PRO-GRACE

PROMOTING A PLANT GENETIC RESOURCE COMMUNITY FOR EUROPE

DELIVERABLE 1.4

Minimum Information about a Biological collection

Call identifier:
PRO-GRACE
Grant agreement no: 101094738

Promoting a Plant Genetic Resource Community for Europe

Deliverable No. D1.4

Minimum Information about a Biological collection

Contractual delivery date:
M20

Actual delivery date:
M21

Responsible partner:
IPK
Contributing partners:
INRAE, UoB, UPV, WR, ENEA, IPGRI, PSR, GCDT, CSIC, CRI, JHI

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101094738.

Grant agreement no.	Horizon Europe – 101094738
Project full title	PRO-GRACE – Promoting a plant genetic resource community for Europe

Deliverable number	**D1.4**
Deliverable title	**Minimum Information about a Plant Genetic Resource**
Type	
Dissemination level	Public
Work package number	WP1
Author(s)	Catherine Hazel Aguilar, Stephan Weise, Markus Oppermann, Anne-Françoise Adam-Blondon, Joanna Magos Brehm, Nigel Maxted, Jaime Prohens, Antonio Granell, Clara Pons, Theo van Hintum, Lorenzo Maggioni, Paul Shaw, Bela Bartha, Vojtech Holubec, Luigi Guarino, Giovanni Giuliano
Keywords	Plant genetic resources, FAIR, data standard, documentation, information management, genetic resource center.

The research leading to these results has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101094738**.**

Table of Contents

ABBREVIATIONS AND ACRONYMS iv

Executive Summary vi

1. Introduction 1

2. Use of Terms 3

3. Current State of PGR Data Landscape 5

3.1 Stakeholders Generating and Using PGR-Related Data **5**

3.1.1 Genetic Resource Centers (GRC) 5

3.1.2 Research Institutions and Universities 7

3.1.3 Non-Governmental and Civil Society Organizations 8

3.1.4 Private sector 9

3.2 Data Types **11**

3.2.1 Passport Data 11

3.2.2 Phenotypic Data 15

3.2.3 Image Data 22

3.2.2 Genomic Data 25

3.3 PGR Information Management Challenges **27**

3.3.1 Data Fragmentation and Inconsistencies in Data Collection Protocols and Formats 27

3.3.2 Compromised and Incomplete Datasets 28

3.3.3 Volume and Complexity 30

3.3.4 Access and Availability 30

4. Assigning Persistent Unique Identifiers for PGR 31

5. Utilizing Controlled Vocabularies and Ontologies for PGR Documentation 34

6. The Concept of Minimum Information Standards in Data-Driven Science 34

7. Rationale for Developing an Integrative Framework: Harmonizing Minimum Information Checklists through MI-PGR 36

8. Core Features of MI-PGR 37

8.1 MI-PGR Level 1: Essential Identification **41**

8.2 MI-PGR Level 2: Detailed Identification **47**

8.3 MI-PGR Level 3: Basic phenotypic characteristics **68**

8.4 MI-PGR Level 4: Detailed Phenotypic Evaluation Traits and Comprehensive Image Data **75**

8.4 MI-PGR Level 5: Molecular Phenotype **84**

8.4 MI-PGR Level 6: Genetic Data **90**

9. Way Forward 99

References 101

ABBREVIATIONS AND ACRONYMS

ABS	Access and Benefit-Sharing
BRC	Biological Resource Center
CBD	Convention on Biological Diversity
C&E	Characterization and evaluation
CGIAR	Consultative Group on International Agricultural Research
CSO	Civil Society Organization
CWR	Crop Wild Relatives
DDBJ	DNA Data Bank of Japan
DiSSCo	Distributed System of Scientific Collections
DOI	Digital Object Identifiers
DwC	Darwin Core
ECPGR	European Cooperative Programme for Genetic Resources
ELIXIR	European Life-Science Infrastructure
EMPHASIS	European Infrastructure for Multi-Scale Plant Phenotyping and Simulation for Food Security in a Changing Climate
ENA	European Nucleotide Archive
ENVO	Environmental Ontology
EURISCO	European Search Catalogue for Plant Genetic Resources
EVA	European Evaluation Network
FAIR	Findable, Accessible, Interoperable, Reusable
FAO	The Food and Agriculture Organization of the United Nations
GBIF	Global Biodiversity Information Facility
GIS	Geographic Information Systems
GRC	Genetic resource center
GRIN	Germplasm Resources Information Network
GSC	Genomic Standards Consortium
INSDC	International Nucleotide Sequence Database Collaboration
ISA-tab	Investigation/Study/Assay tab-delimited format
ITPGRFA	International Treaty on Plant Genetic Resources for Food and Agriculture
MCPD	Multi-Crop Passport Descriptors
MIAME	Minimum Information about a Microarray Experiment
MIAPPE	Minimum Information about a Plant Phenotyping Experiment
MIDS	Minimum Information about a Digital Specimen
MIGS	Minimum Information about a Genome Sequence
MIMAG	Minimum Information about a Metagenome-assembled Genome
MI-PGR	Minimum Information about a Plant Genetic Resource
MIS	Minimum Information Standards
MISAG	Minimum Information about a Single Amplified Genome Sequence
MIxS	Minimum Information about any (x) Sequence
MLS	Multilateral System
NCBI	National Center for Biotechnology Information
NGS	Next-generation Sequencing
OBI	Ontology for Biomedical Investigations
OLS	Ontology Look Up Services
PECO	Plant Experimental Conditions Ontology

