provided as an ISA-Tab archive at Oury *et al.*, 2020c and via the GnpIS BrAPI endpoint (Oury *et al.*, 2020b).

## Poplar dataset (GnpIS)

The poplar dataset is the result of the investigation detailed in Monclus *et al.*, 2012 studying the variation of traits related to phenology, growth and water use efficiency in a full sib family of 360 poplar (*Populus trichocarpa* and *deltoides* crosses that produce the *Populus x generosa species*) individuals in three different locations. It corresponds to a test of clonal material (cuttings) derived from the same *material source* in orchards on three experimental sites in Europe over two years. The *material source* is the genbank accession and each *biological material* in each experimental site aggregates several individuals. The dataset has been organized in three two-year *studies*, which share the same *material source*. They include ten *observed variables* related to plant phenology, growth and water use efficiency, which are all included in the reference Woody Plant Ontology. *Ad hoc* variables were created in the dataset to refer to the measure of a given Woody Plant Ontology variable in different years as described in Pommier *et al.*, 2019b. In this dataset, the trees were not identified by geographical coordinates, but by unique identifiers. The poplar dataset is provided as an ISA-Tab archive at Michotey *et al.*, 2020b, and via the GnpIS BrAPI endpoint (Michotey *et al.*, 2020a).

## Maize datasets (VIB)

Three different maize (*Zea mays*) datasets were provided by VIB.

The first experiment details the assessment of variation in 103 lines of the maize B73xH99 recombinant inbred line (RIL) population in 13 studies in controlled growth chambers, for a set of primarily leaf size traits, complemented with measurements capturing growth dynamics, and cellular measurements (Baute *et al.*, 2015). In the second experiment, 1,636 MAGIC maize RILs were derived from eight genetically diverse founder lines (Dell'Acqua *et al.*, 2015). 529 of those lines were characterized, and a number of traits (ear height, plant height, pollen shedding and transformed grain yield) were determined in two fields. Finally, the third experiment describes the in-depth phenotyping of the fourth leaf at later stages of development in 197 RILs of two different maize populations (Baute *et al.*, 2016). As a follow-up to the previous two experiments, the traits from the former were selected for assessment, which was conducted on the multiparent MAGIC population of the latter.

All three datasets are available via the VIB BrAPI endpoint (Baute *et al.*, 2019a,c; Pea *et al.*, 2019a), and in ISA-Tab format (produced via BrAPI2ISA) at (Baute *et al.*, 2019b), (Pea *et al.*, 2019b) and Baute *et al.*, 2019d.

## Chapter 4: Using the MIAPPE standard to improve reusability of plant phenotyping data: Lessons learned from reusing multi-location potato field trial data

### Supplementary Notes S4.1: Finding relevant phenotypic datasets on the FDP

For this, the user has to navigate to the FDP of their institute. For this, they only need to know the address of the FDP. For this proof of concept, everything is hosted locally, so we have <http://localhost:3131/FDP>. In this case, this page looks like the **Figure S4.2**. Note the presence of a list of catalogs that this FDP holds. In this case, since we are looking for the CxE phenotypic datasets, we navigate to it.

The catalog itself (illustrated on **Figure S4.3**) includes a number of datasets. No further information is given on this page about them, so they have to be manually checked. The one we are interested in turns out to be dataset 1, the metadata for which is shown on **Figure S4.4**. It should be noted that the FDP specification contains no recommendations about the description of datasets. Therefore, to provide essential information about the contents of this phenotypic dataset (biological materials, observed variables, etc.), we have supplemented it with MIAPPE metadata (a schematic view is given in **Figure S4.5**). Because of this supplementation, if this FDP were to be indexed, information about the actual dataset contents could be harvested and enable content-related searches. The metadata page for Dataset 1 is composed of two parts: first, the metadata given in the FDP specification, for the dataset level; second, MIAPPE metadata.

Finally, following the link to the SPARQL distribution, we find the URL of a SPARQL endpoint hosting the dataset of interest (**Figure S4.6**). We can use this to explore it.

**Figure S4.2:** *The FDP metadata, including the address to the catalog of interest (in the red box).*

```

localhost:3131/FDP

@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.
@prefix dcterms: <http://purl.org/dc/terms/>.
@prefix r3d: <http://www.re3data.org/schema/3-0#>.
@prefix foaf: <http://xmlns.com/foaf/0.1/>.
@prefix fdp: <http://rdf.biosemantics.org/ontologies/fdp-o#>.
@prefix dcat: <http://www.w3.org/ns/dcat#>.
@prefix lang: <http://id.loc.gov/vocabulary/iso639-1/>.
@prefix xsd: <http://www.w3.org/2001/XMLSchema#>.
@prefix owl: <http://www.w3.org/2002/07/owl#>.

<http://localhost:3131/FDP> dcterms:title "FAIR Data Point of the Plant Breeding Group, Wageningen UR"
  rdfs:label "FAIR Data Point of the Plant Breeding Group, Wageningen UR" ;
  r3d:institution <http://www.wur.nl/> ;
  dcterms:hasVersion "0.1 beta" ;
  a r3d:Repository ;
  r3d:repositoryIdentifier <http://localhost:3131/FDP#repositoryID> ;
  dcterms:license <http://rdflicense.appspot.com/rdflicense/cc-by-nc-nd3.0> ;
  fdp:metadataIssued "2017-11-06T12:11:00.000000+00:00"^^xsd:dateTime ;
  fdp:metadataModified "2017-11-06T12:11:00.000000+00:00"^^xsd:dateTime ;
  r3d:institutionCountry <http://lexvo.org/id/iso3166/NL> ;
  dcterms:language lang:en ;
  dcterms:description "FAIR Data Point of the Plant Breeding Group, Wageningen UR" ;
  fdp:metadataIdentifier <http://localhost:3131/FDP#metadataID> ;
  dcterms:conformsTo <http://rdf.biosemantics.org/FDP/shex/fdpMetadata> ;
  dcterms:publisher <http://orcid.org/0000-0002-4368-8058> ;
  r3d:dataCatalog <http://localhost:3131/FDP/catalog/phenotypic.ttl> ,
    <http://localhost:3131/FDP/catalog/genotypic.ttl> ,
    <http://localhost:3131/FDP/catalog/genomic.ttl> .

<http://localhost:3131/FDP#metadataID> dcterms:identifier "325c7498-4469-11e7-a919-92ebcb67fe33" ;
  rdf:type <http://purl.org/spar/datacite/ResourceIdentifier> .

<http://localhost:3131/FDP#repositoryID> dcterms:identifier "49b17ed6-4469-11e7-a919-92ebcb67fe33" ;
  rdf:type <http://purl.org/spar/datacite/Identifier> .

<http://www.wur.nl/> a foaf:Organization ;
  foaf:name "Wageningen UR".

<http://orcid.org/0000-0002-4368-8058> rdf:type foaf:Organization ;
  foaf:name "Richard Finkers".

```

**Figure S4.3:** The phenotypic catalog, holding links to the datasets it includes (in the red box).

```

← → ↻ 🏠 localhost:3131/FDP/catalog/phenotypic.ttl

@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.
@prefix dcterms: <http://purl.org/dc/terms/>.
@prefix dcat: <http://www.w3.org/ns/dcat#>.
@prefix foaf: <http://xmlns.com/foaf/0.1/>.
@prefix fdp: <http://rdf.biosemantics.org/ontologies/fdp-o#>.
@prefix lang: <http://id.loc.gov/vocabulary/iso639-1/>.
@prefix xsd: <http://www.w3.org/2001/XMLSchema#>.
@prefix owl: <http://www.w3.org/2002/07/owl#>.

<http://localhost:3131/FDP/catalog/phenotypic> a dcat:Catalog ;
  dcterms:hasVersion "0.1" ;
  rdf:label "Plant Breeding Phenotypic Catalog" ;
  dcterms:title "Plant Breeding Phenotypic Catalog" ;
  dcterms:modified "2018-11-06T12:12:00.000000+00:00"^^xsd:dateTime ;
  fdp:metadataIssued "2018-11-06T12:12:00.000000+00:00"^^xsd:dateTime ;
  fdp:metadataModified "2018-11-06T12:12:00.000000+00:00"^^xsd:dateTime ;
  dcat:themeTaxonomy <http://dbpedia.org/resource/Breeding> ;
  fdp:metadataIdentifier <http://localhost:3131/FDP/catalog/phenotypic#metadataID> ;
  dcterms:conformsTo <http://rdf.biosemantics.org/fdp/shex/catalogMetadata> ;
  dcterms:isPartOf <http://localhost:3131/FDP/> ;
  dcterms:language lang:en ;
  dcterms:publisher <http://orcid.org/0000-0002-4368-8058> ;
  dcat:dataset
    <http://localhost:3131/FDP/dataset/Dataset_1.ttl> ,
    <http://localhost:3131/FDP/dataset/Dataset_2.ttl> ,
    <http://localhost:3131/FDP/dataset/Dataset_3.ttl> .

<http://orcid.org/0000-0002-4368-8058> a foaf:Person ;
  <http://xmlns.com/foaf/0.1/name> "Richard Finkers" .

<http://localhost:3131/FDP/catalog/phenotypic#metadataID> dcterms:identifier "49b181f6-4469-11e7-a919-92ebcb67fe33" ;
  a <http://purl.org/spar/datacite/ResourceIdentifier> .

```

**Figure S4.4:** Metadata for dataset 1. The black frames indicate the FDP dataset metadata specification. Everything else (green frames) is from MIAPPE (incomplete) and has been added here to give an indication as to the specific contents of this dataset.

```

← → ↻ 🏠 localhost:3131/FDP/dataset/Dataset_1.ttl

@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.
@prefix dcterms: <http://purl.org/dc/terms/>.
@prefix dcat: <http://www.w3.org/ns/dcat#>.
@prefix foaf: <http://xmlns.com/foaf/0.1/>.
@prefix fdp: <http://rdf.biosemantics.org/ontologies/fdp-o#>.
@prefix lang: <http://id.loc.gov/vocabulary/iso639-1/>.
@prefix xsd: <http://www.w3.org/2001/XMLSchema#>.

@prefix ppeo: <http://purl.org/ppoe/PEO.owl#>.
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.
@prefix xml: <http://www.w3.org/XML/1998/namespace>.

<http://localhost:3131/FDP/dataset/Dataset_1.ttl> a dcat:dataset ;
  rdfs:label "CxE_Hurtado_data" ;
  dcterms:title "CxE_Hurtado_data" ;
  dcterms:hasVersion "0.1" ;
  dcterms:language lang:en ;
  dcat:theme <http://www.wikidata.org/entity/Q14947594> ;
  fdp:metadataIdentifier <http://localhost:3131/FDP/dataset/Dataset_1.ttl#metadataID> ;
  dcterms:modified "2018-11-06T12:13:00.000000+00:00"^^xsd:dateTime ;
  fdp:metadataIssued "2018-11-06T12:13:00.000000+00:00"^^xsd:dateTime ;
  fdp:metadataModified "2018-11-06T12:13:00.000000+00:00"^^xsd:dateTime ;
  dcterms:conformsTo <http://rdf.biosemantics.org/fdp/shex/datasetMetadata> ;
  dcterms:publisher <http://orcid.org/0000-0001-8209-1900> ;
  dcterms:isPartOf <http://localhost:3131/FDP/catalog/phenotypic> ;
  dcat:distribution
    <http://localhost:3131/FDP/distribution/Pheno_dataset_1.zip.ttl>,
    <http://localhost:3131/FDP/distribution/Pheno_dataset_1_sparql.ttl> .

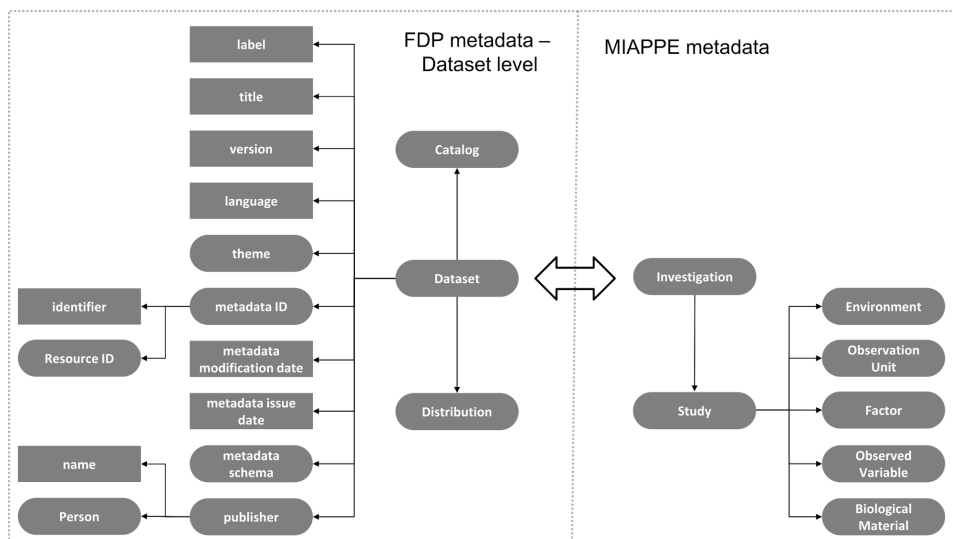
<http://localhost:3131/FDP/dataset/Dataset_1.ttl#metadataID> dcterms:identifier "49b1832c-4469-11e7-a919-92ebcb67fe33" ;
  a <http://purl.org/spar/datacite/ResourceIdentifier> .

<http://orcid.org/0000-0001-8209-1900> foaf:name "Elina Papoutsoglou" ;
  a foaf:Person .

<#investigation/WUR_inv_CE2020> a ppeo:investigation ;
  ppeo:hasAssociatedPublication "https://library.wur.nl/WebQuery/wurpubs/fulltext/240586" ;
  ppeo:hasDescription "FAIRified, partial data from the 2012 thesis of Paula Ximena Hurtado Lopez" ;
  ppeo:hasIdentifier "WUR_inv_CE2020" ;
  ppeo:hasLicense "CC-BY 4.0" ;
  ppeo:hasMIAPPEVersion "1.1"^^xsd:float ;

```

**Figure S4.5:** Illustration of the metadata specification Dataset level of the FDP and of MIAPPE (incomplete). The FDP only includes metadata about the resource, which is not sufficient for describing the content of the dataset. Therefore, it should be supplemented with content-oriented information about the experiments conducted that are described in the dataset. The dataset holds a MIAPPE Investigation, which makes for a good connection point.



**Figure S4.6:** The SPARQL distribution for the dataset of interest. The red box frames the URL of the queryable endpoint itself, which we can use in our scripts.

```

localhost:3131/FDP/distribution/Pheno_dataset_1_sparql.ttl

@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix dcat: <http://www.w3.org/ns/dcat#> .
@prefix dct: <http://purl.org/dc/terms/> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .

<http://localhost:3131/FDP/distribution/Pheno_dataset_1_sparql.ttl> a dcat:Distribution ;
  dct:title "SPARQL endpoint for the phenotypic datasets associated with P. Hurtado's thesis"@en;
  dct:identifier "SPARQL" ;
  dct:hasVersion "1.0" ;
  rdfs:label "Queryable endpoint for data associated with P. Hurtado's thesis"@en;
  dcat:downloadURL <http://localhost:3030/dataset.html?tab=query&ds=/pheno>;
  dct:license <http://rdflicense.appspot.com/rdflicense/cc-by-nc-nd3.0>;
  dct:issued "2018-11-06T12:13:00.000000+00:00"^^xsd:dateTime ;
  dct:modified "2018-11-06T12:13:00.000000+00:00"^^xsd:dateTime ;
  dcat:mediaType "json" .
  
```

## Chapter 5: Extracting knowledge networks from plant scientific literature: Potato tuber flesh color as an exemplary trait

**Table S5.1:** *The confusion matrix displaying the entity detection per article for the full training set of 34 articles.*

| Document ID                        | Total entities per article | True positives | False positives | False negatives |
|------------------------------------|----------------------------|----------------|-----------------|-----------------|
| 10.1007/s10142-008-0083-x          | 188                        | 182            | 1               | 5               |
| 10.1186/1471-2164-11-158           | 280                        | 248            | 5               | 27              |
| 10.1046/j.1365-313X.1996.9050745.x | 126                        | 113            | 0               | 13              |
| 10.1111/j.1744-7348.2003.tb00284.x | 289                        | 275            | 2               | 12              |
| 10.1007/BF02872013                 | 148                        | 140            | 5               | 3               |
| 10.1007/s00122-007-0560-y          | 221                        | 220            | 0               | 1               |
| 10.1104/pp.110.158733              | 303                        | 294            | 5               | 4               |
| ISSN: 0016-6731                    | 38                         | 32             | 0               | 6               |
| 10.1093/jxb/erp394                 | 298                        | 278            | 9               | 11              |
| 10.1186/1471-2229-7-11             | 200                        | 179            | 2               | 19              |
| 10.1371/journal.pone.0000350       | 230                        | 225            | 1               | 4               |
| 10.1093/jxb/eri016                 | 293                        | 288            | 3               | 2               |
| 10.1006/mben.2002.0234             | 253                        | 249            | 1               | 3               |
| 10.1111/j.1399-3054.2007.01016.x   | 257                        | 187            | 44              | 26              |
| 10.1093/jxb/erh121                 | 292                        | 283            | 6               | 3               |
| 10.1111/j.1365-3040.2011.02301.x   | 375                        | 355            | 14              | 6               |
| 10.21273/JASHS.136.4.265           | 275                        | 244            | 19              | 12              |
| 10.17221/460/2013-PSE              | 128                        | 122            | 1               | 5               |
| 10.1016/j.foodchem.2005.11.002     | 143                        | 99             | 0               | 44              |
| 10.1016/j.foodchem.2012.11.114     | 359                        | 232            | 3               | 124             |
| 10.17221/265/2011-PSE              | 222                        | 144            | 5               | 73              |

| Document ID                      | Total entities per article | True positives | False positives | False negatives |
|----------------------------------|----------------------------|----------------|-----------------|-----------------|
| 10.1016/j.foodchem.2014.08.011   | 257                        | 206            | 10              | 41              |
| 10.1016/j.jfca.2013.07.001       | 271                        | 247            | 2               | 22              |
| 10.17221/49/2010-PSE             | 168                        | 126            | 3               | 39              |
| 10.21273/JASHS.118.1.145         | 195                        | 142            | 3               | 50              |
| 10.1111/j.1439-0523.2008.01420.x | 139                        | 134            | 0               | 5               |
| 10.1007/s00122-009-1024-3        | 99                         | 66             | 6               | 27              |
| 10.1007/BF02853712               | 23                         | 20             | 0               | 3               |
| 10.1016/j.foodchem.2006.09.033   | 220                        | 132            | 3               | 85              |
| 10.21273/JASHS.126.6.722         | 222                        | 181            | 2               | 39              |
| 10.1186/1471-2229-6-13           | 272                        | 218            | 0               | 54              |
| 10.1007/BF02986245               | 274                        | 266            | 1               | 7               |
| 10.1007/s00122-014-2349-0        | 340                        | 309            | 1               | 30              |
| 10.1007/s12230-012-9250-7        | 152                        | 136            | 1               | 15              |
| <b>Total:</b>                    | 7550                       | 6572           | 158             | 820             |
| <hr/>                            |                            |                |                 |                 |
| <b>Precision:</b>                | 0.976523031                |                |                 |                 |
| <b>Recall:</b>                   | 0.889069264                |                |                 |                 |
| <b>F1:</b>                       | 0.930746353                |                |                 |                 |