PGR	Plant genetic resources
PO	Plant Ontology
PUID	Persistent unique identifier
RI	Research Infrastructure
TDWG	Biodiversity Information Standards (formerly known as the Taxonomic Databases Working Group)
WIEWS	World Information and Early Warning System
WFP	Wild Food Plants
WMO	World Meteorological Organization

Executive Summary

Plant Genetic Resources (PGR), through the essential raw materials that they provide, enable present-day societies to address complex and multifaceted challenges confronting global food systems. This fact underscores the need for effective stewardship of PGR – an undertaking that can only be attained through a complete understanding of the genetic composition, phenotypic variation, genetic diversity, population structure, and crop-environment interactions associated with crops. To facilitate the process of screening, trait selection, varietal development and ultimately, utilization, detailed and accurate records must be obtained for every accession. One such example is harnessing of the genetic potential of crop wild relatives to develop resilient, climate-smart and nutrient-dense crop varieties that are widely adaptable in diverse ecological niches and that are suited to local consumer preferences, among other advantages.

Utilization of PGR-associated data and the genetic material itself for diverse purposes however, is never as straightforward as it appears. PGR science, being a multifaceted domain, requires trans- and interdisciplinary approaches drawing on expertise from the fields of genetics, molecular biology, seed science, plant breeding, agriculture, information technology, engineering, data science and the social sciences, among others. This domain becomes increasingly complex with the involvement of numerous stakeholders and networks, each bringing different approaches to PGR conservation, management and utilization. These fragmented efforts often result in siloed datasets, where inconsistencies in data collection, curation, reporting, and dissemination create substantial barriers to maximizing their potential for research, breeding and agricultural innovation.

Since the success of any conservation endeavor is contingent upon the availability of high-quality, comprehensive datasets, steps must be undertaken to make the process of data extraction effective, efficient, and if possible, with a minimum expenditure of funds. PGR-associated datasets that adhere to FAIR principles (Findable, accessible, interoperable and reusable) can significantly strengthen evidence-based conservation practices, promote sustainable resource use, and inform policy decisions across organizational and international boundaries to better achieve the end goal of PGR conservation — the strategic and optimal utilization of genetic resources.

This deliverable was developed with an overarching goal of capturing key data attributes to improve the documentation, accessibility and interoperability of PGR data. It presents a comprehensive overview of the current data landscape to set the stage for the proposed minimum reporting guidelines while providing an in-depth discussion of the various PGR-associated data types, along with existing standards and best practices. In addition, it identifies and elucidates existing opportunities, gaps and challenges that producers, curators and consumers of PGR data will come across in the course of data management and use. In particular, this deliverable introduces an integrative framework, named Minimum Information about a Plant Genetic Resource (MI-PGR), that proposes a coherent and harmonized set of guidelines for data collection, representation, annotation, and reporting to better describe and understand a PGR accession. Furthermore, this document outlines actionable steps required to advance this initiative, including its implementation within the future GRACE-RI, and ultimately the transformation of extensive datasets into valuable knowledge, thereby bridging the gap between PGR conservation and utilization.

1. Introduction

Technological advancements, data-driven by nature, have resulted in a surfeit of information that has facilitated scientific breakthroughs. In recent years, data generation has become hugely amplified, primarily driven by robust technologies that churn out huge volumes of data in real-time (Choi, 2019; Niedbala *et al.*, 2023; Shukla *et al*., 2023; Leal, 2024). However, this emerging magnitude of information has concurrently introduced challenges related to data stewardship, dissemination, accessibility, reuse, and interoperability (Wilkinson *et al.,* 2016). This situation is particularly evident in the domain of plant genetic resources (PGR) conservation, management and utilization, which has seen a massive influx of scientific information over the past years (Volk *et al.,* 2021). While holding transformative potential, this entire gamut of PGR information poses substantial obstacles to their coherent integration and practical application. Just like a double-edged sword, on one side this abundance of information opens new frontiers for research and innovation, while on the other causing intractable problems in effective data management, integration and use.

PGR provide the essential raw materials that can be utilized to address complex and multifaceted challenges that characterize global food systems today (Ulian *et al.,* 2020; Weise *et al.,* 2020; Pathirana & Carimi, 2022). Effective conservation, management and sustainable use of these resources require a thorough understanding of genetic composition, diversity, population structure, and environmental interactions; all of which can be gleaned from myriad data sources. Incidentally, the proliferation of quantitative large-scale and high-throughput technologies for genotyping and phenotyping plant accessions in the past two decades resulted in the rise of 'big data' (Arend *et al.,* 2016; Neveu *et al.,* 2018; Andres-Hernandez *et al.,* 2021) and the difficulties associated with its management (Scossa *et al.,* 2021; Lassoued *et al.,* 2021). These technological advancements notwithstanding, PGR conserved in genebanks remain largely underutilized due to difficulties in culling out useful accessions from large and diverse *ex situ* collections (Volk *et al.,* 2021).

The surge in PGR-associated data likewise has resulted in bottlenecks in processing and analysis that ultimately hinder the timely and effective translation of information into actionable insights. This notable disconnect between data generation and its practical application in PGR conservation strategies and subsequent utilization warrants the development of sophisticated information management systems (Ghaffar *et al*., 2020; Weise *et al*., 2020), structural and semantic standards (Andres-Hernandez *et al.*, 2021), and analytical tools (Volk *et al*., 2021; Wafula *et al*., 2023) to support evidence-based conservation practices, sustainable resource use, and policy decisions. Furthermore, data management, curation and integration initiatives must be cognizant of the needs of various actors who generate, access, synthesize and use PGR information across organizational and international boundaries.

In 2016, Wilkinson and colleagues introduced the FAIR principles[1,2] (Findable, Accessible, Interoperable, and Reusable) in response to the growing challenges in effective data handling and information exchange. These guiding principles advocate for systematic data stewardship practices that enhance the utility and impact of data by ensuring it is well-documented, easily accessible, and usable across different platforms and disciplines. Implementing FAIR principles in the context of PGR, while considering domain-specific constraints, can facilitate the efficient integration, access, exchange and use of data, thereby enhancing the collective capacity to sustainably conserve and utilize plant genetic diversity.

[1] https://www.go-fair.org/fair-principles/
[2] https://fairtoolkit.pistoiaalliance.org/fair-guiding-principles/

Interoperability, a crucial aspect of FAIR, specifically promotes the use of standardized data structures and common vocabularies, facilitating the harmonization of differently formatted data from diverse sources. This enables data to be adequately curated and maintained to the highest standards. The development of data standards, however, is a long and intricate process that involves a broad spectrum of stakeholders to ensure that these standards are comprehensive, inclusive and practical.

Presently, there are several well-recognized standards and minimum information checklists (e.g. List of Multi-Crop Passport Descriptors (MCPD)[3] (Alercia *et al.,* 2015), Darwin Core (DwC)[4] (Wieczorek *et al.,* 2012), Minimum Information About a Plant Phenotyping Experiment (MIAPPE)[5] (Ćwiek-Kupczyńska *et al.,* 2016; Papoutsoglou *et al.,* 2020), Minimum Information about any (x) Sequence (MIxS)[6] (Field *et al.,* 2008; Yilmaz *et al.,* 2011) and Audiovisual Core[7] (Morris *et al.,* 2013)) within the agricultural research and conservation circles. These standards were primarily developed independently within their respective disciplinary communities, each addressing the specific needs and priorities of its domain. This independent development process has led to differences in terminologies, structure, metadata schemas, and data elements across the standards. While these delineations are undoubtedly essential for achieving depth and specialization within each domain, they hinder effective data sharing, comparison, and comprehensive analysis. This is especially problematic in the context of PGR, where the efficient utilization of conserved resources is dependent on the ability to seamlessly integrate and analyse datasets that span multiple domains, including agricultural, genetic, phenotypic, environmental, taxonomic, and associated metadata information.