**Table S5.2:** Summary table of the single-year difference in connections between flesh color and its eventual neighbours. It shows the degrees of separation between each flesh color node, and the nodes that eventually became its direct neighbours.

|  | 2009→2010                           | flesh color-like nodes |             |             |                   |                   | min |                     |
|--|-------------------------------------|------------------------|-------------|-------------|-------------------|-------------------|-----|---------------------|
|  |                                     | flesh                  | flesh color | tuber flesh | tuber flesh color | white flesh color |     | yellow-orange color |
| eventual direct neighbours to flesh color-like nodes | CCD                                 | 3→1                    | x→3         | 6→3         | x→3               | x→1               | x→2 | 3→1                 |
|  | CHY                                 | 2                      | x→1         | 5→3         | x→2               | x→3               | x→3 | 2→1                 |
|  | DXS                                 | 1                      | x→3         | 5→3         | x→3               | x→3               | x→3 | 1                   |
|  | PSY                                 | 1                      | x→3         | 5→3         | x→3               | x→3               | x→2 | 1                   |
|  | TP                                  | 3                      | x→5         | 7→4         | x→5               | x→4               | x→4 | 3                   |
|  | abscisic acid                       | 1                      | x→3         | 5→2         | x→3               | x→2               | x→3 | 1                   |
|  | aminocyclopropane-1-carboxylic acid | 1                      | x→4         | 5→4         | x→4               | x→3               | x→3 | 1                   |
|  | anthocyanin                         | 3                      | x→4         | 1           | x→5               | x→5               | x→5 | 1                   |
|  | b-carotene hydroxylase              | 2                      | x→1         | 5→3         | x→1               | x→3               | x→3 | 2→1                 |
|  | bHLH                                | 5→4                    | x→4         | 1           | x→5               | x→5               | x→5 | 1                   |
|  | carotenoid                          | 1                      | x→2         | 4→2         | x→2               | x→3               | x→2 | 1                   |
|  | chlorophyll                         | 1                      | x→3         | 5→3         | x→3               | x→3               | x→3 | 1                   |
|  | ethylene                            | 3                      | x→5         | 7→5         | x→5               | x→4               | x→1 | 3→1                 |
|  | flavonoid                           | 1                      | x→3         | 3           | x→3               | x→3               | x→3 | 1                   |
|  | flavonol                            | x                      | x           | x           | x                 | x                 | x   |                     |
|  | hydroxycinnamic acid                | 1                      | x→4         | 5→4         | x→4               | x→3               | x→4 | 1                   |
|  | lycopene                            | 2                      | x→3         | 5→3         | x→3               | x→2               | x→1 | 2→1                 |
|  | lycopene e-cyclase                  | 2                      | x→1         | 5→2         | x→3               | x→3               | x→3 | 2→1                 |
|  | phenolic                            | 2                      | x→3         | 4→3         | x→3               | x→4               | x→3 | 2                   |
|  | phenylalanine ammonia lyase         | x                      | x           | x           | x                 | x                 | x   |                     |
|  | zeaxanthin epoxidase                | 2                      | x→2         | 5→1         | x→3               | x→3               | x→3 | 2→1                 |

**Supplementary Notes S5.1:** Tracing of the critical connections between ZEP/BCH and flesh color, in 2007 and 2009, as mentioned in the Results section.

Based on 5.2, this section elaborates on the specific source of the connections between ZEP, BCH and color nodes.

In 2007, Diretto *et al.*, 2007b wrote:

- *"Silencing of beta-carotene hydroxylase increases total carotenoid and beta-carotene levels in potato tubers" (paper title)*

In this case, a new connection has been drawn between b-carotene hydroxylase and carotenoid; carotenoid has been a long time neighbour of flesh.

- *"Changes in endogenous gene expression were extensive and partially overlapping with those of LCY-e silenced tubers: CrtISO, LCY-b and ZEP were induced in both cases, indicating that they may respond to the balance between individual carotenoid species."*

Here a new connection is drawn between ZEP and carotenoid.

In 2010, similarly, Kloosterman *et al.*, 2010 and Wolters *et al.*, 2010 stated, respectively:

- *"Elevated expression level of a dominant allele of the beta-carotene hydroxylase (bch) gene was associated with yellow flesh color through mapping of the gene under a major QTL for flesh color on chromosome 3." and "The identified candidate genes for tuber flesh color (bch) and cooking type (tlrp) can provide useful markers for breeding schemes in the future."*
- *"We observed that among eleven beta-carotene hydroxylase 2 (Chy2) alleles only one dominant allele has a major effect, changing white into yellow flesh colour", and "Analysis of zeaxanthin epoxidase (Zep) alleles showed that all (diploid) genotypes with orange tuber flesh were homozygous for one specific Zep allele."*

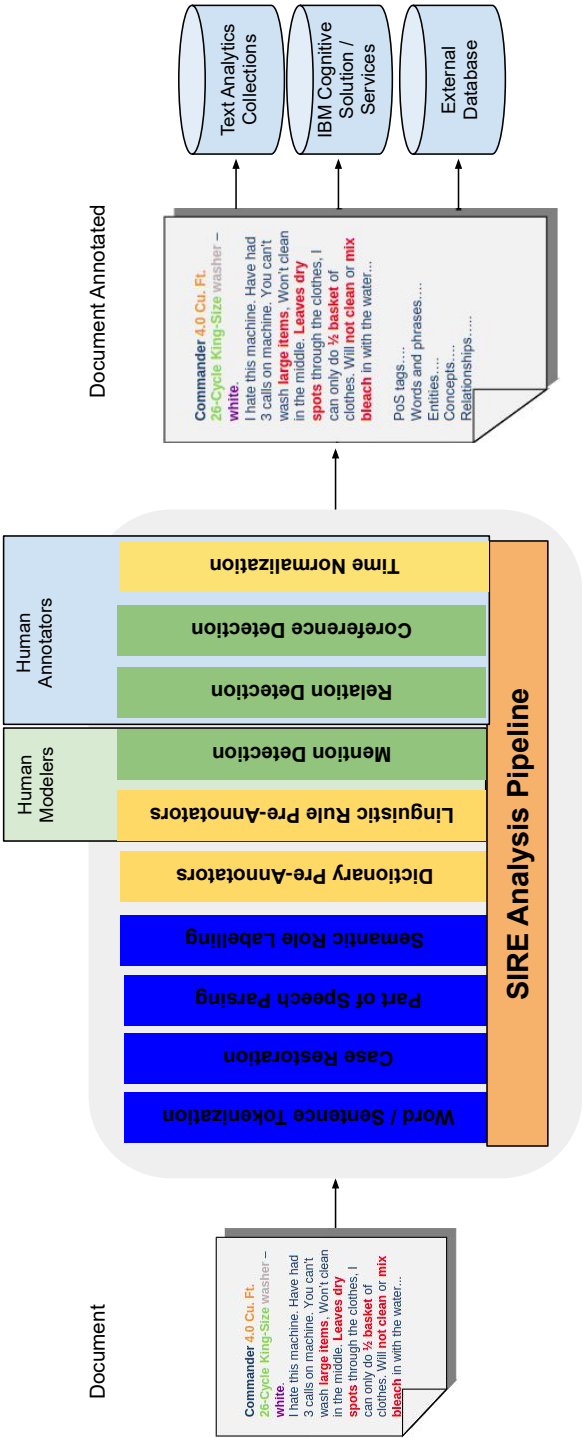
The above sentences enabled Watson to extract direct relationships from ZEP and BCH references to flesh color.

**Table S5.3:** *The list of 34 articles used in the training set.*

| ID (#) | Year of publication | Reference                            | DOI (other link if no DOI is available)   | Other ID                                |
|--------|---------------------|--------------------------------------|---|---|
| 2      | 2008                | Kloosterman <i>et al.</i> , 2008     | 10.1007/s10142-008-0083-x   | PMID: 18504629                          |
| 4      | 2010                | Kloosterman <i>et al.</i> , 2010     | 10.1186/1471-2164-11-158  | PMID: 20210995,<br>PMCID:<br>PMC2843620 |
| 6      | 1996                | Bachem <i>et al.</i> , 1996          | 10.1046/j.1365-313X.1996.9050745.x  | PMID: 8653120                           |
| 7      | 2003                | C. Celis-Gamboa <i>et al.</i> , 2003 | 10.1111/j.1744-7348.2003.tb00284.x  | AGR:IND43669370                         |
| 8      | 2006                | C. Brown <i>et al.</i> , 2006        | 10.1007/BF02872013  | AGR:IND43917866                         |
| 9      | 2007                | Werij <i>et al.</i> , 2007           | 10.1007/s00122-007-0560-y   | PMID: 17492422<br>PMCID:<br>PMC1913181  |
| 10     | 2010                | Campbell <i>et al.</i> , 2010        | 10.1104/pp.110.158733   | -                                       |
| 12     | 1988                | Bonierbale <i>et al.</i> , 1988      | <a href="https://www.genetics.org/content/120/4/1095">https://www.genetics.org/content/120/4/1095</a> | PMCID:<br>PMC1203572,<br>PMID: 17246486 |
| 13     | 2010                | Stushnoff <i>et al.</i> , 2010       | 10.1093/jxb/erp394  | PMID: 20110266,<br>PMCID:<br>PMC2826661 |
| 14     | 2007                | Diretto <i>et al.</i> , 2007b        | 10.1186/1471-2229-7-11  | PMID: 17335571,<br>PMCID:<br>PMC1828156 |
| 15     | 2007                | Diretto <i>et al.</i> , 2007a        | 10.1371/journal.pone.0000350  | PMID: 17406674,<br>PMCID:<br>PMC1831493 |
| 16     | 2005                | Ducreux <i>et al.</i> , 2005         | 10.1093/jxb/eri016  | PMID: 15533882                          |
| 17     | 2002                | Römer <i>et al.</i> , 2002           | 10.1006/mben.2002.0234  | PMID: 12646321                          |
| 18     | 2008                | N. Wang <i>et al.</i> , 2008         | 10.1111/j.1399-3054.2007.01016.x  | AGR:IND44012285                         |
| 19     | 2004                | W. Morris <i>et al.</i> , 2004       | 10.1093/jxb/erh121  | PMID: 15047766                          |
| 20     | 2011                | Zhou <i>et al.</i> , 2011            | 10.1111/j.1365-3040.2011.02301.x  | PMID: 21388418                          |
| 21     | 2011                | K. G. Haynes <i>et al.</i> , 2011    | 10.21273/JASHS.136.4.265  | -                                       |
| 24     | 2013                | Hamouz <i>et al.</i> , 2013          | 10.17221/460/2013-PSE   | -                                       |
| 25     | 2007                | Reyes and Cisneros-Zevallos, 2007    | 10.1016/j.foodchem.2005.11.002  | AGR:IND43869627                         |

| ID (#) | Year of publication | Reference                        | DOI (other link if no DOI is available)                 | Other ID                                |
|--------|---------------------|----------------------------------|---|---|
| 26     | 2013                | Lachman <i>et al.</i> , 2013     | 10.1016/j.foodchem.2012.11.114                          | PMID: 23411230                          |
| 27     | 2011                | Hamouz <i>et al.</i> , 2011      | 10.17221/265/2011-PSE                                   | -                                       |
| 28     | 2015                | Q. Wang <i>et al.</i> , 2015     | 10.1016/j.foodchem.2014.08.011                          | PMID: 25236223                          |
| 29     | 2013                | Hejtmánková <i>et al.</i> , 2013 | 10.1016/j.jfca.2013.07.001                              | AGR:IND601134175                        |
| 30     | 2010                | Hamouz <i>et al.</i> , 2010      | http://www.agriculturejournals.cz/publicFiles/25243.pdf | -                                       |
| 32     | 1993                | C. Brown <i>et al.</i> , 1993    | 10.21273/JASHS.118.1.145                                | -                                       |
| 33     | 2008                | Śliwka <i>et al.</i> , 2008      | 10.1111/j.1439-0523.2008.01420.x                        | AGR:IND44002450                         |
| 34     | 2009                | Y. Zhang <i>et al.</i> , 2009    | 10.1007/s00122-009-1024-3                               | PMID: 19363602,<br>PMCID:<br>PMC2690854 |
| 35     | 1991                | De Jong, 1991                    | 10.1007/BF02853712                                      | -                                       |
| 38     | 2007                | Teow <i>et al.</i> , 2007        | 10.1016/j.foodchem.2006.09.033                          | AGR:IND43886854                         |
| 39     | 2001                | W. Lu <i>et al.</i> , 2001       | 10.21273/JASHS.126.6.722                                | -                                       |
| 41     | 2006                | Diretto <i>et al.</i> , 2006     | 10.1186/1471-2229-6-13                                  | PMID: 17406674,<br>PMCID:<br>PMC1831493 |
| 42     | 2007                | Van Eck <i>et al.</i> , 2007     | 10.1007/BF02986245                                      | AGR:IND43959877                         |
| 43     | 2014                | Campbell <i>et al.</i> , 2014    | 10.1007/s00122-014-2349-0                               | PMID: 24965888                          |
| 44     | 2012                | P. McCord <i>et al.</i> , 2012   | 10.1007/s12230-012-9250-7                               | -                                       |

**Figure S5.2:** A schematic showing the Statistical Information and Relation Extraction (SIRE) pipeline used by Watson.





# References



- Abebe, A., G. Abera, and S. Beyene (2020). Sorption characteristics, growth and yield response of wheat (*Triticum aestivum* L.) to application of essential nutrients on nitisol and vertisol of Central Highland of Ethiopia. *African Journal of Plant Science* 14.3, pp. 108–120. DOI: 10.5897/AJPS2019.1873 (cited on page 85).
- Acharjee, A. (2013). *Systems biology and statistical data integration of ~omics data sets*. PhD thesis. Wageningen University & Research. URL: <https://edepot.wur.nl/258307> (cited on pages 12, 15).
- Acharjee, A., B. Kloosterman, R. G. Visser, and C. Maliepaard (2016). Integration of multi-omics data for prediction of phenotypic traits using random forest. *BMC Bioinformatics* 17.5, p. 180. DOI: 10.1186/s12859-016-1043-4 (cited on pages 97, 100, 103, 108).
- Acharjee, A., B. Kloosterman, R. C. de Vos, J. S. Werij, C. W. Bachem, R. G. Visser, and C. Maliepaard (2011). Data integration and network reconstruction with ~omics data using Random Forest regression in potato. *Analytica Chimica Acta* 705.1-2, pp. 56–63. DOI: 10.1016/j.aca.2011.03.050 (cited on pages 97, 100, 103, 105).
- Adnan, A., J. Diels, J. Jibrin, A. Kamara, A. Shaibu, P. Craufurd, and A. Menkir (2020). CERES-Maize model for simulating genotype-by-environment interaction of maize and its stability in the dry and wet savannas of Nigeria. *Field Crops Research* 253, p. 107826. DOI: 10.1016/j.fcr.2020.107826 (cited on page 130).
- Alercia, A., S. Diulgheroff, and M. Mackay (2015). FAO/Bioversity Multi-Crop Passport Descriptors V. 2.1 [MCPD V. 2.1]. *Bioversity International* 11 p. URL: <https://hdl.handle.net/10568/69166> (cited on page 36).
- Alves, R. S., M. D. V. de Resende, C. F. Azevedo, F. F. e Silva, A. C. P. Nunes, A. P. S. Carneiro, G. A. dos Santos, *et al.* (2020). Optimization of Eucalyptus breeding through random regression models allowing for reaction norms in response to environmental gradients. *Tree Genetics & Genomes* 16.2, pp. 1–8. DOI: 10.1007/s11295-020-01431-5 (cited on page 132).
- Anithakumari, A. M. (2011). *Genetic dissection of drought tolerance in potato*. PhD thesis. Wageningen University & Research. URL: <https://edepot.wur.nl/165211> (cited on pages 12, 15).
- Arend, D., A. Junker, U. Scholz, D. Schüler, J. Wylie, and M. Lange (2016a). PGP repository: a plant phenomics and genomics data publication infrastructure. *Database* Volume 2016, article ID baw033. DOI: 10.1093/database/baw033 (cited on page 154).
- Arend, D., M. Lange, J.-M. Pape, K. Weigelt-Fischer, F. Arana-Ceballos, I. Mücke, C. Klukas, T. Altmann, U. Scholz, and A. Junker (2016b). Quantitative monitoring of *Arabidopsis thaliana* growth and development using high-throughput plant phenotyping. *Scientific Data* 3, article ID 160055. DOI: 10.1038/sdata.2016.55 (cited on page 154).
- Arnaud, E., L. Cooper, R. Shrestha, N. Menda, R. T. Nelson, L. Matteis, M. Skofic, R. Bastow, P. Jaiswal, L. Mueller, *et al.* (2012). Towards a reference plant trait ontology for modeling knowledge of plant traits and phenotypes. *International Conference on Knowledge Engineering and Ontology Development* 2, pp. 220–225. DOI: 10.5220/0004138302200225 (cited on pages 21, 33).
- Atemezing, G., O. Corcho, D. Garijo, J. Mora, M. Poveda-Villalón, P. Rozas, D. Vila-Suero, and B. Villazón-Terrazas (2013). Transforming meteorological data into



- linked data. *Semantic Web* 4.3, pp. 285–290. DOI: 10.3233/SW-120089 (cited on pages 76, 88).
- Aubert, C., P. L. Buttigieg, M.-A. Laporte, M. Devare, and E. Arnaud (2017). *CGIAR Agronomy Ontology*. URL: <http://purl.obolibrary.org/obo/agro.owl> (visited on 03/03/2021) (cited on pages 34, 39).
- Bachem, C. W., R. S. Van Der Hoeven, S. M. De Bruijn, D. Vreugdenhil, M. Zabeau, and R. G. Visser (1996). Visualization of differential gene expression using a novel method of RNA fingerprinting based on AFLP: Analysis of gene expression during potato tuber development. *The Plant Journal* 9.5, pp. 745–753. DOI: 10.1046/j.1365-313X.1996.9050745.x (cited on page 165).
- Bairoch, A., R. Apweiler, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, *et al.* (2005). The universal protein resource (UniProt). *Nucleic Acids Research* 33.suppl\_1, pp. D154–D159. DOI: 10.1093/nar/gki070 (cited on page 19).
- Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature News* 533.7604, p. 452. DOI: 10.1038/533452a (cited on page 10).
- Baran, J., M. Gerner, M. Haeussler, G. Nenadic, and C. M. Bergman (2011). pubmed2ensembl: a resource for mining the biological literature on genes. *PLOS One* 6.9, e24716. DOI: 10.1371/journal.pone.0024716 (cited on page 98).
- Baute, J., J. D. Block, and D. Inzé (2019a). *Zea mays biparental RIL population - growth chamber phenotyping data (PIPPA BrAPI endpoint) [Data set]*. URL: <https://pippa.psb.ugent.be/BrAPIPPA/brapi/v1/trials/1> (cited on pages 44, 156).
- Baute, J., J. D. Block, and D. Inzé (Nov. 2019b). *Zea mays biparental RIL population - growth chamber phenotyping data [Data set]*. Zenodo. DOI: 10.5281/zenodo.3553692 (cited on pages 44, 156).
- Baute, J., J. D. Block, and D. Inzé (2019c). *Zea mays MAGIC RIL population - growth chamber phenotyping data (PIPPA BrAPI endpoint) [Data set]*. URL: <https://pippa.psb.ugent.be/BrAPIPPA/brapi/v1/trials/2> (cited on pages 44, 156).
- Baute, J., J. D. Block, and D. Inzé (Nov. 2019d). *Zea mays MAGIC RIL population - growth chamber phenotyping data [Data set]*. Zenodo. DOI: 10.5281/zenodo.3553768 (cited on pages 44, 156).
- Baute, J., D. Herman, F. Coppens, J. De Block, B. Slabbinck, M. Dell'Acqua, M. E. Pè, S. Maere, H. Nelissen, and D. Inzé (2015). Correlation analysis of the transcriptome of growing leaves with mature leaf parameters in a maize RIL population. *Genome Biology* 16, p. 168. DOI: 10.1186/s13059-015-0735-9 (cited on pages 45, 156).
- Baute, J., D. Herman, F. Coppens, J. De Block, B. Slabbinck, M. Dell'Acqua, M. E. Pè, S. Maere, H. Nelissen, and D. Inzé (2016). Combined Large-Scale Phenotyping and Transcriptomics in Maize Reveals a Robust Growth Regulatory Network. *Plant Physiology* 170 (3), pp. 1848–1867. DOI: 10.1104/pp.15.01883 (cited on pages 45, 156).
- Benson, D. A., I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, B. A. Rapp, and D. L. Wheeler (2000). GenBank. *Nucleic Acids Research* 28.1, pp. 15–18. DOI: 10.1093/nar/28.1.15 (cited on pages 11, 19, 34).

- Bernal-Vasquez, A.-M., A. Gordillo, M. Schmidt, and H.-P. Piepho (2017). Genomic prediction in early selection stages using multi-year data in a hybrid rye breeding program. *BMC Genetics* 18.1, pp. 1–17. DOI: 10.1186/s12863-017-0512-8 (cited on page 116).
- Beyan, O., A. Choudhury, J. van Soest, O. Kohlbacher, L. Zimmermann, H. Stenzhorn, M. R. Karim, M. Dumontier, S. Decker, L. O. B. da Silva Santos, *et al.* (2020). Distributed analytics on sensitive medical data: The Personal Health Train. *Data Intelligence* 2.1-2, pp. 96–107. DOI: 10.1162/dint\_a\_00032 (cited on page 22).
- Bolser, D., D. M. Staines, E. Pritchard, and P. Kersey (2016). Ensembl plants: integrating tools for visualizing, mining, and analyzing plant genomics data. *Plant Bioinformatics*, pp. 115–140. DOI: 10.1007/978-1-4939-3167-5\_6 (cited on pages 12, 19).
- Bonierbale, M. W., R. L. Plaisted, and S. D. Tanksley (1988). RFLP maps based on a common set of clones reveal modes of chromosomal evolution in potato and tomato. *Genetics* 120.4, pp. 1095–1103. URL: <https://www.genetics.org/content/120/4/1095> (cited on page 165).
- Bornmann, L. and R. Mutz (2015). Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology* 66.11, pp. 2215–2222. DOI: 10.1002/asi.23329 (cited on page 10).
- BrAPI2ISA contributors (2020). *BrAPI2ISA*. URL: <https://github.com/elixir-europe/plant-brapi-to-isa> (visited on 03/03/2021) (cited on page 42).
- Brazma, A., P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C. A. Ball, H. C. Causton, *et al.* (2001). Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. *Nature Genetics* 29.4, pp. 365–371. DOI: 10.1038/ng1201-365 (cited on pages 11, 33).
- Brazma, A., H. Parkinson, U. Sarkans, M. Shojatalab, J. Vilo, N. Abeygunawardena, E. Holloway, M. Kapushesky, P. Kemmeren, G. G. Lara, *et al.* (2003). ArrayExpress—a public repository for microarray gene expression data at the EBI. *Nucleic Acids Research* 31.1, pp. 68–71. DOI: 10.1093/nar/gkg091 (cited on pages 11, 19).
- BreedBase team (2020). *BreedBase*. URL: <https://breedbase.org/> (visited on 02/10/2020) (cited on page 33).
- Brown, C., C. Edwards, C.-P. Yang, and B. Dean (1993). Orange flesh trait in potato: Inheritance and carotenoid content. *Journal of the American Society for Horticultural Science* 118.1, pp. 145–150. DOI: 10.21273/JASHS.118.1.145 (cited on page 166).
- Brown, C., T. Kim, Z. Ganga, K. Haynes, D. De Jong, M. Jahn, I. Paran, and W. De Jong (2006). Segregation of total carotenoid in high level potato germplasm and its relationship to beta-carotene hydroxylase polymorphism. *American Journal of Potato Research* 83.5, pp. 365–372. DOI: 10.1007/BF02872013 (cited on pages 97, 165).
- Brown, D., I. Van den Bergh, S. de Bruin, L. Machida, and J. van Etten (2020). Data synthesis for crop variety evaluation. A review. *Agronomy for Sustainable Development* 40.4, pp. 1–20. DOI: 10.1007/s13593-020-00630-7 (cited on page 12).

- Bustin, S. A., V. Benes, J. A. Garson, J. Hellemans, J. Huggett, M. Kubista, R. Mueller, T. Nolan, M. W. Pfaffl, G. L. Shipley, *et al.* (2009). The MIQE Guidelines: *Minimum Information for Publication of Quantitative Real-Time PCR Experiments*. *Clinical Chemistry* 55 (4), pp. 611–622. DOI: 10.1373/clinchem.2008.112797 (cited on page 33).
- Buttigieg, P. L., N. Morrison, B. Smith, C. J. Mungall, and S. E. Lewis (2013). The environment ontology: contextualising biological and biomedical entities. *Journal of Biomedical Semantics* 4, p. 43. DOI: 10.1186/2041-1480-4-43 (cited on page 34).
- Cakmak, I. (2002). Plant nutrition research: Priorities to meet human needs for food in sustainable ways. *Plant and soil* 247.1, pp. 3–24. DOI: 10.1023/A:1021194511492 (cited on page 11).
- Campbell, R., L. J. Ducreux, W. L. Morris, J. A. Morris, J. C. Suttle, G. Ramsay, G. J. Bryan, P. E. Hedley, and M. A. Taylor (2010). The metabolic and developmental roles of carotenoid cleavage dioxygenase4 from potato. *Plant Physiology* 154.2, pp. 656–664. DOI: 10.1104/pp.110.158733 (cited on page 165).
- Campbell, R., S. D. Pont, J. A. Morris, G. McKenzie, S. K. Sharma, P. E. Hedley, G. Ramsay, G. J. Bryan, and M. A. Taylor (2014). Genome-wide QTL and bulked transcriptomic analysis reveals new candidate genes for the control of tuber carotenoid content in potato (*Solanum tuberosum* L.) *Theoretical and Applied Genetics* 127.9, pp. 1917–1933. DOI: 10.1007/s00122-014-2349-0 (cited on page 166).
- Carreño-Quintero, N. (2013). *Potato genetical genomics: investigating the genetic basis of primary metabolism and its relationship to the phenotype*. PhD thesis. Wageningen University & Research. URL: <https://edepot.wur.nl/285013> (cited on pages 12, 15).
- Celis-Gamboa, B. C. (2002). *The life cycle of the potato (Solanum tuberosum L.): from crop physiology to genetics*. PhD thesis. Wageningen University & Research. URL: <https://edepot.wur.nl/199013> (cited on pages 12, 15).
- Celis-Gamboa, C., P. Struik, E. Jacobsen, and R. Visser (2003). Temporal dynamics of tuber formation and related processes in a crossing population of potato (*Solanum tuberosum*). *Annals of Applied Biology* 143.2, pp. 175–186. DOI: 10.1111/j.1744-7348.2003.tb00284.x (cited on pages 73, 75, 165).
- Chattopadhyay, K., S. K. Mohanty, J. Vijayan, B. C. Marndi, A. Sarkar, K. A. Molla, K. Chakraborty, S. Ray, and R. K. Sarkar (2020). Genetic dissection of component traits for salinity tolerance at reproductive stage in rice. *Plant Molecular Biology Reporter*, pp. 1–17. DOI: 10.1007/s11105-020-01257-4 (cited on page 128).
- Chaves, I., C. M. Miguel, D. Faria, and B. V. Costa (2020a). *Enabling reusability of plant phenomic datasets with MIAPPE 1.1 - Supplementary dataset iBET (PHENO BrAPI endpoint) [Data set]*. URL: <https://brapi.biodata.pt/brapi/v1/trials/2> (cited on pages 44, 154).
- Chaves, I., C. M. Miguel, D. Faria, and B. V. Costa (2020b). *Enabling reusability of plant phenomic datasets with MIAPPE 1.1 - Supplementary dataset iBET [Data set]*. Version V2. Portail Data INRAE. DOI: 10.15454/AH6U4A (cited on pages 44, 154).

- Chen, Y., J. E. Argentinis, and G. Weber (2016). IBM Watson: How cognitive computing can be applied to big data challenges in life sciences research. *Clinical Therapeutics* 38.4, pp. 688–701. DOI: 10.1016/j.clinthera.2015.12.001 (cited on page 98).
- Chivenge, P., B. Vanlauwe, and J. Six (2011). Does the combined application of organic and mineral nutrient sources influence maize productivity? A meta-analysis. *Plant and Soil* 342.1, pp. 1–30. DOI: 10.1007/s11104-010-0626-5 (cited on page 116).
- Cho, H., W. Choi, and H. Lee (2017). A method for named entity normalization in biomedical articles: application to diseases and plants. *BMC Bioinformatics* 18.1, p. 451. DOI: 10.1186/s12859-017-1857-8 (cited on page 98).
- Choi, W., B. Kim, H. Cho, D. Lee, and H. Lee (2016). A corpus for plant-chemical relationships in the biomedical domain. *BMC Bioinformatics* 17.1, p. 386. DOI: 10.1186/s12859-016-1249-5 (cited on page 98).
- Clough, M. E., G. C. Yencho, B. Christ, W. DeJong, D. Halseth, K. Haynes, M. Henninger, C. Hutchinson, M. Kleinhenz, G. A. Porter, *et al.* (2010). An interactive online database for potato varieties evaluated in the eastern United States. *HortTechnology* 20.1, pp. 250–256. DOI: 10.21273/HORTTECH.20.1.250 (cited on page 19).
- Coelho, I. F., R. S. Alves, M. A. Peixoto, L. P. R. Teodoro, P. E. Teodoro, J. F. N. Pinto, E. F. dos Reis, L. L. Bhering, *et al.* (2020). Multi-trait multi-environment diallel analyses for maize breeding. *Euphytica* 216.9, pp. 1–17. DOI: 10.1007/s10681-020-02677-9 (cited on page 130).
- Collins, S., F. Genova, N. Harrower, S. Hodson, S. Jones, L. Laaksonen, D. Mitchen, R. Petrauskaitė, and P. Wittenburg (2018). Turning FAIR into reality: Final report and action plan from the European Commission expert group on FAIR data. URL: <https://hdl.handle.net/20.500.12259/103794> (cited on page 123).
- Cook, H. V. and L. J. Jensen (2019). A guide to dictionary-based text mining. *Bioinformatics and Drug Discovery*, pp. 73–89. DOI: 10.1007/978-1-4939-9089-4\_5 (cited on page 98).
- Cooper, L., A. Meier, M.-A. Laporte, J. L. Elser, C. Mungall, B. T. Sinn, D. Cavaliere, S. Carbon, N. A. Dunn, B. Smith, B. Qu, J. Preece, E. Zhang, S. Todorovic, G. Gkoutos, J. H. Doonan, D. W. Stevenson, E. Arnaud, and P. Jaiswal (2018). The Planteome database: an integrated resource for reference ontologies, plant genomics and phenomics. *Nucleic Acids Research* 46.D1, pp. D1168–D1180. DOI: 10.1093/nar/gkx1152 (cited on pages 33, 34, 54).
- Coppens, F., N. Wuyts, D. Inzé, and S. Dhondt (2017). Unlocking the potential of plant phenotyping data through integration and data-driven approaches. *Current Opinion in Systems Biology* 4, pp. 58–63. DOI: 10.1016/j.coisb.2017.07.002 (cited on pages 12, 73).
- Costa, A., I. Barbosa, C. Roussado, J. Graça, and H. Spiecker (2016). Climate response of cork growth in the Mediterranean oak (*Quercus suber* L.) woodlands of southwestern Portugal. *Dendrochronologia* 38, pp. 72–81. DOI: 10.1016/j.dendro.2016.03.007 (cited on page 154).
- Costa-Neto, G., R. Fritsche-Neto, and J. Crossa (2021). Nonlinear kernels, dominance, and envirotyping data increase the accuracy of genome-based prediction in multi-environment trials. *Heredity* 126.1, pp. 92–106. DOI: 10.1038/s41437-020-00353-1 (cited on page 130).

- Courtot, M., L. Cherubin, A. Faulconbridge, D. Vaughan, M. Green, D. Richardson, P. Harrison, P. L. Whetzel, H. Parkinson, and T. Burdett (2019). BioSamples database: an updated sample metadata hub. *Nucleic Acids Research* 47 (D1), pp. D1172–D1178. DOI: 10.1093/nar/gky1061 (cited on page 39).
- Curty, R. G., K. Crowston, A. Specht, B. W. Grant, and E. D. Dalton (2017). Attitudes and norms affecting scientists' data reuse. *PLOS One* 12.12, e0189288. DOI: 10.1371/journal.pone.0189288 (cited on page 11).
- Ćwiek-Kupczyńska, H. (2018). Striving for Semantics of Plant Phenotyping Data. *Semantics, Analytics, Visualization SAVE-SD 2017, SAVE-SD 2018*, pp. 161–169. DOI: 10.1007/978-3-030-01379-0\_12 (cited on page 32).
- Ćwiek-Kupczyńska, H., T. Altmann, D. Arend, E. Arnaud, D. Chen, G. Cornut, F. Fiorani, W. Frohmborg, A. Junker, C. Klukas, *et al.* (2016). Measures for interoperability of phenotypic data: minimum information requirements and formatting. *Plant Methods* 12.1, pp. 1–18. DOI: 10.1186/s13007-016-0144-4 (cited on pages 20, 33, 56, 87).
- Dalló, S. C., A. D. Zdziarski, L. G. Woyann, A. S. Milioli, R. Zanella, V. d. B. B. Batti, and G. Benin (2020). Key locations for soybean genotype assessment in South Brazil region. *Semina: Ciências Agrárias* 41.3, pp. 767–782. DOI: 10.5433/1679-0359.2020v41n3p767 (cited on page 132).
- Das, A., S. Gupta, A. K. Parihar, D. Singh, R. Chand, A. Pratap, K. D. Singha, and K. P. S. Kushwaha (2020). Delineating genotype×environment interactions towards durable resistance in mungbean against *Cercospora* leaf spot (*Cercospora canescens*) using GGE biplot. *Plant Breeding* 139.3, pp. 639–650. DOI: 10.1111/pbr.12789 (cited on page 134).
- DataCite Metadata Working Group (2014). DataCite Metadata Schema Documentation for the Publication and Citation of Research Data. *DataCite e.V. Version 3.1*. DOI: 10.5438/0010 (cited on page 36).
- De Jong, H. (1991). Inheritance of anthocyanin pigmentation in the cultivated potato: A critical review. *American Potato Journal* 68.9, pp. 585–593. DOI: 10.1007/BF02853712 (cited on page 166).
- Deist, T. M., F. J. Dankers, P. Ojha, M. S. Marshall, T. Janssen, C. Faivre-Finn, C. Masciocchi, V. Valentini, J. Wang, J. Chen, *et al.* (2020). Distributed learning on 20 000+ lung cancer patients—The Personal Health Train. *Radiotherapy and Oncology* 144, pp. 189–200. DOI: 10.1016/j.radonc.2019.11.019 (cited on page 23).
- Dell'Acqua, M., D. M. Gatti, G. Pea, F. Cattonaro, F. Coppens, G. Magris, A. L. Hlaing, H. H. Aung, H. Nelissen, J. Baute, *et al.* (2015). Genetic properties of the MAGIC maize population: a new platform for high definition QTL mapping in *Zea mays*. *Genome Biology* 16, p. 167. DOI: 10.1186/s13059-015-0716-z (cited on pages 45, 156).
- Dijk, A. D. J. van, G. Kootstra, W. Kruijer, and D. de Ridder (2020). Machine learning in plant science and plant breeding. *iScience*, p. 101890. DOI: 10.1016/j.isci.2020.101890 (cited on page 126).
- Ding, R., C. N. Arighi, J.-Y. Lee, C. H. Wu, and K. Vijay-Shanker (2015). pGenN, a gene normalization tool for plant genes and proteins in scientific literature. *PLOS One* 10.8, e0135305. DOI: 10.1371/journal.pone.0135305 (cited on page 98).

- Diouf, I., L. Derivot, S. Koussevitzky, Y. Carretero, F. Bitton, L. Moreau, and M. Causse (2020). Genetic basis of phenotypic plasticity and genotype  $\times$  environment interactions in a multi-parental tomato population. *Journal of Experimental Botany* 71.18, pp. 5365–5376. DOI: 10.1093/jxb/eraa265 (cited on page 129).
- Diretto, G., S. Al-Babili, R. Tavazza, V. Papacchioli, P. Beyer, and G. Giuliano (2007a). Metabolic engineering of potato carotenoid content through tuber-specific overexpression of a bacterial mini-pathway. *PLOS One* 2.4, e350. DOI: 10.1186/1471-2229-7-11 (cited on page 165).
- Diretto, G., R. Tavazza, R. Welsch, D. Pizzichini, F. Mourgues, V. Papacchioli, P. Beyer, and G. Giuliano (2006). Metabolic engineering of potato tuber carotenoids through tuber-specific silencing of lycopene epsilon cyclase. *BMC Plant Biology* 6.1, pp. 1–11. DOI: 10.1186/1471-2229-6-13 (cited on page 166).
- Diretto, G., R. Welsch, R. Tavazza, F. Mourgues, D. Pizzichini, P. Beyer, and G. Giuliano (2007b). Silencing of beta-carotene hydroxylase increases total carotenoid and beta-carotene levels in potato tubers. *BMC Plant Biology* 7.1, p. 11. DOI: 10.1186/1471-2229-7-11 (cited on pages 105, 164, 165).
- Doan, A., N. F. Noy, and A. Y. Halevy (2004). Introduction to the special issue on semantic integration. *ACM Sigmod Record* 33.4, pp. 11–13. DOI: 10.1145/1041410.1041412 (cited on page 54).
- Dowell, R. D., R. M. Jokerst, A. Day, S. R. Eddy, and L. Stein (2001). The distributed annotation system. *BMC Bioinformatics* 2.1, pp. 1–7. DOI: 10.1186/1471-2105-2-7 (cited on page 54).
- Ducreux, L. J., W. L. Morris, P. E. Hedley, T. Shepherd, H. V. Davies, S. Millam, and M. A. Taylor (2005). Metabolic engineering of high carotenoid potato tubers containing enhanced levels of  $\beta$ -carotene and lutein. *Journal of Experimental Botany* 56.409, pp. 81–89. DOI: 10.1093/jxb/eri016 (cited on page 165).
- EBI (2020). *European Variation Archive*. URL: <https://www.ebi.ac.uk/eva/> (visited on 03/03/2021) (cited on pages 12, 125).
- Eck, H. J. van (1995). *Localisation of morphological traits on the genetic map of potato using RFLP and isozyme markers*. PhD thesis. Wageningen University & Research. URL: <https://edepot.wur.nl/206450> (cited on pages 12, 15).
- Eck, H. J. van, J. R. van der Voort, J. Draaistra, P. van Zandvoort, E. van Enckevort, B. Segers, J. Peleman, E. Jacobsen, J. Helder, and J. Bakker (1995). The inheritance and chromosomal localization of AFLP markers in a non-inbred potato offspring. *Molecular Breeding* 1.4, pp. 397–410. DOI: 10.1007/BF01248417 (cited on page 15).
- Endara, L., H. Cui, and J. G. Burleigh (2018). Extraction of phenotypic traits from taxonomic descriptions for the tree of life using natural language processing. *Applications in Plant Sciences* 6.3, e1035. DOI: 10.1002/aps3.1035 (cited on page 98).
- Fahlgren, N., M. A. Gehan, and I. Baxter (2015). Lights, camera, action: high-throughput plant phenotyping is ready for a close-up. *Current Opinion in Plant Biology* 24, pp. 93–99. DOI: 10.1016/j.pbi.2015.02.006 (cited on page 12).
- FAIR-CxE contributors (2020). *FAIR-CxE Github repository*. URL: [https://github.com/PBR/FAIR\\_CxE](https://github.com/PBR/FAIR_CxE) (visited on 03/03/2021) (cited on pages 76, 77, 83, 89–91).