This deliverable presents a proposed framework for a coherent and harmonized set of guidelines that identify the critical information necessary to describe a PGR accession. These guidelines build upon established standards to create a common language, format, and structure, ensuring that the essential data needed by scientists, breeders, and other stakeholders are accurately captured and made accessible. It is important to note that while this document provides a detailed set of minimum reporting guidelines and recommendations, it remains a proposal at this stage. Transitioning to a recognized standard will require substantial further development, including extensive discussions with a range of actors involved in various PGR-related initiatives, and close collaboration with existing networks dedicated to data standardization.

Moreover, this document provides a comprehensive overview of the current data landscape to set the stage for the proposed minimum reporting guidelines. It includes an in-depth discussion on the various data types associated with PGR, along with existing standards and best practices. The document identifies and elucidates existing gaps, challenges, and opportunities to provide further context and justification for these guidelines. Additionally, it discusses the critical role of a “minimum information standard” in data-driven science, emphasizing how such standards, while termed "minimum," actually encompass a broad range of data elements essential for enhancing the reliability, accessibility, and interoperability of PGR data. Subsequently, it presents a strategy to capture key data attributes aimed at improving the documentation, access, and interoperability of PGR data. The document concludes with a detailed outline of the next steps required to advance this initiative.

[3] https://alliancebioversityciat.org/publications-data/faobioversity-multi-crop-passport-descriptors-v21-mcpd-v21-december-2015
[4] https://dwc.tdwg.org/
[5] https://www.miappe.org/
[6] https://www.gensc.org//pages/about.html
[7] https://ac.tdwg.org/

2. Use of Terms

This section provides precise definitions for key terms and concepts related to PGR, as used throughout this deliverable. It serves as a reference to facilitate a shared understanding among all stakeholders, thereby reducing ambiguity and enhancing the document's overall coherence.

Accession	a sample of seeds, planting materials or plants representing a PGR resource, either a wild population, a landrace, a breeding line, or an obsolete or improved cultivar, which is conserved in a genebank. Each accession should be distinct and, in terms of genetic integrity, as close as possible to the sample provided originally[8].
Accession Number	A unique identifier that is assigned by the curator when an accession is entered into a gene bank. This identifier should never be assigned to another accession[9].
Biological Resource Centers (BRC)	An essential part of the infrastructure underpinning biotechnology. They consist of service providers and repositories of the living cells, genomes of organisms and information relating to heredity and the functions of biological systems. BRCs contain collections of culturable organisms, replicable parts of these (e.g. genomes, cDNA), viable but not yet culturable organisms of cells and tissues as well as data bases containing molecular, physiological and structural information relevant to these collections and relevant informatics[10].
Crop wild relatives (CWR)	Plant taxa closely related to crops (or any socio-economically valuable species), which may be crop progenitors and to which the CWR may contribute beneficial traits, such as pest or disease resistance, yield improvement or stability. They are generally defined in terms of any wild taxon belonging to the same genus (or closely related genera) as the crop. A more practical definition is based on the ease with which taxa cross with the crop or using taxonomic placement as a proxy, describes CWR as taxa that belong to Gene Pools 1 or 2, or Taxon Groups 1 to 4 of the crop[11].
Database	An organized set of interrelated data assembled for a specific purpose and held in one or more storage media[9].
Descriptor	Attribute, characteristic or measurable trait that is observed in an accession of a genebank[12].
Documentation	The organized collection of records that describe structure, purpose, operation, maintenance, and data requirements[9].
Ex situ conservation	The conservation of components of biological diversity outside their natural habitats[13]. It involves the location, sampling, transfer and storage of samples of the target taxa away from their native habitats or cultivation sites.[14]
Genebank	A facility dedicated to the conservation of genetic material *ex situ*. It conserves PGR collections under medium or long-term storage conditions, in the form of seeds in cold rooms, plants in the field, and tissues *in vitro* or cryopreserved[8].

[8]https://www.fao.org/fileadmin/user_upload/wiews/docs/Metadata-02-05-01_PGR.pdf
[9]FAO. 2014. Genebank Standards for Plant Genetic Resources for Food and Agriculture. Rev. ed. Rome.
[10]OECD, 2007
[11]Maxted *et al.*, 2006
[12]Bioversity International, 2007
[13] CBD, Art.2
[14]Maxted *et al.,* 1997
[15]https://www.fao.org/4/Y2775E/Y2775E00.htm