- FAIRsharing.org: MIAPPE (2020). *Minimum Information about Plant Phenotyping Experiment*. URL: <https://doi.org/10.25504/FAIRsharing.nd9ce9> (visited on 03/03/2021) (cited on page 48).
- Fang, F. C., R. G. Steen, and A. Casadevall (2012). Misconduct accounts for the majority of retracted scientific publications. *Proceedings of the National Academy of Sciences* 109.42, pp. 17028–17033. DOI: 10.1073/pnas.1212247109 (cited on page 116).
- FAO, IFAD, UNICEF, WFP, and WHO (2018). The state of food security and nutrition in the world 2018: building climate resilience for food security and nutrition. *Food & Agriculture Organization*. URL: <http://www.fao.org/3/I9553EN/i9553en.pdf> (cited on page 73).
- Federhen, S. (2012). The NCBI taxonomy database. *Nucleic Acids Research* 40.D1, pp. D136–D143. DOI: 10.1093/nar/gkr1178 (cited on pages 11, 19).
- Ferreira Coelho, I., M. A. Peixoto, J. Santana Pinto Coelho Evangelista, R. Silva Alves, S. Sales, M. D. V. d. Resende, J. F. Naves Pinto, E. Fialho dos Reis, and L. L. Bhering (2020). Multiple-trait, random regression, and compound symmetry models for analyzing multi-environment trials in maize breeding. *PLOS One* 15.11, e0242705. DOI: 10.1371/journal.pone.0242705 (cited on page 128).
- Ferrucci, D. A. (2012). Introduction to "This is Watson". *IBM Journal of Research and Development* 56.3.4, pp. 1–1. DOI: 10.1147/JRD.2012.2184356 (cited on page 98).
- Field, D., G. Garrity, T. Gray, N. Morrison, J. Selengut, P. Sterk, T. Tatusova, N. Thomson, M. J. Allen, S. V. Angiuoli, et al. (2008). The minimum information about a genome sequence MIMS specification. *Nature Biotechnology* 26, pp. 541–547. DOI: 10.1038/nbt1360 (cited on page 33).
- Fielding, R. T. and R. N. Taylor (2002). Principled design of the modern web architecture. *ACM Transactions on Internet Technology (TOIT)* 2.2, pp. 115–150. DOI: 10.1145/514183.514185 (cited on page 54).
- Finkers, R. (2018). "Farm Data Train". In: *Scientific Symposium FAIR Data Sciences for Green Life Sciences*, pp. 1–1. DOI: 10.18174/FAIRdata2018.16292 (cited on page 23).
- Flavell, R. B. (2017). Innovations continuously enhance crop breeding and demand new strategic planning. *Global Food Security* 12, pp. 15–21. DOI: 10.1016/j.gfs.2016.10.001 (cited on page 54).
- Florian, R., A. Ittycheriah, H. Jing, and T. Zhang (2003). Named entity recognition through classifier combination. *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pp. 168–171 (cited on page 109).
- Furbank, R. T. and M. Tester (2011). Phenomics—technologies to relieve the phenotyping bottleneck. *Trends in Plant Science* 16.12, pp. 635–644. DOI: 10.1016/j.tplants.2011.09.005 (cited on page 12).
- Galea, D., I. Laponogov, and K. Veselkov (2018). Exploiting and assessing multi-source data for supervised biomedical named entity recognition. *Bioinformatics* 34.14, pp. 2474–2482. DOI: 10.1093/bioinformatics/bty152 (cited on page 98).
- Getahun, B. B. (2017). *Genetic diversity of potato for nitrogen use efficiency under low input conditions in Ethiopia*. PhD thesis. Wageningen University & Research. DOI: 10.18174/420903 (cited on pages 12, 15).

- Ghouila, A., G. H. Siwo, J.-B. D. Entfellner, S. Panji, K. A. Button-Simons, S. Z. Davis, F. M. Fadlelmola, M. T. Ferdig, N. Mulder, T. Bensellak, *et al.* (2018). Hackathons as a means of accelerating scientific discoveries and knowledge transfer. *Genome Research* 28.5, pp. 759–765. DOI: 10.1101/gr.228460.117 (cited on pages 54, 60).
- Gibney, E. and R. Van Noorden (2013). Scientists losing data at a rapid rate. *Nature News*. DOI: 10.1038/nature.2013.14416 (cited on page 10).
- Giuliano, G. (2014). Plant carotenoids: genomics meets multi-gene engineering. *Current Opinion in Plant Biology* 19, pp. 111–117. DOI: 10.1016/j.pbi.2014.05.006 (cited on page 97).
- Goodman, A., A. Pepe, A. W. Blocker, C. L. Borgman, K. Cranmer, M. Crosas, R. Di Stefano, Y. Gil, P. Groth, M. Hedstrom, *et al.* (2014). Ten simple rules for the care and feeding of scientific data. *PLOS Computational Biology* 10.4, e1003542. DOI: 10.1371/journal.pcbi.1003542 (cited on page 14).
- Guo, J., S. Pradhan, D. Shahi, J. Khan, J. Mcbreen, G. Bai, J. P. Murphy, and M. A. Babar (2020). Increased prediction accuracy using combined genomic information and physiological traits in a soft wheat panel evaluated in multi-environments. *Scientific Reports* 10.1, pp. 1–12. DOI: 10.1038/s41598-020-63919-3 (cited on page 132).
- Hahn, U., K. B. Cohen, Y. Garten, and N. H. Shah (2012). Mining the pharmacogenomics literature—a survey of the state of the art. *Briefings in Bioinformatics* 13.4, pp. 460–494. DOI: 10.1093/bib/bbs018 (cited on page 98).
- Halewood, M., T. Chiurugwi, R. Sackville Hamilton, B. Kurtz, E. Marden, E. Welch, F. Michiels, J. Mozafari, M. Sabran, N. Patron, *et al.* (2018). Plant genetic resources for food and agriculture: opportunities and challenges emerging from the science and information technology revolution. *New Phytologist* 217.4, pp. 1407–1419. DOI: 10.1111/nph.14993 (cited on page 67).
- Hamouz, K., J. Lachman, K. Hejtmánková, K. Pazderu, M. Čížek, and P. Dvořák (2010). Effect of natural and growing conditions on the content of phenolics in potatoes with different flesh colour. *Plant, Soil and Environment* 56.8, pp. 368–374. URL: <http://www.agriculturejournals.cz/publicFiles/25243.pdf> (cited on page 166).
- Hamouz, K., J. Lachman, K. Pazderu, K. Hejtmankova, J. Cimr, J. Musilova, V. Pivec, M. Orsak, and A. Svobodova (2013). Effect of cultivar, location and method of cultivation on the content of chlorogenic acid in potatoes with different flesh colour. *Plant, Soil and Environment* 59.10, pp. 465–471. DOI: 10.17221/460/2013-PSE (cited on page 165).
- Hamouz, K., J. Lachman, K. Pazderu, J. Tomášek, K. Hejtmánková, and V. Pivec (2011). Differences in anthocyanin content and antioxidant activity of potato tubers with different flesh colour. *Plant, Soil and Environment* 57.10, pp. 478–485. DOI: 10.17221/265/2011-PSE (cited on page 166).
- Harmston, N., W. Filsell, and M. P. Stumpf (2010). What the papers say: Text mining for genomics and systems biology. *Human Genomics* 5.1, p. 17. DOI: 10.1186/1479-7364-5-1-17 (cited on page 97).
- Hassani-Pak, K., A. Singh, M. Brandizi, J. Hearnshaw, S. Amberkar, A. L. Phillips, J. H. Doonan, and C. Rawlings (2020). KnetMiner: a comprehensive approach for



- supporting evidence-based gene discovery and complex trait analysis across species. *bioRxiv*. DOI: 10.1101/2020.04.02.017004 (cited on page 121).
- Haug, K., R. M. Salek, P. Conesa, J. Hastings, P. De Matos, M. Rijnbeek, T. Mahendraker, M. Williams, S. Neumann, P. Rocca-Serra, *et al.* (2013). MetaboLights—an open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic Acids Research* 41.D1, pp. D781–D786. DOI: 10.1093/nar/gks1004 (cited on page 11).
- Haynes, K. G., B. A. Clevidence, D. Rao, and B. T. Vinyard (2011). Inheritance of carotenoid content in tetraploid  $\times$  diploid potato crosses. *Journal of the American Society for Horticultural Science* 136.4, pp. 265–272. DOI: 10.21273/JASHS.136.4.265 (cited on page 165).
- Hejtmánková, K., Z. Kotíková, K. Hamouz, V. Pivec, J. Vacek, and J. Lachman (2013). Influence of flesh colour, year and growing area on carotenoid and anthocyanin content in potato tubers. *Journal of Food Composition and Analysis* 32.1, pp. 20–27. DOI: 10.1016/j.jfca.2013.07.001 (cited on page 166).
- Heslot, N. D. (2014). Optimal use of phenotypic data for breeding using genomic predictions. *Cornell Theses and Dissertations*. URL: <https://hdl.handle.net/1813/36138> (cited on page 116).
- Hirsch, C. D., J. P. Hamilton, K. L. Childs, J. Cepela, E. Crisovan, B. Vaillancourt, C. N. Hirsch, M. Habermann, B. Neal, and C. R. Buell (2014). Spud DB: A resource for mining sequences, genotypes, and phenotypes to accelerate potato breeding. *The Plant Genome* 7.1, pp. 2013–12. DOI: 10.3835/plantgenome2013.12.0042 (cited on page 19).
- Hirschberg, J. and C. D. Manning (2015). Advances in natural language processing. *Science* 349.6245, pp. 261–266. DOI: 10.1126/science.aaa8685 (cited on page 97).
- Huala, E., A. W. Dickerman, M. Garcia-Hernandez, D. Weems, L. Reiser, F. LaFond, D. Hanley, D. Kiphart, M. Zhuang, W. Huang, *et al.* (2001). The Arabidopsis Information Resource (TAIR): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant. *Nucleic Acids Research* 29.1, pp. 102–105. DOI: 10.1093/nar/29.1.102 (cited on page 19).
- Huang, C.-C. and Z. Lu (2015). Community challenges in biomedical text mining over 10 years: success, failure and the future. *Briefings in Bioinformatics* 17.1, pp. 132–144. DOI: 10.1093/bib/bbv024 (cited on page 97).
- Hudzenko, V., O. Demydov, V. Kavunets, L. Kachan, V. Ishchenko, and M. Sardak (2020). Assessment of ecological stability in yield for breeding of spring barley cultivars with increased adaptive potential. *Regulatory Mechanisms in Biosystems* 11.3, pp. 425–430. DOI: 10.15421/022065 (cited on page 133).
- Hunt, C. H., B. J. Hayes, F. A. Van Eeuwijk, E. S. Mace, and D. R. Jordan (2020). Multi-environment analysis of sorghum breeding trials using additive and dominance genomic relationships. *Theoretical and Applied Genetics* 133.3, pp. 1009–1018. DOI: 10.1007/s00122-019-03526-7 (cited on page 132).
- Hurtado-Lopez, P. X. (2012). *Investigating genotype by environment and QTL by environment interactions for developmental traits in potato*. PhD thesis. Wageningen University & Research. URL: <https://edepot.wur.nl/240586> (cited on pages 12, 15, 25, 32, 73, 76, 83).

- Hurtado-Lopez, P. X., B. B. Tessema, S. K. Schnabel, C. Maliepaard, C. van der Linden, P. Eilers, J. Jansen, F. van Eeuwijk, and R. Visser (2015). Understanding the genetic basis of potato development using a multi-trait QTL analysis. *Euphytica* 204.1, pp. 229–241. DOI: 10.1007/s10681-015-1431-2 (cited on page 15).
- Inácio, V., P. M. Barros, A. Costa, C. Roussado, E. Gonçalves, R. Costa, J. Graça, M. M. Oliveira, and L. Morais-Cecílio (2017). Differential DNA methylation patterns are related to phellogen origin and quality of *Quercus suber* cork. *PLOS One* 12 (1), e0169018. DOI: 10.1371/journal.pone.0169018 (cited on pages 45, 154).
- ISA-Tab for plant phenotyping contributors (2020). *ISA-Tab for plant phenotyping*. URL: <https://github.com/MIAPPE/ISA-Tab-for-plant-phenotyping> (visited on 03/03/2021) (cited on page 41).
- Jacobs, J., H. Van Eck, P. Arens, B. Verkerk-Bakker, B. te Lintel Hekkert, H. Bastiaanssen, A. El-Kharbotly, A. Pereira, E. Jacobsen, and W. Stiekema (1995). A genetic map of potato (*Solanum tuberosum*) integrating molecular markers, including transposons, and classical markers. *Theoretical and Applied Genetics* 91.2, pp. 289–300. DOI: 10.1007/BF00220891 (cited on pages 12, 73).
- Jacobsen, A., R. Kaliyaperumal, L. O. B. da Silva Santos, B. Mons, E. Schultes, M. Roos, and M. Thompson (2020). A generic workflow for the data FAIRification process. *Data Intelligence* 2.1-2, pp. 56–65. DOI: 10.1162/dint\_a\_00028 (cited on page 22).
- Jaiswal, P., S. Avraham, K. Ilic, E. A. Kellogg, S. McCouch, A. Pujar, L. Reiser, S. Y. Rhee, M. M. Sachs, M. Schaeffer, et al. (2005). Plant Ontology (PO): a controlled vocabulary of plant structures and growth stages. *Comparative and Functional Genomics* 6.7-8, pp. 388–397. DOI: 10.1002/cfg.496 (cited on pages 21, 33).
- Jang, G., T. Lee, S. Hwang, C. Park, J. Ahn, S. Seo, Y. Hwang, and Y. Yoon (2018). PISTON: Predicting drug indications and side effects using topic modeling and natural language processing. *Journal of Biomedical Informatics* 87, pp. 96–107. DOI: 10.1016/j.jbi.2018.09.015 (cited on page 98).
- Jarquín, D., R. Howard, J. Crossa, Y. Beyene, M. Gowda, J. W. Martini, G. Covarrubias Pazaran, J. Burgueño, A. Pacheco, M. Grondona, et al. (2020a). Genomic prediction enhanced sparse testing for multi-environment trials. *G3: Genes, Genomes, Genetics* 10.8, pp. 2725–2739. DOI: 10.1534/g3.120.401349 (cited on page 130).
- Jarquín, D., R. Howard, Z. Liang, S. K. Gupta, J. C. Schnable, and J. Crossa (2020b). Enhancing hybrid prediction in pearl millet using genomic and/or multi-environment phenotypic information of inbreds. *Frontiers in Genetics* 10, p. 1294. DOI: 10.3389/fgene.2019.01294 (cited on page 133).
- Jiwuba, L., A. Danquah, I. Asante, E. Blay, J. Onyeka, E. Danquah, and C. Egesi (2020). Genotype by environment interaction on resistance to cassava green mite associated traits and effects on yield performance of cassava genotypes in Nigeria. *Frontiers in Plant Science* 11, p. 1344. DOI: 10.3389/fpls.2020.572200 (cited on page 129).
- Jongedijk, E. (1991). *Desynapsis and FDR 2N-megaspore formation in diploid potato: potentials and limitations for breeding and for the induction of diplosporic apomixis*. PhD thesis. Wageningen University & Research. URL: <https://edepot.wur.nl/440899> (cited on pages 12, 15).

- Jonquet, C., A. Toulet, E. Arnaud, S. Aubin, E. D. Yeumo, V. Emonet, J. Graybeal, M.-A. Laporte, M. A. Musen, V. Pesce, *et al.* (2018). AgroPortal: A vocabulary and ontology repository for agronomy. *Computers and Electronics in Agriculture* 144, pp. 126–143. DOI: 10.1016/j.compag.2017.10.012 (cited on page 34).
- Juliana, P., R. P. Singh, H.-J. Braun, J. Huerta-Espino, L. Crespo-Herrera, T. Payne, J. Poland, S. Shrestha, U. Kumar, A. K. Joshi, *et al.* (2020). Retrospective quantitative genetic analysis and genomic prediction of global wheat yields. *Frontiers in Plant Science* 11, p. 1328. DOI: 10.3389/fpls.2020.580136 (cited on page 129).
- Junker, A. (2020). *Raw images files from quantitative monitoring of 484 Arabidopsis thaliana plants using high-throughput plant phenotyping (MIAPPE 1.1 update) [Data set]*. e!DAL - Plant Genomics and Phenomics Research Data Repository (PGP), IPK Gatersleben, Seeland OT Gatersleben, Germany. DOI: 10.5447/IPK/2020/3 (cited on pages 44, 154).
- Junker, A. and M. Li (2020). *Phenotypic assessment of growth and coloration dynamics as well as photosynthetic efficiency parameters in barley HvASL mutants and wild type plants [Data set]*. e!DAL - Plant Genomics and Phenomics Research Data Repository (PGP), IPK Gatersleben, Seeland OT Gatersleben, Germany. DOI: 10.5447/IPK/2020/4 (cited on pages 44, 155).
- Junker, A., M. M. Muraya, K. Weigelt-Fischer, F. Arana-Ceballos, C. Klukas, A. E. Melchinger, R. C. Meyer, D. Riewe, and T. Altmann (2015). Optimizing experimental procedures for quantitative evaluation of crop plant performance in high throughput phenotyping systems. *Frontiers in Plant Science* 5, p. 770. DOI: 10.3389/fpls.2014.00770 (cited on pages 45, 154).
- Junker, A., K. Weigelt-Fischer, T. Altmann, and C. Klukas (2020). *Raw images files from quantitative monitoring of 484 Arabidopsis thaliana plants using high-throughput plant phenotyping*. URL: <https://doi.org/10.5447/IPK/2016/7> (visited on 03/03/2021) (cited on page 154).
- Kambhatla, N. (2004). Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, 22–es. DOI: 10.3115/1219044.1219066 (cited on page 109).
- Kanehisa, M., M. Furumichi, M. Tanabe, Y. Sato, and K. Morishima (2016). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research* 45.D1, pp. D353–D361. DOI: 10.1093/nar/gkw1092 (cited on page 111).
- Kanehisa, M., S. Goto, S. Kawashima, Y. Okuno, and M. Hattori (2004). The KEGG resource for deciphering the genome. *Nucleic Acids Research* 32.suppl\_1, pp. D277–D280. DOI: 10.1093/nar/gkh063 (cited on page 19).
- Khan, M., F. Mohammad, F. Khan, S. Ahmad, and I. Ullah (2020). Additive main effect and multiplicative interaction analysis for grain yield in bread wheat. *Journal of Animal and Plant Sciences* 30.3, pp. 677–684. DOI: 10.36899/JAPS.2020.3.0080 (cited on page 131).
- Kim, C., V. Zhu, J. Obeid, and L. Lenert (2019). Natural language processing and machine learning algorithm to identify brain MRI reports with acute ischemic stroke. *PLOS One* 14.2, e0212778. DOI: 10.1371/journal.pone.0212778 (cited on page 98).

- King, G. J., S. Amoah, and S. Kurup (2010). Exploring and exploiting epigenetic variation in crops. *Genome* 53.11, pp. 856–868. DOI: 10.1139/G10-059 (cited on page 32).
- Kloosterman, B. (2006). *Transcriptomic analysis of potato tuber development and tuber quality traits using microarray technology*. PhD thesis. Wageningen University & Research. URL: <https://edepot.wur.nl/121811> (cited on pages 12, 15).
- Kloosterman, B., J. A. Abelenda, M. d. M. C. Gomez, M. Oortwijn, J. M. de Boer, K. Kowitwanich, B. M. Horvath, H. J. van Eck, C. Smaczniak, S. Prat, *et al.* (2013). Naturally occurring allele diversity allows potato cultivation in northern latitudes. *Nature* 495.7440, pp. 246–250. DOI: 10.1038/nature11912 (cited on page 15).
- Kloosterman, B., D. De Koeyer, R. Griffiths, B. Flinn, B. Steuernagel, U. Scholz, S. Sonnewald, U. Sonnewald, G. J. Bryan, S. Prat, *et al.* (2008). Genes driving potato tuber initiation and growth: identification based on transcriptional changes using the POCI array. *Functional & Integrative Genomics* 8.4, pp. 329–340. DOI: 10.1007/s10142-008-0083-x (cited on page 165).
- Kloosterman, B., M. Oortwijn, T. America, R. de Vos, R. G. Visser, C. W. Bachem, *et al.* (2010). From QTL to candidate gene: genetical genomics of simple and complex traits in potato using a pooling strategy. *BMC Genomics* 11.1, pp. 1–16. DOI: 10.1186/1471-2164-11-158 (cited on pages 164, 165).
- Kluyver, T., B. Ragan-Kelley, F. Pérez, B. E. Granger, M. Bussonnier, J. Frederic, K. Kelley, J. B. Hamrick, J. Grout, S. Corlay, *et al.* (2016). *Jupyter Notebooks—a publishing format for reproducible computational workflows*. Vol. 2016 (cited on page 123).
- Krajewski, P., D. Chen, H. Ćwiek, A. D. van Dijk, F. Fiorani, P. Kersey, C. Klukas, M. Lange, A. Markiewicz, J. P. Nap, *et al.* (2015). Towards recommendations for metadata and data handling in plant phenotyping. *Journal of Experimental Botany* 66.18, pp. 5417–5427. DOI: 10.1093/jxb/erv271 (cited on pages 20, 33, 56, 87, 116).
- Krajewski, P. and H. Ćwiek-Kupczyńska (2020). *MIAPPE Impact*. URL: <https://www.miappe.org/publications/#impact> (visited on 03/03/2021) (cited on page 33).
- Kuzniar, A., R. Kaliyaperumal, C. Martinez-Ortiz, and C. Geng (Sept. 2020). *FAIR Data Point*. Zenodo. DOI: 10.5281/zenodo.4059590 (cited on pages 76, 88).
- Lachman, J., K. Hamouz, J. Musilová, K. Hejtmánková, Z. Kotířková, K. Pazderu, J. Domkářová, V. Pivec, and J. Cimr (2013). Effect of peeling and three cooking methods on the content of selected phytochemicals in potato tubers with various colour of flesh. *Food Chemistry* 138.2-3, pp. 1189–1197. DOI: 10.1016/j.foodchem.2012.11.114 (cited on page 166).
- Lamprecht, A.-L., L. Garcia, M. Kuzak, C. Martinez, R. Arcila, E. Martin Del Pico, V. Dominguez Del Angel, S. van de Sandt, J. Ison, P. A. Martinez, *et al.* (2020). Towards FAIR principles for research software. *Data Science* 3.1, pp. 37–59 (cited on page 123).
- Lapatas, V., M. Stefanidakis, R. C. Jimenez, A. Via, and M. V. Schneider (2015). Data integration in biological research: an overview. *Journal of Biological Research-Thessaloniki* 22.1, p. 9. DOI: 10.1186/s40709-015-0032-5 (cited on page 33).
- Lassila, O., R. R. Swick, *et al.* (1998). Resource description framework (RDF) model and syntax specification (cited on page 88).

- Lee, J. M., G. F. Davenport, D. Marshall, T. N. Ellis, M. J. Ambrose, J. Dicks, T. J. van Hintum, and A. J. Flavell (2005). GERMINATE. A generic database for integrating genotypic and phenotypic information for plant genetic resource collections. *Plant Physiology* 139.2, pp. 619–631. DOI: 10.1104/pp.105.065201 (cited on page 116).
- Leinonen, R., R. Akhtar, E. Birney, L. Bower, A. Cerdeno-Tárraga, Y. Cheng, I. Cleland, N. Faruque, N. Goodgame, R. Gibson, *et al.* (2010). The European nucleotide archive. *Nucleic Acids Research* 39.suppl\_1, pp. D28–D31. DOI: 10.1093/nar/gkq967 (cited on pages 12, 19, 125).
- Li, M., G. Hensel, M. Melzer, A. Junker, H. Tschiersch, D. Arend, J. Kumlehn, T. Börner, and N. Stein (2019). Mutation of the ALBOSTRIANS ohnologous gene HvCMF3 impairs chloroplast development and thylakoid architecture in barley due to reduced plastid translation. *bioRxiv*, p. 756833. DOI: 10.1101/756833 (cited on pages 45, 154).
- Lu, W., K. Haynes, E. Wiley, and B. Clevidence (2001). Carotenoid content and color in diploid potatoes. *Journal of the American Society for Horticultural Science* 126.6, pp. 722–726. DOI: 10.21273/JASHS.126.6.722 (cited on page 166).
- McCord, M. C., J. W. Murdock, and B. K. Boguraev (2012). Deep parsing in Watson. *IBM Journal of Research and Development* 56.3.4, pp. 3–1. DOI: 10.1147/JRD.2012.2185409 (cited on page 109).
- McCord, P., L. Zhang, and C. Brown (2012). The incidence and effect on total tuber carotenoids of a recessive zeaxanthin epoxidase allele (Zep1) in yellow-fleshed potatoes. *American Journal of Potato Research* 89.4, pp. 262–268. DOI: 10.1007/s12230-012-9250-7 (cited on page 166).
- Al-Mehemdi, A. F., M. M. Elsayhookie, and M. H. Al-Issawi (2020). Analysis of genotype-environment interaction in fennel using Sudoku design. *Asian Journal of Agriculture & Biology* 8.1, pp. 61–68. DOI: 10.35495/ajab.2019.07.314 (cited on page 134).
- Merkel, D. (2014). Docker: lightweight linux containers for consistent development and deployment. *Linux Journal* 2014.239, p. 2 (cited on page 123).
- Meyer, S., A. Nagel, and C. Gebhardt (2005). PoMaMo—a comprehensive database for potato genome data. *Nucleic Acids Research* 33.suppl\_1, pp. D666–D670. DOI: 10.1093/nar/gki018 (cited on page 19).
- MIAPPE contributors (2020a). *MIAPPE Github repository*. URL: <https://github.com/MIAPPE/MIAPPE> (visited on 03/03/2021) (cited on pages 35, 48).
- MIAPPE contributors (2020b). *MIAPPE v1.1 training spreadsheet*. URL: [https://github.com/MIAPPE/MIAPPE/tree/master/MIAPPE\\_Checklist-Data-Model-v1.1/MIAPPE\\_templates](https://github.com/MIAPPE/MIAPPE/tree/master/MIAPPE_Checklist-Data-Model-v1.1/MIAPPE_templates) (visited on 03/03/2021) (cited on page 42).
- Michotey, C. and I. Chaves (2020). *Woody Plant Ontology*. URL: [http://www.cropontology.org/ontology/C0\\_357/Woody%20Plant%20Ontology](http://www.cropontology.org/ontology/C0_357/Woody%20Plant%20Ontology) (visited on 03/03/2021) (cited on pages 44, 154).
- Michotey, C., I. Chaves, C. Anger, V. Jorge, F. Ehrenmann, F. Jean, and L. Opgenoorth (2019). Woody Plant Ontology. Version V1. *Portail Data INRAE*. DOI: 10.15454/JB2WCE (cited on page 44).
- Michotey, C., V. Jorge, and R. Monclus (2020a). *Integrating genome annotation and QTL position to identify candidate genes for productivity, architecture and water-use efficiency in Populus spp - Supplementary dataset (GnPLS BrAPI endpoint) [Data set]*. URL: <https://urgi.versailles.inra.fr/faidare/brapi/v1/trial>

- s/aHR0cDovL2R4LmRvaS5vcmcvMTAuMTE4Ni8xNDcxLTIyMjktMTItMTcz (cited on pages 44, 155).
- Michotey, C., V. Jorge, and R. Monclus (2020b). *Integrating genome annotation and QTL position to identify candidate genes for productivity, architecture and water-use efficiency in Populus spp - Supplementary dataset [Data set] [Data set]*. Version V1. Portail Data INRAE. DOI: 10.15454/EASUQV (cited on pages 44, 155).
- Millet, E. J., W. Kruijer, A. Coupel-Ledru, S. A. Prado, L. Cabrera-Bosquet, S. Lacube, A. Charcosset, C. Welcker, F. van Eeuwijk, and F. Tardieu (2019). Genomic prediction of maize yield across European environmental conditions. *Nature Genetics* 51.6, pp. 952–956. DOI: 10.1038/s41588-019-0414-y (cited on pages 12, 32).
- Milne, I., P. Shaw, G. Stephen, M. Bayer, L. Cardle, W. T. Thomas, A. J. Flavell, and D. Marshall (2010). Flapjack—graphical genotype visualization. *Bioinformatics* 26.24, pp. 3133–3134. DOI: 10.1093/bioinformatics/btq580 (cited on pages 55, 56).
- Mitrović, B., B. Drašković, D. Stanisavljević, M. Perišić, P. Čanak, I. Mitrović, and S. Tančić-Živanov (2020). Environmental modeling of interaction variance for grain yield of medium early maturity maize hybrids. *Genetika* 52.1, pp. 367–378. DOI: 10.2298/GENSR2001367M (cited on page 133).
- Mohammadi, R., B. Sadeghzadeh, M. M. Poursiahbidi, and M. M. Ahmadi (2020). Integrating univariate and multivariate statistical models to investigate genotype  $\times$  environment interaction in durum wheat. *Annals of Applied Biology*. DOI: 10.1111/aab.12648 (cited on page 128).
- Momotaz, A., R. W. Davidson, D. Zhao, P. McCord, H. S. Sandhu, M. Baltazar, M. S. Islam, and O. Coto Arbelo (2020). Genotype-by-environment interaction analysis across three crop cycles in sugarcane. *Journal of Crop Improvement*, pp. 1–15. DOI: 10.1080/15427528.2020.1817220 (cited on page 129).
- Monclus, R., J.-C. Leplé, C. Bastien, P.-F. Bert, M. Villar, N. Marron, F. Brignolas, and V. Jorge (2012). Integrating genome annotation and QTL position to identify candidate genes for productivity, architecture and water-use efficiency in *Populus spp*. *BMC Plant Biology* 12, p. 173. DOI: 10.1186/1471-2229-12-173 (cited on pages 45, 155).
- Morris, W., L. Ducreux, D. Griffiths, D. Stewart, H. Davies, and M. Taylor (2004). Carotenogenesis during tuber development and storage in potato. *Journal of Experimental Botany* 55.399, pp. 975–982. DOI: 10.1093/jxb/erh121 (cited on page 165).
- Mueller, L. A., T. H. Solow, N. Taylor, B. Skwarecki, R. Buels, J. Binns, C. Lin, M. H. Wright, R. Ahrens, Y. Wang, et al. (2005). The SOL Genomics Network. A comparative resource for *Solanaceae* biology and beyond. *Plant Physiology* 138.3, pp. 1310–1317. DOI: 10.1104/pp.105.060707 (cited on page 19).
- Mushayi, M., H. Shimelis, J. Derera, A. I. Shayanowako, and I. Mathew (2020). Multi-environmental evaluation of maize hybrids developed from tropical and temperate lines. *Euphytica* 216, pp. 1–14. DOI: 10.1007/s10681-020-02618-6 (cited on page 131).
- Navas-Lopez, J. F., J. Cano, R. de la Rosa, L. Velasco, and L. Leon (2020). Genotype by environment interaction for oil quality components in olive tree. *European Journal of Agronomy* 119, p. 126115. DOI: 10.1016/j.eja.2020.126115 (cited on page 129).

- Neches, R., R. E. Fikes, T. Finin, T. Gruber, R. Patil, T. Senator, and W. R. Swartout (1991). Enabling technology for knowledge sharing. *AI Magazine* 12.3, pp. 36–36. DOI: 10.1609/aimag.v12i3.902 (cited on page 10).
- Neveu, P., A. Tireau, N. Hilgert, V. Nègre, J. Mineau-Cesari, N. Bricchet, R. Chapuis, I. Sanchez, C. Pommier, B. Charnomordic, *et al.* (2019). Dealing with multi-source and multi-scale information in plant phenomics: the ontology-driven Phenotyping Hybrid Information System. *New Phytologist* 221, pp. 588–601. DOI: 10.1111/nph.15385 (cited on pages 33, 39).
- Niu, J., S. Zheng, X. Shi, Y. Si, S. Tian, Y. He, and H.-Q. Ling (2020). Fine mapping and characterization of the awn inhibitor B1 locus in common wheat (*Triticum aestivum* L.) *The Crop Journal* 8.4, pp. 613–622. DOI: 10.1016/j.cj.2019.12.005 (cited on page 85).
- Noy, N. F., N. H. Shah, P. L. Whetzel, B. Dai, M. Dorf, N. Griffith, C. Jonquet, D. L. Rubin, M.-A. Storey, C. G. Chute, *et al.* (2009). BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Research* 37.Issue suppl\_2, W170–W173. DOI: 10.1093/nar/gkp440 (cited on page 33).
- Oliveira, A. A. de, M. F. Resende, L. F. V. Ferrão, R. R. Amadeu, L. J. M. Guimarães, C. T. Guimarães, M. M. Pastina, and G. R. A. Margarido (2020). Genomic prediction applied to multiple traits and environments in second season maize hybrids. *Heredity* 125.1, pp. 60–72. DOI: 10.1038/s41437-020-0321-0 (cited on page 130).
- Oliveira, I. C. M., J. H. S. Guilhen, P. C. de Oliveira Ribeiro, S. A. Gezan, R. E. Schaffert, M. L. F. Simeone, C. M. B. Damasceno, J. E. de Souza Carneiro, P. C. S. Carneiro, R. A. da Costa Parrella, *et al.* (2020). Genotype-by-environment interaction and yield stability analysis of biomass sorghum hybrids using factor analytic models and environmental covariates. *Field Crops Research* 257, p. 107929. DOI: 10.1016/j.fcr.2020.107929 (cited on page 128).
- Oury, F.-X., E. Heumez, B. Rolland, J. Auzanneau, P. Bérard, M. Brancourt-Hulmel, X. Charrier, H. Chiron, C. Depatureaux, L. Falchetto, O. Gardet, S. Gilles, A. Giraud, C. Lecomte, J.-Y. Morlais, P. Pluchard, D. Tropée, M. Trottet, P. Walczak, G. Doussinault, M. Rousset, and G. Charmet (2018). Winter wheat (*Triticum aestivum* L.) phenotypic data from the multiannual, multilocal field trials of the INRA Small Grain Cereals Network. Version V5. *Portail Data INRAE*. DOI: 10.15454/1.4489666216568333E12 (cited on pages 45, 155).
- Oury, F.-X., E. Heumez, B. Rolland, J. Auzanneau, P. Bérard, M. Brancourt-Hulmel, X. Charrier, H. Chiron, C. Depatureaux, L. Falchetto, O. Gardet, S. Gilles, A. Giraud, C. Lecomte, J.-Y. Morlais, P. Pluchard, D. Tropée, M. Trottet, P. Walczak, G. Doussinault, M. Rousset, and G. Charmet (2020a). *Winter wheat (Triticum aestivum L.) phenotypic data from the multiannual, multilocal field trials of the INRA Small Grain Cereals Network*. URL: <https://urgi.versailles.inra.fr/ephep/ephep/viewer.do#dataResults/trialSetIds=8> (visited on 03/03/2021) (cited on page 43).
- Oury, F.-X., C. Pommier, and G. Charmet (2020b). *Enabling reusability of plant phenomic datasets with MIAPPE 1.1 - Supplementary dataset INRA Wheat (GnplS BrAPI endpoint) [Data set]*. URL: <https://urgi.versailles.inra.fr/faidare/brapi/v1/trials/dXJu01VSR0kvdHJpYWwvNw==> (cited on pages 44, 155).