Genetic resources conservation	The conservation of species, populations, individuals or parts of individuals, by *in situ* or *ex situ* methods, to sustain a diversity of genetic materials for present and future generations[15].
Genotype	The genetic constitution (gene makeup) of an organism; The pair of alleles at a particular locus, e.g., AA, Aa or aa; The sum total of all pairs of alleles at all loci that contribute to the expression of a quantitative trait [14].
In situ conservation	The conservation of ecosystems and natural habitats and the maintenance and recovery of viable populations of species in their natural surroundings and, in the case of domesticated or cultivated species, in the surroundings where they have developed their distinctive properties[13]. It involves the location, designation, management and monitoring of populations of the target taxa in their native habitats or cultivation sites.[14]
Landrace	A landrace is a dynamic population of a cultivated plant species that has historical origin, distinct identity and lacks formal crop improvement, as well as often being genetically diverse, locally adapted and associated with traditional farming systems and often has cultural associations[16].
Passport data	Basic information about the origin of an accession, such as details recorded at the collecting site, pedigree or other relevant information that assists in the identification of an accession[9].
Phenotype	The external appearance of a plant that results from the interaction of its genetic composition (genotype) with the environment[9].
Plant genetic resources (PGR)	Defined in the International Undertaking on Plant Genetic Resources (FAO, 1983) to mean the reproductive or vegetative propagating material of the following categories of plants: (i) cultivated varieties (cultivars) in current use and newly developed varieties; (ii) obsolete cultivars; (iii) primitive cultivars (landraces); (iv) wild and weed species, near relatives of cultivated varieties; and (v) special genetic stocks (including elite and current breeder's lines and mutants).[14] The genetic material of plants, which is of value as a resource for present and future generations of people[17].
PGR/ Germplasm collection	A systematically organized assemblage of plant genetic materials maintained for the purposes of conservation, research, breeding and utilization under defined conditions.
Sustainable use	Use of resources in a way and at a rate that does not lead to the long-term degradation of the environment, thereby maintaining its potential to meet the needs and aspirations of present and future generations[18].
Wild food plants (WFP)	Non-cultivated plant species that are harvested from the wild to be consumed as food or drink[19].

[16]Maxted et al., 2020
[17]IPGRI, 1993
[18]Glossary of terms to negotiators or MEAs, 2007
[19] Teixidor-Toneu *et al.,* 2023

3. Current State of PGR Data Landscape

Effective management and utilization of PGR involve navigating complex and diverse multidisciplinary data. Over the years, substantial efforts have been directed toward the systematic collection, conservation, characterization, evaluation, and documentation of PGR worldwide (Frankel & Bennett, 1970; Maxted *et al.,* 1997: Engels & Visser, 2003; Gepts, 2006; Khoury *et al.,* 2010; Engels & Ebert, 2021; Volk *et al.,* 2021; Lusty *et al.,* 2021). These efforts have generated distinct types and varied amounts of data, each requiring tailored approaches (Weise *et al.,* 2020). In addition, technologies that generate and analyse large quantities of phenotypic, genetic, and environmental data have evolved rapidly, especially with the advent of the omics revolution (Volk *et al.,* 2021). This technology-based data deluge has made the PGR data landscape much more information-rich. Adding another layer of complexity to the landscape are the diverse actors (viz. scientists, breeders, educators, and other stakeholders) who generate, access, integrate, synthesize, and utilize widely dispersed PGR data across various platforms, spanning organizational and international boundaries.

The circumstances described herein reflect the current state of the PGR data landscape along with significant advancements and enduring challenges that PGR practitioners have to contend with as a matter of routine. Understanding these complexities is, therefore, crucial in developing flexible and robust data frameworks that can accommodate diverse and heterogeneous data sets, institutional capacities, technological advancements, and varied stakeholder needs.

3.1 Stakeholders Generating and Using PGR-Related Data

3.1.1 Genetic Resource Centers (GRC)

Historically, *ex situ* GRC are specialized facilities dedicated to the conservation and management of PGR outside of their natural habitats. While *ex situ* conservation does not replicate the natural evolutionary processes occurring *in situ*, it became the principal approach to PGR conservation due to its effectiveness, the extensive range of genetic materials it can safeguard and the lack of effective exemplars of *in situ* / on-farm conservation (Maxted *et al.*, 2020). Here the term GRC is preferentially used over genebank, as is argued by Maxted et al. (2016), using GRC implies a more comprehensive PGR role than purely genebanking. Although the GRC staff role may vary from country to country, the existing genebank's remit (ECPGR European Genebank Managers Network, 2024; https://www.ecpgr.org/about/genebank-managers-network) is extensive including (a) engaging in international, national and local policy development, (b) national conservation planning, (c) target population national network management, (d) target population characterization and evaluation, (e) ensuring user access to in situ conserved resources (via the ex situ backup sample), and even (f) providing leadership of the PGR *In Situ* Population Management Committee (see PRO-GRACE Deliverable 1.3). The extended role implied by using GRC would substantially and positively raise the genebanks role in conservation, food security and ecosystem services provision. It would also mean as argued by Maxted and Magos (2023) a potential doubling of the diversity available for breeders and other stakeholder's use, which must be regarded as an existential genebank core activity.