- Oury, F.-X., C. Pommier, and G. Charmet (2020c). *Enabling reusability of plant phenomic datasets with MIAPPE 1.1 - Supplementary dataset INRA Wheat [Data set]*. Version V1. Portail Data INRAE. DOI: 10.15454/1AFKZ2 (cited on pages 44, 155).
- Pandey, M. K., S. Chaudhari, D. Jarquin, P. Janila, J. Crossa, S. C. Patil, S. Sundravadana, D. Khare, R. S. Bhat, T. Radhakrishnan, *et al.* (2020). Genome-based trait prediction in multi-environment breeding trials in groundnut. *Theoretical and Applied Genetics* 133.11, pp. 3101–3117. DOI: 10.1007/s00122-020-03658-1 (cited on page 130).
- Papatheodorou, I., N. A. Fonseca, M. Keays, Y. A. Tang, E. Barrera, W. Bazant, M. Burke, A. Füllgrabe, A. M.-P. Fuentes, N. George, *et al.* (2018). Expression Atlas: gene and protein expression across multiple studies and organisms. *Nucleic Acids Research* 46.D1, pp. D246–D251. DOI: 10.1093/nar/gkx1158 (cited on page 19).
- Papoutsoglou, E. A., D. Faria, D. Arend, E. Arnaud, I. N. Athanasiadis, I. Chaves, F. Coppens, G. Cornut, B. V. Costa, H. Ćwiek-Kupczyńska, B. Driesbeke, R. Finkers, A. Junker, G. J. King, P. Krajewski, M. Lange, M.-A. Laporte, C. Michotey, M. Oppermann, R. Ostler, H. Poorter, R. Ramírez-Gonzalez, J. C. Reif, P. Rocca-Serra, S.-A. Sansone, U. Scholz, F. Tardieu, C. Uauy, B. Usadel, R. G. Visser, S. Weise, P. J. Kersey, C. M. Miguel, A.-F. Adam-Blondon, and C. Pommier (2020a). Enabling reusability of plant phenomic datasets with MIAPPE 1.1. *New Phytologist* 227.1, pp. 260–273. DOI: 10.1111/nph.16544 (cited on pages 29, 73, 74, 76, 87).
- Papoutsoglou, E. A., D. Faria, D. Arend, E. Arnaud, I. N. Athanasiadis, I. Chaves, F. Coppens, G. Cornut, B. V. Costa, H. Ćwiek-Kupczyńska, B. Driesbeke, R. Finkers, A. Junker, G. J. King, P. Krajewski, M. Lange, M.-A. Laporte, C. Michotey, M. Oppermann, R. Ostler, H. Poorter, R. Ramírez-Gonzalez, J. C. Reif, P. Rocca-Serra, S.-A. Sansone, U. Scholz, F. Tardieu, C. Uauy, B. Usadel, R. G. Visser, S. Weise, P. J. Kersey, C. M. Miguel, A.-F. Adam-Blondon, and C. Pommier (2020b). Enabling reusability of plant phenomic datasets with MIAPPE 1.1 - Supplementary dataset. Version V1. DOI: 10.15454/1YXVZV (cited on pages 44, 154).
- Papoutsoglou, E. A. and G. Singh (2020a). *Test set - 4023 PubMed abstracts (for manuscript: Extracting knowledge networks from plant scientific literature: Potato tuber flesh color as an exemplary trait)*. Zenodo. DOI: 10.5281/zenodo.3999597 (cited on page 108).
- Papoutsoglou, E. A. and G. Singh (2020b). *WatsonPotato*. Github. URL: <https://github.com/PBR/WatsonPotato> (visited on 03/03/2021) (cited on page 111).
- Park, T.-H. (2005). *Identification, characterization and high-resolution mapping of resistance genes to Phytophthora infestans in potato*. PhD thesis. Wageningen University & Research. URL: <https://edepot.wur.nl/121660> (cited on pages 12, 15).
- Pauli, D., S. C. Chapman, R. Bart, C. N. Topp, C. J. Lawrence-Dill, J. Poland, and M. A. Gore (2016). The quest for understanding phenotypic variation via integrated approaches in the field environment. *Plant Physiology* 172.2, pp. 622–634. DOI: 10.1104/pp.16.00592 (cited on page 12).
- Pea, G., A. L. Hlaing, H. H. Aung, and M. E. Pè (2019a). *Zea mays MAGIC RIL population - field trial phenotyping data (PIPPA BrAPI endpoint) [Data set]*. URL:



- <https://pippa.psb.ugent.be/BrAPIPPA/brapi/v1/trials/3> (cited on pages 44, 156).
- Pea, G., A. L. Hlaing, H. H. Aung, and M. E. Pè (2019b). *Zea mays MAGIC RIL population - field trial phenotyping data [Data set]*. Zenodo. DOI: 10.5281/zenodo.3553749 (cited on pages 44, 156).
- Pettifer, S., J. Ison, M. Kalaš, D. Thorne, P. McDermott, I. Jonassen, A. Liaquat, J. M. Fernández, J. M. Rodriguez, I. Partners, *et al.* (2010). The EMBRACE web service collection. *Nucleic Acids Research* 38.suppl\_2, W683–W688. DOI: 10.1093/nar/gkq297 (cited on page 54).
- Pierce, H. H., A. Dev, E. Statham, and B. E. Bierer (2019). Credit data generators for data reuse. *Nature*. DOI: 10.1038/d41586-019-01715-4 (cited on page 22).
- Pieruschka, R. and U. Schurr (2019). Plant Phenotyping: Past, Present, and Future. *Plant Phenomics* 2019, pp. 1–6. DOI: 10.34133/2019/7507131 (cited on page 73).
- PIPPA team (2020). *PIPPA - PSB Interface for Plant Phenotype Analysis*. URL: <https://pippa.psb.ugent.be/> (visited on 03/03/2021) (cited on page 33).
- Piowar, H. A., R. S. Day, and D. B. Fridsma (2007). Sharing detailed research data is associated with increased citation rate. *PLOS One* 2.3, e308. DOI: 10.1371/journal.pone.0000308 (cited on page 124).
- Pommier, C., H. Ćwiek-Kupczyńska, D. Faria, E. Papoutsoglou, B. V. Bruno C, M.-A. Laporte, I. Chaves, P. Neveu, G. Cornut, M. Ruiz, and P. Larmande (2020). *Plant Phenotype Experiment Ontology (PPEO)*. URL: <http://purl.org/pp eo> (visited on 03/03/2021) (cited on pages 36, 87).
- Pommier, C., T. Letellier, J. Pietragalla, M. A. Laporte, E. Arnaud, J. Le Gouis, and R. Shrestha (2019a). Wheat Crop Ontology. Version V1. *Portail Data INRAE*. DOI: 10.15454/3EDMCP (cited on pages 44, 154, 155).
- Pommier, C., C. Michotey, G. Cornut, P. Roumet, E. Duchêne, R. Flores, A. Lebreton, M. Alaux, S. Durand, E. Kimmel, *et al.* (2019b). Applying FAIR principles to plant phenotypic data management in GnpIS. *Plant Phenomics* 2019, article ID 1671403. DOI: 10.34133/2019/1671403 (cited on pages 32, 36, 155).
- Poorter, H., J. Bühler, D. van Dusschoten, J. Climent, and J. A. Postma (2012). Pot size matters: a meta-analysis of the effects of rooting volume on plant growth. *Functional Plant Biology* 39.11, pp. 839–850. DOI: 10.1071/FP12049 (cited on page 116).
- Pundir, S., M. J. Martin, and C. O'Donovan (2017). UniProt protein knowledgebase. *Protein Bioinformatics*, pp. 41–55. DOI: 10.1093/nar/gkw1099 (cited on page 111).
- Radić, V., I. Balalić, M. Zćirić, Vasiljević, S. Jocić, and A. Marjanović-Jeromela (2020). Genotype×environment interaction of some traits in sunflower (*Helianthus Annuus L.*) lines. *Applied Ecology and Environmental Research* 18.1, pp. 1707–1719. DOI: 10.15666/aeer/1801\_17071719 (cited on page 134).
- Rahimi-Eichi, V., M. Okamoto, T. Garnett, P. Eckermann, B. Darrier, M. Riboni, and P. Langridge (2020). Strengths and weaknesses of national variety trial data for multi-environment analysis: a case study on grain yield and protein content. *Agronomy*, 10(5)–53. DOI: 10.3390/agronomy10050753 (cited on page 131).
- Rathnakumar, A., S. S. Manohar, H. L. Nadaf, S. C. Patil, M. P. Deshmukh, P. Thirumalaisamy, N. Kumar, H. Lalwani, P. Nagaraju, B. Yenagi, *et al.* (2020). G×E interactions in QTL introgression lines of Spanish-type groundnut (*Arachis hypogaea*

- L.) *Euphytica* 216, pp. 1–20. DOI: 10.1007/s10681-020-02613-x (cited on page 131).
- Rayner, T. F., P. Rocca-Serra, P. T. Spellman, H. C. Causton, A. Farne, E. Holloway, R. A. Irizarry, J. Liu, D. S. Maier, M. Miller, *et al.* (2006). A simple spreadsheet-based, MIAME-supportive format for microarray data: MAGE-TAB. *BMC Bioinformatics* 7, p. 489. DOI: 10.1186/1471-2105-7-489 (cited on page 33).
- Research Informatics Unit (RIU), CIP (2020). *Potato Ontology*. URL: [http://www.cropontology.org/ontology/C0\\_330/Potato](http://www.cropontology.org/ontology/C0_330/Potato) (visited on 03/03/2021) (cited on pages 124, 125).
- Reyes, L. F. and L. Cisneros-Zevallos (2007). Degradation kinetics and colour of anthocyanins in aqueous extracts of purple-and red-flesh potatoes (*Solanum tuberosum* L.) *Food Chemistry* 100.3, pp. 885–894. DOI: 10.1016/j.foodchem.2005.11.002 (cited on page 165).
- Rife, T. W. and J. A. Poland (2014). Field Book: An open-source application for field data collection on Android. *Crop Science* 54.4, pp. 1624–1627. DOI: 10.2135/cropsci2013.08.0579 (cited on page 55).
- Robert, P., J. Le Gouis, R. Rincet, B. Consortium, *et al.* (2020). Combining crop growth modeling with trait-assisted prediction improved the prediction of genotype by environment interactions. *Frontiers in Plant Science* 11, p. 827. DOI: 10.3389/fpls.2020.00827 (cited on page 131).
- Robins, J. G., C. W. Rigby, and K. B. Jensen (2020). Genotype  $\times$  environment interaction patterns in rangeland variety trials of cool-season grasses in the western United States. *Agronomy* 10.5, p. 623. DOI: 10.3390/agronomy10050623 (cited on page 131).
- Rocca-Serra, P., M. Brandizi, E. Maguire, N. Sklyar, C. Taylor, K. Begley, D. Field, S. Harris, W. Hide, O. Hofmann, *et al.* (2010). ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level. *Bioinformatics* 26.18, pp. 2354–2356. DOI: 10.1093/bioinformatics/btq415 (cited on pages 33, 36, 87).
- Römer, S., J. Lübeck, F. Kauder, S. Steiger, C. Adomat, and G. Sandmann (2002). Genetic engineering of a zeaxanthin-rich potato by antisense inactivation and co-suppression of carotenoid epoxidation. *Metabolic Engineering* 4.4, pp. 263–272. DOI: 10.1006/mben.2002.0234 (cited on page 165).
- Ruas, M., V. Guignon, G. Sempere, J. Sardos, Y. Hueber, H. Duvergey, A. Andrieu, R. Chase, C. Jenny, T. Hazekamp, *et al.* (2017). MGIS: managing banana (*Musa spp.*) genetic resources information and high-throughput genotyping data. *Database* 2017. DOI: 10.1093/database/bax046 (cited on pages 62, 67).
- Rubiales, D., E. Barilli, and F. Flores (2020). Broomrape as a major constraint for grass pea (*Lathyrus sativus*) production in mediterranean rain-fed environments. *Agronomy* 10.12, p. 1931. DOI: 10.3390/agronomy10121931 (cited on page 128).
- Rubiales, D. and F. Flores (2020). Adaptation of one-flowered vetch (*Vicia articulata* Hornem.) to Mediterranean rain fed conditions. *Agronomy* 10.3, p. 383. DOI: 10.3390/agronomy10030383 (cited on page 132).
- Sales, H., J. Nunes, and M. C. Vaz Patto (2020). Achievements and challenges towards a sustainable conservation and Use of 'Galega vulgar' *Olea europaea* variety. *Agronomy* 10.10, p. 1467. DOI: 10.3390/agronomy10101467 (cited on page 129).

- Sansone, S.-A., P. McQuilton, P. Rocca-Serra, A. Gonzalez-Beltran, M. Izzo, A. L. Lister, and M. Thurston (2019). FAIRsharing as a community approach to standards, repositories and policies. *Nature Biotechnology* 37, pp. 358–367. DOI: 10.1038/s41587-019-0080-8 (cited on pages 18, 33).
- Sansone, S.-A., P. Rocca-Serra, D. Field, E. Maguire, C. Taylor, O. Hofmann, H. Fang, S. Neumann, W. Tong, L. Amaral-Zettler, *et al.* (2012). Toward interoperable bioscience data. *Nature Genetics* 44, pp. 121–126. DOI: 10.1038/ng.1054 (cited on page 33).
- Selby, P., R. Abbeloos, J. E. Backlund, M. Basterrechea Salido, G. Bauchet, O. E. Benites-Alfaro, C. Birkett, V. C. Calaminos, P. Carceller, G. Cornut, *et al.* (2019). BrAPI — an application programming interface for plant breeding applications. *Bioinformatics*. DOI: 10.1093/bioinformatics/btz190 (cited on pages 34, 36, 41, 51, 73, 87).
- Sempéré, G., F. Philippe, A. Dereeper, M. Ruiz, G. Sarah, and P. Larmande (2016). Gigwa—Genotype investigator for genome-wide analyses. *GigaScience* 5.1, s13742–016. DOI: 10.1186/s13742-016-0131-8 (cited on page 62).
- Shakoor, N., S. Lee, and T. C. Mockler (2017). High throughput phenotyping to accelerate crop breeding and monitoring of diseases in the field. *Current Opinion in Plant Biology* 38, pp. 184–192. DOI: 10.1016/j.pbi.2017.05.006 (cited on page 12).
- Shannon, P., A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research* 13.11, pp. 2498–2504. DOI: 10.1101/gr.1239303 (cited on page 112).
- Sharma, V., W. Law, M. J. Balick, and I. N. Sarkar (2017). Harnessing biomedical natural language processing tools to identify medicinal plant knowledge from historical texts. *AMIA Annual Symposium Proceedings* 2017, p. 1537 (cited on page 98).
- Shi, Z., I. Zhovannik, A. Traverso, F. J. Dankers, T. M. Deist, P. Kalendralis, R. Monshouwer, J. Bussink, R. Fijten, H. J. Aerts, *et al.* (2019). Distributed radiomics as a signature validation study using the Personal Health Train infrastructure. *Scientific Data* 6.1, pp. 1–8. DOI: 10.1038/s41597-019-0241-0 (cited on page 23).
- Shrestha, R. (2020). *IBP Crop Research Ontology CO\_715*. URL: [http://www.cropontology.org/ontology/CO\\_715/Crop%20Research](http://www.cropontology.org/ontology/CO_715/Crop%20Research) (visited on 03/03/2021) (cited on page 39).
- Shrestha, R., L. Matteis, M. Skofic, A. Portugal, G. McLaren, G. Hyman, and E. Arnaud (2012). Bridging the phenotypic and genetic data useful for integrated breeding through a data annotation using the Crop Ontology developed by the crop communities of practice. *Frontiers in Physiology* 3, p. 326. DOI: 10.3389/fphys.2012.00326 (cited on pages 21, 33, 54, 67, 111).
- Sielemann, K., A. Hafner, and B. Pucker (2020). The reuse of public datasets in the life sciences: potential risks and rewards. *PeerJ* 8, e9954. DOI: 10.7717/peerj.9954 (cited on page 11).
- Singh, G. (2019). *Genomics data integration for knowledge discovery using genome annotations from molecular databases and scientific literature*. PhD thesis. Wageningen University & Research. DOI: 10.18174/505685 (cited on page 112).

- Singh, G., A. Kuzniar, E. M. van Mulligen, A. Gavai, C. W. Bachem, R. G. Visser, and R. Finkers (2018). QTLTableMiner++: semantic mining of QTL tables in scientific articles. *BMC Bioinformatics* 19.1, p. 183. DOI: 10.1186/s12859-018-2165-7 (cited on page 98).
- Singh, G., E. A. Papoutsoglou, F. Keijts-Lalleman, B. Vencheva, M. Rice, R. G. Visser, C. W. Bachem, and R. Finkers (2021). Extracting knowledge networks from plant scientific literature: Potato tuber flesh color as an exemplary trait. *BMC Plant Biology* 21.198. DOI: 10.1186/s12870-021-02943-5 (cited on page 95).
- Singh, G. and E. A. Papoutsoglou (2019a). *Cytoscape session for the potato knowledge graph extracted with IBM Watson's supervised NLP model*. Zenodo. DOI: 10.5281/zenodo.3275105 (cited on pages 99, 102).
- Singh, G. and E. A. Papoutsoglou (2019b). *IBM Watson's NLP model for annotating potato literature*. Zenodo. DOI: 10.5281/zenodo.3260364 (cited on page 112).
- Śliwka, J., I. Wasilewicz-Flis, H. Jakuczun, and C. Gebhardt (2008). Tagging quantitative trait loci for dormancy, tuber shape, regularity of tuber shape, eye depth and flesh colour in diploid potato originated from six *Solanum* species. *Plant Breeding* 127.1, pp. 49–55. DOI: 10.1111/j.1439-0523.2008.01420.x (cited on page 166).
- Small Grain Genetic Resource Centre (2020). *Small Grain Genetic Resource Centre*. URL: <https://urgi.versailles.inra.fr/siregal/siregal/card.do?id=1&dbName=siregal&className=genres.administration.GrcImpl> (visited on 03/03/2021) (cited on page 155).
- Sodedji, F. A. K., S. Agbahoungba, S.-P. A. Nguetta, E. E. Agoyi, M. A. T. Ayenan, S. H. Sossou, C. Mamadou, A. E. Assogbadjo, and D. Kone (2020). Resistance to legume pod borer (*Maruca vitrata* Fabricius) in cowpea: genetic advances, challenges, and future prospects. *Journal of Crop Improvement* 34.2, pp. 238–267. DOI: 10.1080/15427528.2019.1680471 (cited on page 132).
- Souza, M. H. de, J. D. Pereira Júnior, S. D. M. Steckling, J. Mencalha, F. d. S. Dias, J. R. d. A. S. d. C. Rocha, P. C. S. Carneiro, and J. E. d. S. Carneiro (2020). Adaptability and stability analyses of plants using random regression models. *PLOS One* 15.12, e0233200. DOI: 10.1371/journal.pone.0233200 (cited on page 128).
- Spindel, J. E. and S. R. McCouch (2016). When more is better: how data sharing would accelerate genomic selection of crop plants. *New Phytologist* 212.4, pp. 814–826. DOI: 10.1111/nph.14174 (cited on pages 12, 32, 67).
- SPTO: *Solanaceae Phenotype Ontology* (2018). URL: <http://bioportal.bioontology.org/ontologies/SPTO?p=classes&conceptid=root> (visited on 03/03/2021) (cited on page 109).
- Statistics Ontology Project (2020). *Statistics Ontology (STATO)*. URL: <http://statontology.org/> (visited on 03/03/2021) (cited on page 34).
- Steinbach, D., M. Alaux, J. Amselem, N. Choisne, S. Durand, R. Flores, A.-O. Keliet, E. Kimmel, N. Lapalu, I. Luyten, *et al.* (2013). GnplS: an information system to integrate genetic and genomic data from plants and fungi. *Database* 2013, article ID bat058. DOI: 10.1093/database/bat058 (cited on page 33).
- Stushnoff, C., L. J. Ducreux, R. D. Hancock, P. E. Hedley, D. G. Holm, G. J. McDougall, J. W. McNicol, J. Morris, W. L. Morris, J. A. Sungurtas, *et al.* (2010). Flavonoid profiling and transcriptome analysis reveals new gene–metabolite correlations in

- tubers of *Solanum tuberosum* L. *Journal of Experimental Botany* 61.4, pp. 1225–1238. DOI: 10.1093/jxb/erp394 (cited on page 165).
- Sulli, M., G. Mandolino, M. Sturaro, C. Onofri, G. Diretto, B. Parisi, and G. Giuliano (2017). Molecular and biochemical characterization of a potato collection with contrasting tuber carotenoid content. *PLOS One* 12.9, e0184143. DOI: 10.1371/journal.pone.0184143 (cited on page 97).
- Swanckaert, J., D. Akansake, K. Adofo, K. Acheremu, B. De Boeck, R. Eyzaguirre, W. J. Grüneberg, J. W. Low, and H. Campos (2020). Variance component estimations and mega-environments for sweetpotato breeding in West Africa. *Crop Science* 60.1, pp. 50–61. DOI: 10.2298/GENSR2001367M (cited on page 134).
- Szalay, A. and J. Gray (2006). Science in an exponential world. *Nature* 440.7083, pp. 413–414. DOI: 10.1038/440413a (cited on page 10).
- Tardieu, F., L. Cabrera-Bosquet, T. Pridmore, and M. Bennett (2017). Plant phenomics, from sensors to knowledge. *Current Biology* 27.15, R770–R783. DOI: 10.1016/j.cub.2017.05.055 (cited on page 32).
- Taylor, C. F., N. W. Paton, K. S. Lilley, P.-A. Binz, R. K. Julian, A. R. Jones, W. Zhu, R. Apweiler, R. Aebersold, E. W. Deutsch, et al. (2007). The minimum information about a proteomics experiment (MIAPE). *Nature Biotechnology* 25.8, pp. 887–893. DOI: 10.1038/nbt1329 (cited on page 11).
- Tenopir, C., S. Allard, K. Douglass, A. U. Aydinoglu, L. Wu, E. Read, M. Manoff, and M. Frame (2011). Data sharing by scientists: practices and perceptions. *PLOS One* 6.6, e21101. DOI: 10.1371/journal.pone.0021101 (cited on pages 10, 32).
- Tenopir, C., N. M. Rice, S. Allard, L. Baird, J. Borycz, L. Christian, B. Grant, R. Olendorf, and R. J. Sandusky (2020). Data sharing, management, use, and reuse: Practices and perceptions of scientists worldwide. *PLOS One* 15.3, e0229003. DOI: 10.1371/journal.pone.0229003 (cited on page 11).
- Teow, C. C., V.-D. Truong, R. F. McFeeters, R. L. Thompson, K. V. Pecota, and G. C. Yencho (2007). Antioxidant activities, phenolic and  $\beta$ -carotene contents of sweet potato genotypes with varying flesh colours. *Food chemistry* 103.3, pp. 829–838. DOI: 10.1016/j.foodchem.2006.09.033 (cited on page 166).
- Tessema, B. B. (2017). *Genetic studies towards elucidation of drought tolerance of potato*. PhD thesis. Wageningen University & Research. DOI: 10.18174/413763 (cited on pages 12, 15).
- The COPO team (2020). *COPO – Collaborative Open Plant Omics*. URL: <https://copo-project.org/> (visited on 03/03/2021) (cited on page 48).
- The Gene Ontology Consortium (2019). The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Research* 47.D1, pp. D330–D338. DOI: 10.1093/nar/gky1055 (cited on pages 33, 48).
- The ISA Team (2020). *ISA Commons*. URL: <https://www.isacommons.org/> (visited on 03/03/2021) (cited on page 41).
- The UniProt Consortium (2019). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research* 47.D1, pp. D506–D515. DOI: 10.1093/nar/gky1049 (cited on page 34).
- Tom, H. (Sept. 1994). The Geographic Information Systems (GIS) Standards Infrastructure. *StandardView* 2.3, pp. 133–142. ISSN: 1067-9936. DOI: 10.1145/202749.202755 (cited on page 122).

- Tong, H. and Z. Nikoloski (2021). Machine learning approaches for crop improvement: Leveraging phenotypic and genotypic big data. *Journal of Plant Physiology* 257, p. 153354. DOI: 10.1016/j.jplph.2020.153354 (cited on page 116).
- Van Berloo, R., R. Hutten, H. Van Eck, and R. Visser (2007). An online potato pedigree database resource. *Potato Research* 50.1, pp. 45–57. DOI: 10.1007/s11540-007-9028-3 (cited on page 19).
- Van Eck, J., B. Conlin, D. Garvin, H. Mason, D. Navarre, and C. Brown (2007). Enhancing beta-carotene content in potato by RNAi-mediated silencing of the beta-carotene hydroxylase gene. *American Journal of Potato Research* 84.4, pp. 331–342. DOI: 10.1007/BF02986245 (cited on page 166).
- Van Landeghem, S., S. De Bodt, Z. J. Drebert, D. Inzé, and Y. Van de Peer (2013). The potential of text mining in data integration and network biology for plant research: a case study on *Arabidopsis*. *The Plant Cell* 25.3, pp. 794–807. DOI: 10.1105/tpc.112.108753 (cited on page 98).
- Vandenbussche, P.-Y., G. A. Atemezing, M. Poveda-Villalón, and B. Vatant (2017). Linked Open Vocabularies (LOV): a gateway to reusable semantic vocabularies on the Web. *Semantic Web* 8.3, pp. 437–452. DOI: 10.3233/SW-160213 (cited on page 87).
- Villa, F., S. Balbi, I. N. Athanasiadis, and C. Caracciolo (2017). Semantics for interoperability of distributed data and models: Foundations for better-connected information. *F1000Research* 6. DOI: 10.12688/f1000research.11638.1 (cited on page 124).
- Vines, T. H., A. Y. Albert, R. L. Andrew, F. Débarre, D. G. Bock, M. T. Franklin, K. J. Gilbert, J.-S. Moore, S. Renaut, and D. J. Rennison (2014). The availability of research data declines rapidly with article age. *Current Biology* 24.1, pp. 94–97. DOI: 10.1016/j.cub.2013.11.014 (cited on pages 10, 32).
- W3C (2020a). *JSON-LD 1.0*. URL: <https://www.w3.org/TR/json-ld/> (visited on 03/03/2021) (cited on page 43).
- W3C (2020b). *RDF 1.1 Concepts and Abstract Syntax*. URL: <https://www.w3.org/TR/rdf11-concepts/> (visited on 03/03/2021) (cited on pages 18, 43).
- Wallis, J. C., E. Rolando, and C. L. Borgman (2013). If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology. *PLOS One* 8.7, e67332. DOI: 10.1371/journal.pone.0067332 (cited on page 11).
- Wang, C., A. Kalyanpur, J. Fan, B. K. Boguraev, and D. Gondek (2012). Relation extraction and scoring in DeepQA. *IBM Journal of Research and Development* 56.3.4, pp. 9–1. DOI: 10.1147/JRD.2012.2187239 (cited on page 109).
- Wang, J., C. Defrenne, M. L. McCormack, L. Yang, D. Tian, Y. Luo, E. Hou, T. Yan, Z. Li, W. Bu, et al. (2021). Fine-root functional trait responses to experimental warming: a global meta-analysis. *New Phytologist*. DOI: 10.1111/nph.17279 (cited on page 116).
- Wang, N., W. Fang, H. Han, N. Sui, B. Li, and Q.-W. Meng (2008). Overexpression of zeaxanthin epoxidase gene enhances the sensitivity of tomato PSII photoinhibition to high light and chilling stress. *Physiologia Plantarum* 132.3, pp. 384–396. DOI: 10.1111/j.1399-3054.2007.01016.x (cited on page 165).
- Wang, Q., Y. Cao, L. Zhou, C.-Z. Jiang, Y. Feng, and S. Wei (2015). Effects of postharvest curing treatment on flesh colour and phenolic metabolism in fresh-cut

- potato products. *Food Chemistry* 169, pp. 246–254. DOI: 10.1016/j.foodchem.2014.08.011 (cited on page 166).
- Werij, J. S. (2011). *Genetic analysis of potato tuber quality traits*. PhD thesis. Wageningen University & Research. URL: <https://edepot.wur.nl/183746> (cited on pages 12, 15).
- Werij, J. S., B. Kloosterman, C. Celis-Gamboa, C. R. De Vos, T. America, R. G. Visser, and C. W. Bachem (2007). Unravelling enzymatic discoloration in potato through a combined approach of candidate genes, QTL, and expression analysis. *Theoretical and Applied Genetics* 115.2, pp. 245–252. DOI: 10.1007/s00122-007-0560-y (cited on page 165).
- White, J. W., L. Hunt, K. J. Boote, J. W. Jones, J. Koo, S. Kim, C. H. Porter, P. W. Wilkens, and G. Hoogenboom (2013). Integrated description of agricultural field experiments and production: The ICASA Version 2.0 data standards. *Computers and Electronics in Agriculture* 96, pp. 1–12. DOI: 10.1016/j.compag.2013.04.003 (cited on page 48).
- Wilkinson, M. D., M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, *et al.* (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3.1, pp. 1–9. DOI: 10.1038/sdata.2016.18 (cited on pages 14, 32, 56, 73).
- Wilkinson, M. D. and M. Links (2002). BioMOBY: an open source biological web services proposal. *Briefings in Bioinformatics* 3.4, pp. 331–341. DOI: 10.1093/bib/3.4.331 (cited on page 54).
- Wilkinson, M. D., S.-A. Sansone, E. Schultes, P. Doorn, L. O. B. da Silva Santos, and M. Dumontier (2018). A design framework and exemplar metrics for FAIRness. *Scientific Data* 5.1, pp. 1–4. DOI: 10.1038/sdata.2018.118 (cited on page 123).
- Willemsen, J. (2018). *The identification of allelic variation in potato*. PhD thesis. Wageningen University & Research. DOI: 10.18174/459655 (cited on pages 12, 15).
- Wolstencroft, K., O. Krebs, J. L. Snoep, N. J. Stanford, F. Bacall, M. Golebiewski, R. Kuzyakiv, Q. Nguyen, S. Owen, S. Soiland-Reyes, *et al.* (2017). FAIRDOMHub: a repository and collaboration environment for sharing systems biology research. *Nucleic Acids Research* 45.D1, pp. D404–D407. DOI: 10.1093/nar/gkw1032 (cited on page 48).
- Wolters, A.-M. A., J. G. Uitdewilligen, B. A. Kloosterman, R. C. Hutten, R. G. Visser, and H. J. van Eck (2010). Identification of alleles of carotenoid pathway genes important for zeaxanthin accumulation in potato tubers. *Plant Molecular Biology* 73.6, pp. 659–671. DOI: 10.1007/s11103-010-9647-y (cited on pages 105, 164).
- Xin, J., C. Afrasiabi, S. Lelong, J. Adesara, G. Tsueng, A. I. Su, and C. Wu (2018). Cross-linking BioThings APIs through JSON-LD to facilitate knowledge exploration. *BMC Bioinformatics* 19.1, pp. 1–7. DOI: 10.1186/s12859-018-2041-5 (cited on page 67).
- Yang, W., H. Feng, X. Zhang, J. Zhang, J. H. Doonan, W. D. Batchelor, L. Xiong, and J. Yan (2020). Crop phenomics and high-throughput phenotyping: past decades, current challenges, and future perspectives. *Molecular Plant* 13.2, pp. 187–214. DOI: 10.1016/j.molp.2020.01.008 (cited on page 12).

- Yilmaz, P., R. Kottmann, D. Field, R. Knight, J. R. Cole, L. Amaral-Zettler, J. A. Gilbert, I. Karsch-Mizrachi, A. Johnston, G. Cochrane, *et al.* (2011). Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nature Biotechnology* 29.5, pp. 415–420. DOI: 10.1038/nbt.1823 (cited on page 11).
- Zaban, A., M. Veteläinen, C. Celis-Gamboa, R. van Berloo, H. Häggman, and R. Visser (2006). Physiological and genetic aspects of a diploid potato population in the Netherlands and Northern Finland. *Suomen maataloustieteellisen seuran tiedote* 21, pp. 1–7. DOI: 10.33354/smst.76724 (cited on pages 73, 75).
- Zamir, D. (2013). Where have all the crop phenotypes gone? *PLOS Biology* 11.6, e1001595. DOI: 10.1371/journal.pbio.1001595 (cited on page 32).
- Zewdu, Z., T. Abebe, T. Mitiku, F. Worede, A. Dessie, A. Berie, and M. Atnaf (2020). Performance evaluation and yield stability of upland rice (*Oryza sativa* L.) varieties in Ethiopia. *Cogent Food & Agriculture* 6.1, p. 1842679. DOI: 10.1080/23311932.2020.1842679 (cited on page 133).
- Zhang, W., J. Hu, Y. Yang, and Y. Lin (2020). One compound approach combining factor-analytic model with AMMI and GGE biplot to improve multi-environment trials analysis. *Journal of Forestry Research* 31.1, pp. 123–130. DOI: 10.1007/s11676-018-0846-8 (cited on page 133).
- Zhang, Y., C. S. Jung, and W. S. De Jong (2009). Genetic analysis of pigmented tuber flesh in potato. *Theoretical and Applied Genetics* 119.1, pp. 143–150. DOI: 10.1007/s00122-009-1024-3 (cited on page 166).
- Zhou, X., R. McQuinn, Z. Fei, A.-M. A. Wolters, J. van Eck, C. Brown, J. J. Giovannoni, and L. Li (2011). Regulatory control of high levels of carotenoid accumulation in potato tubers. *Plant, Cell & Environment* 34.6, pp. 1020–1030. DOI: 10.1111/j.1365-3040.2011.02301.x (cited on page 165).







**Summary**

**Acknowledgements**

**About the author**

**Education statement**



# Summary

The increasing nutritional demands of the world as well as the need for crops that perform reliably, in spite of diverse environmental conditions (abiotic and biotic stresses and variable weather conditions), put the plant sciences at the forefront of domains where progress is urgently needed. To be able to do so, plant phenotyping and genotyping are extremely important. Especially in plant phenotyping, research is met with challenges related to poor data management, and thereby inefficient exploitation - let alone reuse of datasets. The challenges to phenotypic data reuse and integration arise due to the highly distributed nature of data in the domain (as there are no central plant phenotypic data repositories) and their multifaceted heterogeneity. The variety of experimental goals and the sheer number of species studied may necessitate different approaches (e.g. for crops, model organisms, forest trees). Experiments may be conducted in open fields, greenhouses or other locations, follow different designs and produce different types of data (e.g. visual observation of a score, images, manual and automatic measurements, molecular assays). Even when everything else matches, the data files produced may have different formats and structures, which is a challenge for data integration. Moreover, good data documentation practices are often lacking, which hinders interpretation and reuse. In the vast majority of cases, plant phenotyping datasets are used only once, solely to address the research question for which they were originally generated. It is the exception, rather than the rule, when different datasets, produced by different, uncoordinated parties, are analyzed to generate further knowledge. Even rarer, though much more useful, are cases where independently created datasets are integrated for the purpose of meta-analyses or improvement of statistical and predictive models. Such work is crucial, for example, for multi-environment studies investigating the adaptability of crops to different conditions. This relative rarity of meta-analyses and integrative studies indicates that researchers conduct experiments and collect data anew for every new study they wish to undertake, which is in many cases a suboptimal use of resources. This may not be a serious issue on a low level (i.e., single experiments) but on a higher level where multiple independent experiments may be reused and integrated in e.g. multi-environment trials, this has a greater impact.

The challenges mentioned in the previous paragraph for plant phenotyping are not specific, but generic for the data life cycle in research. To address this challenge, the FAIR (Findable, Accessible, Interoperable, Reusable) data principles have been proposed as guidelines to alleviate generic reusability bottlenecks. However, FAIR data principles require domain-specific solutions. With them, datasets become more easily discoverable, interpretable, integratable and reusable. Furthermore, the principles emphasize that there should be an equal focus on human and machine readability, so that automated techniques can facilitate every step of the process. It is up to each community to devise ways to implement the FAIR principles. In this thesis, we investigated the application of the FAIR data principles in the domain of plant phenotyping. Our initial research question focused on a core requirement of FAIR, domain-relevant community standards. We identified and tackled shortcomings of the MIAPPE (Minimum Information About a Plant Phenotyping Experiment) metadata standard, which was initially presented as

a flat checklist. We produced a new, refined version, MIAPPE 1.1, which can cover experiments involving a broader range of plant species (including forest trees), boasts improved documentation, and can now support FAIR data through its explicit data model and ontology (**Chapter 2**). We tested the new version of the standard by using it to describe a wide range of different plant phenotyping experiments which proved that it can sufficiently accommodate the metadata of those experiments in a variety of formats.

For our second research question, we addressed the needs of machine readable data exchange for plant breeding information systems with the plant Breeding API (BrAPI), a standardized RESTful API (Application Programming Interface) specification, developed by and for the community (**Chapter 3**). Unlike MIAPPE, which is strictly a metadata standard for phenotyping experiments, BrAPI has a broader scope, which covers phenotypic and genotypic data alike. BrAPI can now be used to interact uniformly with breeding systems, fetching essential genotypic, phenotypic and organizational information, and BrAPI-compliant endpoints can support modular applications for a variety of use cases. Finally, we ensured that BrAPI includes community-relevant metadata by following the MIAPPE community standard and ensuring that its essential attributes were present. BrAPI is only one of the MIAPPE implementations: in **Chapter 2**, to make the metadata standard easier to adopt, we provided more of them for different usage contexts. We developed the Plant Phenotype Experiment Ontology for the RDF (Resource Description Framework) implementation, and a configuration supporting ISA-Tab (Investigation Study Assay-Tabular) archives. Therefore, all of these implementations can now communicate MIAPPE-compliant datasets, in fulfillment of the FAIR data requirements for reuse (domain-relevant community standards).

For our third research question, we retrace some of the steps of a previous project, which revolved around the integration and reuse of heterogeneous data (phenotypic, genotypic, environmental) from potato experiments. Reuse was challenging in that project mainly due to a lack of organized metadata, which is a central requirement for FAIR data. Otherwise, resolving the heterogeneity in the presentation of data to arrive at a common format was time consuming and, in some cases, ambiguous. To improve this, in **Chapter 4**, we report steps toward better reusability of the data. Relevant subsets of the datasets were made FAIR and placed on a FAIR Data Point, which can be used to discover, acquire and reproducibly reuse the data. This process proved that the MIAPPE standard can support this integration, and highlighted difficulties that may arise when documentation and metadata are not compiled when an experiment is first conducted. It also emphasized that attributes supported by MIAPPE can be used to integrate datasets from different domains (phenotyping, environment), a type of integration crucial to investigations of crop stability. The FAIR Data Point provides the location of an RDF version (distribution) of the phenotypic dataset. We show that, by using a Jupyter notebook that interacts with it, we can easily create different views of the data, and that combining it with (environmental) data obtained from external resources is trivial.

Finally, we took a different approach toward data integration, findability and reusability. The core concern was the accelerating pace of research publications and the limited time that researchers can devote to consuming large volumes of text. Whereas databases and other structured information sources can be readily explored, articles - which primar-

ily consist of unstructured text - do not enjoy the same benefit. To present researchers with a more efficient means toward hypothesis generation, we constructed knowledge networks based on relationships extracted with natural language processing (NLP) methods, in particular IBM's Watson suite. Using potato tuber flesh color as the trait of interest, we conducted a time analysis to test the viability of our approach, discovering that latent connections hinting at new genotype-phenotype associations between particular metabolites, proteins and genes existed already for longer periods in literature before they were experimentally confirmed. Our knowledge networks included new and testable genes two years ahead of the actual publications (**Chapter 5**).

This thesis contributes to state-of-the-art methods for making plant phenotyping data FAIR. With metadata standards to aid interpretation and reusability, and better means for computer-readable data exchange, an infrastructure can be set up to benefit farmers, academic and industry stakeholders. Not only can better data management pave the way for more reuse and more powerful analyses and models, improving the landscape for plant research and the outlook for advances in the domain; it can also help with gaining new insights which would not have been possible without the linked datasets. We show using the carrot rather than using the stick that, by having FAIR plant phenotyping data, we can enhance re-use and further integration of existing datasets and enable a new era of data-driven research.

# Acknowledgements

A PhD usually takes a long time to complete, and is probably best described as a rollercoaster of successes and failures. I would like to express my appreciation to everyone who helped me (mostly) avoid the latter while achieving and appreciating the former. I am grateful for the opportunity to put this long list on paper in this final section. It will be difficult to do justice to the support I enjoyed over the years. If your name is conspicuously missing from this list, it is because no words could capture my feelings of gratitude to you!

I would like to start by expressing my deepest appreciation to my supervisors.