Ex situ conservation techniques employed in GRC are determined by the specific biological characteristics of each PGR accession. These techniques may include storing orthodox seeds at low temperatures, maintaining living plants in fields, orchards or greenhouses, storing plant materials under slow growth conditions *in vitro*, or using cryopreservation techniques for long-term conservation of plant materials (FAO, 2014). Globally, there are more than 800 *ex situ* GRC, collectively conserving around 5.4 million accessions (FAO, 2022). Notably, approximately 79% of these accessions are conserved as seeds, while the rest are maintained in fields and *in vitro*.

Many efforts to conserve PGR are implemented at the national level by a range of organizations and institutions within tailored national programs (Frison & Demers, 2014). These national GRC, which are adapted to their specific agricultural and ecological contexts, form the backbone of the global *ex situ* conservation network, and reflect the commitment of individual countries to preserving their agricultural heritage and supporting global food security (Hodgkin *et al.*, 2012). According to the 2021 report on Sustainable Development Goal (SDG) Indicator 2.5, 4,872,408 accessions are conserved in base collections under medium- and long-term storage conditions in national genebanks in 115 countries (FAO, 2022). In Europe alone, there are more than 400 GRC, each employing one or more conservation strategies and often integrating multiple approaches to optimize conservation of over 2.1 million accessions (EURISCO[20], as of July 2024). Among them are the Leibniz Institute of Plant Genetics and Crop Plant Research (IPK)[21] in Germany which maintains over 150,000 accessions; the Centre for Genetic Resources (CGN)[22] in the Netherlands, managing nearly 24,000 accessions from over 150 countries; and the Spanish Plant Genetic Resources Centre (CRF)[23] in Spain, which holds more than 75,000 accessions; while the Millenium Seed Bank at the Royal Botanic Gardens Kew stores 98,567 seed collections of 39,681 species sourced from 190 countries[24].

Regional collaborations further enhance the effectiveness of *ex situ* conservation efforts by complementing and supporting the work of national programs. The Nordic Genetic Resource Center (NordGen)[25] serves Denmark, Finland, Iceland, Norway, Sweden Faroe Islands, Greenland and the Aland Islands, maintaining about 33,000 accessions of approximately 450 different plant species. The coordination of GRC across Europe is facilitated by the European Cooperative Programme for Plant Genetic Resources (ECPGR)[26]. Founded in 1980, ECPGR has been central to coordinating technical activities and collaborative frameworks involving most European countries. Through initiatives like the European Search Catalogue for Plant Genetic Resources (EURISCO) (Weise *et al.,* 2017; Kreide *et al.,* 2019; Kotni *et al.,* 2022) the European Genebank Integrated System (AEGIS)[27] (ECPGR, 2009; Engels & Maggioni, 2018), and the European Genebank Managers Network[28], ECPGR ensures the comprehensive documentation and integration of plant genetic resources across the continent. It also supports crop-specific and thematic working groups that develop and share innovative methods, tools, concepts and best practices. Additionally, it spearheaded the establishment of the European Evaluation Network (EVA) [29] for PGRFA to promote a standardized approach to evaluating and utilizing these resources in research and breeding programs (ECPGR, 2021). At the international level, the genebanks of the Consortium of International Agricultural Research Centers (CGIAR) (viz., AfricaRice, Bioversity International, CIAT, CIMMYT, CIP, ICARDA, ICRAF, ICRISAT, IITA, ILRI, IRRI), alongside the World Vegetable Center (WorldVeg) and the International Center for Biosaline Agriculture (ICBA), manage germplasm collections on behalf of the global community. These collections predominantly contain materials that are in the public domain, governed by legal agreements with the International Treaty on Plant Genetic Resources for Food and Agriculture (ITPGRFA), and primarily represent species listed in Annex I of the Treaty[30].

Inevitably, *ex situ* GRC generate vast amounts of heterogeneous data due to the diverse genetic resources they conserve and their extensive scope of conservation activities. The volume of data generated is significant, with each of the millions of accessions associated with information about its origin, phenotypic profile and genetic characteristics among others (van Etten *et al.,* 2023).

[20] http://eurisco.ecpgr.org
[21] https://www.ipk-gatersleben.de/en/research/genebank
[22] https://wur.nl/cgn
[23] https://www.inia.es/en-en/units/Institutes%20and%20Centres/CRF/Paginas/Home.aspx
[24] https://www.kew.org/science/collections-and-resources/research-facilities/millennium-seed-bank
[25] https://www.nordgen.org/our-work/genebank/
[26] https://www.ecpgr.org/
[27] https://www.ecpgr.org/aegis
[28] https://www.ecpgr.org/about/genebank-managers-network
[29] https://www.ecpgr.org/eva
[30] https://www.fao.org/plant-treaty/areas-of-work/the-multilateral-system/annex1/en/

Data collection methods range from traditional field observations to advanced genomic sequencing techniques, each with distinct standards and practices (Engels & Ebert, 2021). Curation practices also differ widely, with some centers maintaining highly detailed and meticulously curated datasets, while others may have more basic records due to resource constraints. Reporting and sharing of curated data also vary significantly among GRC. Some centers have advanced systems for regular reporting and making their data accessible to the global community through databases and online platforms (Opperman *et al.,* 2015; Postman *et al.,* 2010). Other centers may have less sophisticated reporting mechanisms, sometimes limiting data sharing to local or national databases due to resource limitations.

These differences in data collection, management, curation, reporting, and sharing practices reflect the unique contexts, resources, and capacities of each GRC. Despite these efforts, the lack of adequate information on many accessions and limited access to available data remain pervasive issues, resulting in significant gaps in publicly accessible PGR information (Frison & Demers, 2014; Halewood *et al.*, 2018).In addition to *ex situ*, *in situ* conservation efforts, including genetic reserves, protected areas, and on-farm conservation, play an equally vital role in conserving PGR within their natural habitats or on-farm (Maxted *et al.,* 1997; Maxted *et al.,* 2002; Maxted *et al.*, 2020). Europe, in particular, has made significant strides in this area through various initiatives and projects at the national level, aimed at enhancing the conservation of CWR, WFP and landraces (LR). Alongside national efforts, several EU-funded projects are working towards mainstreaming *in situ* conservation. Initiatives such as PGR Forum[31], PGR Secure[32], Farmer's Pride[33] and Dynaversity[34] focused on CWR and LR inventories, creating comprehensive networks of conservation sites and integrating efforts across Europe. For instance, Farmer's Pride has developed a European CWR priority list of 863 taxa related to human and animal food crops (Kell *et al.*, 2005).

Although these initiatives have led to substantial progress, practically *in situ* conservation implementation is limited, very few genetic reserves or active *in situ* conservation projects have been implemented and there are remaining gaps in comprehensive data on genetic diversity within natural habitats and the effectiveness of existing conservation measures. *In situ* conservation efforts are expected to generate a wide range of data, including genetic diversity within and among populations of conserved plant species, ecological interactions, population dynamics, habitat characteristics, threat assessments, and long-term monitoring data. Despite the challenges in implementing *in situ* conservation, Maxted and Magos (2023) suggest that if properly implemented, it has the potential to at least double the diversity available for breeders and other users.

3.1.2 Research Institutions and Universities

Numerous universities and research institutions are at the forefront of generating PGR characterization and evaluation data. These institutions, engaged in fundamental research, translational studies, pre-breeding and breeding programs, maintain medium- to short-term active collections that are crucial for a wide range of scientific activities.

For these institutions, research objectives vary extensively, reflecting the broad scientific scope within the field of PGR. A significant focus is on understanding crop genetic diversity, domestication, and evolutionary history to inform the identification and selection of appropriate genetic panels for genome-wide association studies (Dong *et al*., 2023; Rabieyan *et al*., 2023; Wang *et al*., 2023; Alam and Purugganan, 2024). Some institutions are dedicated to improving specific crops by enhancing desirable traits such as yield or resistance to pests (Pathirana & Carimi, 2022). Others explore the genetic diversity found in crop wild relatives, aiming to identify beneficial traits that can be introduced into cultivated varieties (Brozynska *et al.,* 2016; Tirnaz *et al.,* 2022).