**Richard (Visser)**, thank you for granting me the possibility to do this PhD. You made me feel welcome to the department. Your scope and depth of knowledge about everything related to plant breeding have always amazed me, so your reactions to my work and acknowledgement that it is indeed important have been truly validating. You made time for me even under tight deadlines and helped me with your honest feedback and open lines of questioning. I am a better researcher for that. Thank you!

**Richard (Finkers)**, where to start? Thank you for entrusting this PhD project to me, for your mentorship, always making sure that I had the support I needed, giving me the freedom to pursue my interests, and for believing in my abilities and judgement. You have been instrumental in my development as a researcher, with our conversations, the conferences you encouraged me to attend, and the people you introduced me to. You also contributed to my growth as a person, as I observed your inspiring leadership and contagious laid-back disposition.

**Ioannis**, we met close to the start of my engineering studies, a very long time ago... You introduced me to ontologies and the semantic web: topics that, as it turns out, have had a lasting impact on me! You granted me the opportunity to undertake my diploma thesis with you as a supervisor, and the rest, as they say, is history. Without you, I would not have become aware of, or dared start this PhD. You have consistently pushed me to achieve results I thought were beyond my grasp, helped me expand my academic horizons, and provided me with guidance whenever I needed it. You also gave me the opportunity to try teaching, which was an interesting experience. Thank you so much!

I know that I have been privileged to have such an amazing supervision team. No matter how little progress I had to discuss in some of our meetings, I invariably left feeling motivated and with a grin on my face.

The previous chapters list many authors and I am certainly grateful to everyone who contributed. I would like to highlight some individuals and apologise to the rest, or else this section would be much longer than it already is. First, I wish to thank everyone in the Elixir Plant Science group. I knew next to nothing when I first started joining the meetings and it took me a while to start catching up, but everyone was really welcoming and accommodating. **Paul Kersey** and **Célia Miguel** were crucial for shaping this environment. Thank you!

**Cyril Pommier:** Your leadership and contributions to the domain have been a huge inspiration. I feel extremely fortunate to have been able to work with you as much as I have. I am deeply thankful for the opportunities you presented me with, and for the trust you have shown in me. Moreover, thank you for sharing your experiences over conference dinners: they were always something to look forward to!

**Daniel Faria,** thank you so much. You taught me a lot about ontologies and you showed me how much I still have to learn. I look back on the days I spent at the IGC, working with you on PPEO, as one of the most focused periods of my academic career. Thank you also for your writing advice!

**Guillaume,** I had fun and learned from you when we worked together at BrAPI hackathons. Thank you! **Bert,** I believe that you are the person who spent the most time reading my documentation for the MIAPPE ISA-Tab configuration. Thank you for your efforts and for being a great conference companion. **Živa** and **Kristina,** it was always great to see you around Elixir meetings. I learned a lot of interesting things about potato from you. **Anne-Françoise** and **Inês Chaves,** I always appreciated your presence, not only for your helpful feedback and patience, but also for your openness and friendliness. Thank you!

I would also like to thank the BrAPI community. The hackathons were not only productive but also fun. **Peter Selby,** thank you for your enthusiastic coordination of all BrAPI-related things! The goals I worked toward were supported and appreciated in the environment you helped cultivate.

**Philippe Rocca-Serra,** thank you for sharing your knowledge of ISA. Your patience made my work so much easier!

**Hanna Ćwiek,** it has been a true pleasure working with you on everything MIAPPE-related. I appreciated our interactions in work meetings, and I enjoyed our outings in Montpellier. Thank you!

**Marie-Angélique Laporte,** your work on AGRO and related ontologies, among others, has been an inspiration. Thank you for our pleasant interactions at BrAPI hackathons and conferences.

I would like to give special thanks to the whole Elixir-NL, DTL and GO-FAIR teams, with certain members of which I enjoyed many interactions. **Rob Hooft,** thank you for pushing for so many developments. I found your work on the Data Stewardship Wizard inspiring. **Louis Bonino, Mark Thompson, Erik Schultes, Marco Roos, Rajaram Kaliyaperumal** and **Kees Burger,** I learned from you in various workshops. Thank you for your interesting presentations, friendliness and willingness to help!

On the IBM side, the Watson paper brought a fruitful collaboration with **Mark Rice, Frederique Keijts-Lalleman** and **Bilyana Vencheva.** Thank you for your ever positive attitudes and your help.

Of course, there are also many people that made my PhD experience as good as it was within my home department of Plant Breeding. Many thanks to all colleagues!

I would like to especially thank the terrific secretaries, **Nicole, Daniëlle** and **Letty.** Without you, the department would fall apart before long. You helped me promptly with everything I needed, answered all of my questions and just made things work for



me, even when I was clueless. Moreover, you did everything with a smile. Thank you! **Christian Bachem**, thank you for your help with the Watson work, your guidance, and for sharing so many interesting stories about potatoes and more.

I would like to acknowledge my officemates, among others **Aviv, Yerisf, Carolina, Michiel, Matthijs** and **Manos** (it was great to have a Greek person around!): thank you all for your contribution to a positive working environment.

I am also grateful to all PBR Bioinformatics team colleagues, past and present: I first joined without knowing anything about plants or bioinformatics. Over the last five years, I have managed to change that, even if just a little bit, thanks to your talks and informative answers. I looked forward to our Tuesday morning meetings, and even more to the (remote) coffee chats we had in the Covid era. Without further ado, **Anand, Arnold, Brian, Chengcheng, Danny, Fernanda, Gurnoor, Martijn, Matthijs, Natascha, Patrick**: thank you!

**Patrick**, you played a critical role in helping me acclimatise when I first came to the university. Our work on BrAPI, and your patient explanations and guidance meant a lot to me.

**Matthijs**, it was great to have an officemate who understood my strange work, and I really enjoyed our chats. You have been helpful and I hope to one day be as knowledgeable and efficient as you are.

**Gurnoor**, the only other PhD student in the group for a long time: thank you so much! You showed me around when I first started, we had many interesting discussions about our work, and together we attended conferences and experienced frustrations, excitement and success. I am fortunate to have shared part of my PhD with you.

**Argyris**, we met when I started my diploma thesis in Greece. It has been a pleasure working with you, watching you become a Dr and moving forward with your career. Our interactions always left me in a good mood. Thank you!

A special thanks goes of course to my **paranymphs**. Thank you for your support! I look forward to overcoming this last hurdle of my PhD with you by my side (at a safe distance).

**Katharina**, we met early in our PhDs. Your involvement with the eLabjournal gave us the opportunity to get to know each other better, for which I am extremely grateful. You have become my closest friend on this journey and an inspiration and, no matter where we end up, I look forward to celebrating many more milestones with you.

**Natascha**, we met a bit later and it was not until the pandemic hit that I started getting to know you a bit better. That was clearly a mistake: there is never a dull moment with you around. Thank you also for meticulously proofreading my introduction chapter. You have now embarked on your own journey toward a PhD so, even though you don't need it, good luck!

**Óscar, Daniel, João, Joey, Nick and Rob**, my odd group of international friends: Though we have not managed to meet much in recent years, you have all been a much needed constant in my life for well over a decade. That is no minor achievement! Thank you for your unwavering presence —day and night— and support, and for all the good times.

**Stephan**, my partner in crime: thank you for your continued support and for always making me laugh, sometimes (often?) in spite of myself. Best of luck finishing your PhD! I'd also like to express my gratitude to Stephan's family (especially **Michael, Brigitte** and **Linde-Irisa**): Hartelijk bedankt dat jullie me vanaf het eerste moment geaccepteerd hebben, en dat jullie me thuis hebben laten voelen.

Last but not least, I would like to thank my family. Aunts, uncles and cousins, back in Greece and outside of it, thank you for believing in me!

**Agni** (μαμά), **Giannis** (παμπά), this thesis is dedicated to you. None of this would have been possible without you. Thank you for always being there for me, unconditionally and patiently supporting my decisions, and taking pride in my achievements. There is no greater gift I could have asked for!

# About the author

Evangelia Anastasia (aka Eliana) Papoutsoglou was born in Thessaloniki, Greece, in 1992. Throughout her early education, she discovered that she was a practical person and appreciated the sciences, but developed a bias against most humanities (except for languages, the utilitarian one). She studied English (evidently), German (somehow earning an Abitur from the German School of Thessaloniki) and Japanese [eventually triumphing (over not too many competitors) in a national competition]. Finally, she realized that her aptitudes were suited to non-human languages as well.

Motivated by her growing fascination with technology, she followed a 5-year programme in Electrical and Computer Engineering at the Democritus University of Thrace in Xanthi, Greece, graduating in 2015 and specializing in Electronic and Information Systems. Her diploma (or MSc-equivalent) thesis took her on an adventure to understand and manage data heterogeneity in air pollution sensor data, in collaboration with the Swiss Tropical and Public Health institute, under supervision of Ioannis Athanasiadis. This piqued her continued interest in data management, integration and semantics.

In 2016, Eliana moved to the Netherlands to join the department of Plant Breeding at Wageningen University & Research and start her journey toward a PhD. Though not a biologist/plant scientist/breeder herself, she became enthusiastic about data management and generation for breeding purposes. Because of that, she was successful in collaborating with other researchers to advance the FAIR data principles for plant phenotyping. The results of this project culminated in the present thesis, with her supervisors Richard Visser, Richard Finkers, and Ioannis Athanasiadis once again.

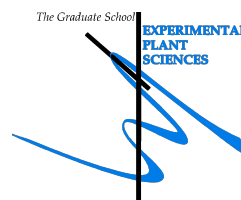
Eliana's PhD work also took her on literal journeys, as she had ample opportunity to travel. She participated in working groups through Elixir (the European research infrastructure for life science data) in different countries, attended conferences and workshops, met all kinds of different people, and got to feel quite international.

Nowadays, her research interests lie in interoperability, linked data, ontologies, the FAIR principles and machine learning. She likes puzzles of different kinds, fiction and animals (and spent two weeks of her PhD with a rescue gosling).



# Education Statement of the Graduate School

## Experimental Plant Sciences



**Issued to:** Evangelia A. Papoutsoglou  
**Date:** 16 June 2021  
**Group:** Plant Breeding  
**University:** Wageningen University & Research

| 1) Start-Up Phase   | <u>date</u>  | <u>cp</u> |
|---|--------------|-----------|
| <b>► First presentation of your project</b><br>Automatic semantic integration for the effective reuse of genotypic and phenotypic data, following the FAIR data principles      | 31 Jan, 2017 | 1.5       |
| <b>► Writing or rewriting a project proposal</b><br>Automatic semantic integration for the effective reuse of genotypic and phenotypic data, following the FAIR data principles | 03 Apr, 2017 | 6.0       |
| <b>► MSc courses</b><br>PBR-22303 Plant Breeding  | 2017         | 3.0       |

*Subtotal Start-Up Phase*

10.5

| 2) Scientific Exposure   | <u>date</u>  | <u>cp</u>  |
|--|--|--|
| <b>► EPS PhD student days</b><br>EPS PhD student days ("Get2Gether"), Soest (NL)<br>EPS PhD student days ("Get2Gether"), Soest (NL)  | 15-16 Feb, 2018<br>11-12 Feb, 2019   | 0.6<br>0.6   |
| <b>► EPS theme symposia</b><br>EPS Theme 1 Symposium, "Developmental Biology of Plants", Wageningen (NL)<br>EPS Theme 2 Symposium, "Interactions between Plants and Biotic Agents", Utrecht (NL)   | 05 Feb, 2020<br>04 Feb, 2020   | 0.3<br>0.3   |
| <b>► Lunteren Days and other national platforms</b><br>Annual Meeting "Experimental Plant Sciences", Lunteren (NL)<br>Annual Meeting "Experimental Plant Sciences", Lunteren (NL)<br>National eScience Symposium 2016, Amsterdam (NL)<br>National eScience Symposium 2017, Amsterdam (NL)<br>Bioinformatics and Systems Biology (BioSB) conference 2017, Lunteren (NL)<br>Bioinformatics and Systems Biology (BioSB) conference 2019, Lunteren (NL)  | 10-11 Apr, 2017<br>08-09 Apr, 2019<br>13 Oct, 2016<br>12 Oct, 2017<br>04-05 Apr, 2017<br>02-03 Apr, 2019   | 0.6<br>0.6<br>0.3<br>0.3<br>0.6<br>0.6   |
| <b>► Seminars (series), workshops and symposia</b><br>FAIR Data Stewardship Workshop, Wageningen (NL)<br>IBM Watson Knowledge Studio lab, Amsterdam (NL)<br>DTL Programmers meeting, Utrecht (NL)<br>SWAT4LS conference hackathon - tutorial days, Amsterdam (NL)<br>Breeding API hackathon, Montpellier (FR)<br>DTL Programmers meeting, Utrecht (NL)<br>Lorentz Workshop: How to make data FAIR for open science, Leiden (NL)<br>BYOD BrAPI Workshop, Ghent (BE)<br>Breeding API hackathon, Seattle (WA, USA)<br>RDFeno (MIAPPE) workshop (ELIXIR capacity exchange), Lisbon (PT)<br>FarmHack, Leeuwarden (NL)<br>Seminar: Access to (plant) data, Wageningen (NL) | 04 Nov, 2016<br>09 Nov, 2016<br>25 Nov, 2016<br>5, 8 Dec, 2016<br>12-16 Dec, 2016<br>20 Jan, 2017<br>15-19 May, 2017<br>30 May - 01 Jun, 2017<br>12-16 Jun, 2017<br>26-28 Sep, 2017<br>24-25 Nov, 2017<br>11 Dec, 2017 | 0.3<br>0.3<br>0.2<br>0.6<br>1.5<br>0.2<br>1.5<br>0.9<br>1.5<br>0.9<br>0.6<br>0.2 |

|   |                       |      |
|---|-----------------------|------|
| Breeding API meeting, Wageningen (NL)   | 12-14 Dec, 2017       | 0.9  |
| Breeding API hackathon, Versailles (FR)   | 04-09 Feb, 2018       | 1.5  |
| Elixir Capacity Exchange meeting, Lisbon (PT)   | 16-17 Apr, 2018       | 0.6  |
| SWAT4LS conference - tutorial day, Antwerp (BE)   | 03 Dec, 2018          | 0.3  |
| Breeding API hackathon, Wageningen (NL)   | 29 Apr - 03 May, 2019 | 1.5  |
| ► <b>Seminar plus</b>   |                       |      |
| ► <b>International symposia and congresses</b>  |                       |      |
| SWAT4LS conference, Amsterdam (NL)  | 06-07 Dec, 2016       | 0.6  |
| Elixir All Hands meeting, Rome (IT)   | 21-23 Mar, 2017       | 0.9  |
| International Semantic Web Conference 2017, Vienna (AT)   | 21-25 Oct, 2017       | 1.5  |
| PhenoHarmoniS 2018, Montpellier (FR)  | 14-17 May, 2018       | 1.2  |
| Elixir All Hands meeting, Berlin (DE)   | 04-07 Jun, 2018       | 1.2  |
| Elixir All Hands meeting, Lisbon (PT)   | 17-20 Jun, 2019       | 1.2  |
| Elixir All Hands meeting (virtual)  | 08-10 Jun, 2020       | 0.9  |
| ► <b>Presentations</b>  |                       |      |
| Talk: "Elixir Plant Use Case" (Session D4: ELIXIR: Data Interoperability & Plants), at BioSB 2017 conference  | 04 Apr, 2017          | 1.0  |
| Poster: "Added value from datasets: The C×E potato use case", at BioSB 2017 conference  | 04 Apr, 2017          | 1.0  |
| Poster: "Toward better data sharing methods for genebanks", at the International Semantic Web Conference 2017 (Semantics for Biodiversity Workshop)                         | 22 Oct, 2017          | 1.0  |
| Talk: "Towards FAIR: Standardizing plant phenotyping (meta)data with MIAPPE", at WUR B-Wise bioinformatics seminar  | 1 Oct, 2019           | 1.0  |
| Talk: "Beyond reproducibility: improving the reusability of plant phenotyping data with MIAPPE", at DTL Focus meeting: "Metadata for data reusability: eNotebook standards" | 31 Oct, 2019          | 1.0  |
| Talk: "MIAPPE 1.1: Building upon an existing standard for better plant phenomics data FAIRness", at Elixir All Hands 2020   | 10 Jun, 2020          | 1.0  |
| ► <b>3rd year interview</b>   |                       |      |
| ► <b>Excursions</b>   |                       |      |
| EPS PhD Council Company Visit: Tomato World   | 14 Oct, 2016          | 0.3  |
| <i>Subtotal Scientific Exposure</i>   |                       | 32.1 |

|  |                     |           |
|--|---------------------|-----------|
| <b>3) In-Depth Studies</b>   | <u>date</u>         | <u>cp</u> |
| ► <b>Advanced scientific courses &amp; workshops</b>   |                     |           |
| BioSB course: Algorithms for Biological Networks (including project), Wageningen (NL)  | 25-29 Jun, 2018     | 3.0       |
| edX online course with certificate: "Deep learning with TensorFlow" (IBM DL0120EN)   | 2020 - 23 Feb, 2021 | 0.7       |
| ► <b>Journal club</b>  |                     |           |
| ► <b>Individual research training</b>  |                     |           |
| Training and practice on semantic technologies - via Elixir capacity exchange, at Instituto Gulbenkian de Ciência, Lisbon (PT) | 18-24 Apr, 2018     | 1.5       |
| <i>Subtotal In-Depth Studies</i>   |                     | 5.2       |

| <b>4) Personal Development</b>  | <u>date</u>           | <u>cp</u> |
|---|-----------------------|-----------|
| ▶ <b>General skill training courses</b>                               |                       |           |
| EPS Introduction course, Wageningen (NL)                              | 16 Feb, 2017          | 0.3       |
| WGS course: Infographics and Iconography, Wageningen (NL)             | 26 Nov, 2019          | 0.3       |
| WGS course: Project and Time Management, Wageningen (NL)              | 05 Nov - 13 Dec, 2019 | 1.5       |
| WGS course: Career Perspectives, Wageningen (NL)                      | 09 Nov - 07 Dec, 2020 | 1.6       |
| ▶ <b>Organisation of meetings, PhD courses or outreach activities</b> |                       |           |
| ▶ <b>Membership of EPS PhD Council</b>                                |                       |           |

*Subtotal Personal Development*

3.7

| <b>5) Teaching &amp; Supervision Duties</b> | <u>date</u> | <u>cp</u> |
|---|-------------|-----------|
| ▶ <b>Courses</b>                            |             |           |
| INF-33306 Linked Data                       | 2019        | 2.8       |
| ▶ <b>Supervision of BSc/MSc students</b>    |             |           |

*Subtotal Teaching & Supervision Duties*

2.8

| <b>TOTAL NUMBER OF CREDIT POINTS*</b>  | <b>54.3</b> |
|--|-------------|
| <p>Herewith the Graduate School declares that the PhD candidate has complied with the educational requirements set by the Educational Committee of EPS with a minimum total of 30 ECTS credits.</p> <p><i>* A credit represents a normative study load of 28 hours of study.</i></p> |             |

The research described in this thesis was financially supported by a grant from Plant Breeding, Wageningen University & Research.

Financial support from Wageningen University for printing this thesis is gratefully acknowledged.

**Cover:** K. Hanika, E. A. Papoutsoglou  
**Layout:** E. A. Papoutsoglou  
**Printing:** ProefschriftMaken ([proefschriftmaken.nl](http://proefschriftmaken.nl))