[31]European Crop Wild Relative Diversity Assessment and Conservation Forum

[32]http://pgrsecure.org/

[33]https://more.bham.ac.uk/farmerspride/

[34]http://dynaversity.eu/

Additionally, trait-specific research, such as studies on drought tolerance (Li *et al.,* 2019; Wang *et al.,* 2021) disease resistance (Hinterberger *et al.,* 2022), and nutrient efficiency (Ali *et al.,* 2018), produces specialized data sets. This variation in research priorities results in a wide range of data types, from detailed genetic sequences to phenotypic observations and ecological interactions. Many of these data types (e.g. phenotyping data) lack a dedicated international archive. Even when archives exist, data deposition can be challenging and often suffers from poor metadata documentation. Each data type requires different collection methods, storage formats, and analysis techniques, which are not always compatible with those used by other institutions (Deng *et al.,* 2023). Different experimental approaches, such as phenotypic assessments and various genotyping techniques, produce data that vary widely in conditions, scales, and formats. For example, genotyping may involve a range of sequencing platforms, from traditional Sanger sequencing to next-generation sequencing (NGS) technologies like Illumina and Oxford Nanopore (Yang *et al.,* 2020; Scossa *et al.,* 2021; Hu *et al.,* 2021).

The inherent differences in mandates and objectives also mean that data management practices are often tailored to specific institutional needs. Consequently, diverse datasets are frequently managed through disparate systems, including offline databases, proprietary software, and localized online repositories that are primarily accessible to their own researchers. This siloed approach limits data sharing and hinders broader comprehensive, integrative analyses across different datasets. Furthermore, the challenge of centralizing data is compounded by the difficulty in depositing and updating information in centralized archives, which requires coherent metadata and consistent management practices. In most cases, projects with limited timeframes often lead to the creation of temporary databases that are decommissioned or archived once the project concludes, leaving valuable data without long-term accessibility or sustainability. Without robust data repositories, management plans, and sustainability measures, short-term initiatives frequently fall short in terms of maintaining and utilizing their data effectively. As a result, while universities and research institutions make substantial contributions to the PGR data landscape through their specialized research efforts, much of the valuable data they generate remains underutilized and inadequately integrated within the broader scientific community.

3.1.3 Non-Governmental and Civil Society Organizations

Several non-governmental organizations (NGOs) and civil society organizations (CSOs) are actively involved in various PGR-related initiatives. One notable example is ProSpecieRara[35] in Switzerland, an organization dedicated to preserving the genetic diversity of plants and animals. ProSpecieRara engages in activities such as maintaining heirloom varieties and promoting their use in modern agriculture. In Austria, ARCHE NOAH[36] supports *in situ* conservation by creating a network of seed savers and advocating for traditional and heirloom varieties in agricultural practices. Similarly, Réseau Semences Paysannes[37] in France is a network of farmers and organizations that promotes the use of traditional seeds and farming practices. Meanwhile, Rete Semi Rurali (RSR)[38], comprising over 30 member-associations, was established to tackle the challenges confronting traditional agricultural systems in Italy. RSR works closely with local farmers to conserve traditional crop varieties on-farm, facilitates seed saving and exchange, and manages agro-biodiversity zones. It also advocates for regulatory frameworks and agricultural policies that empower farmers to reclaim control over seed development and use. In Czech, the Czech Association of Nature Protectors unites more than 25 enthusiastic owners of old temperate fruit orchards. Among its members, EC Meluzína leads the initiative on saving regional varieties of fruit trees, which aims to conserve high-stem meadow orchards, tree rows and isolated trees within Czech landscape EC Meluzina runs a database of ancient fruit landraces and germplasm plots[39], encompassing over 6,000 fruit trees that have been documented, identified and geospatially localized. Their data is integrated into the national documentation system, GRIN Czech. Furthermore, the Czech Genebank is preparing to collaborate with EC Meluzina to enhance on-farm conservation of fruit genetic resources.

[35] https://www.prospecierara.ch/
[36] https://www.arche-noah.at/
[37] https://www.semencespaysannes.org/
[38] https://rsr.bio/
[39] http://www.plantsdata.com/genofondy.aspx

At a pan-European scale, the European Coordination Let's Liberate Diversity (ECLLD)[40] involves a comprehensive collaboration among diverse NGO stakeholders, including farmer organizations, researchers and seed networks. This international non-profit organization aims to develop and promote the dynamic management of cultivated biodiversity on farms and in gardens, facilitate knowledge exchange, and influence policies to ensure the sustainable use and conservation of PGR. Founded in 2005 and formally established in 2012, ECLLD currently operates across 20 European countries with a robust network of 21 members and 170 national organizations.

These NGOs/CSOs are likewise heavily involved in managing community seedbanks (CSB), often referred to as seed libraries or seed reserves. Through CSB and participatory breeding programs, these organizations collect extensive agronomic data, including detailed records on crop performance such as yield, growth habits, and responses to biotic and abiotic stresses. Additionally, they gather data on farmer preferences, which provides insights into the traits most valued by farmers, such as taste, yield, and ease of cultivation. NGOs and CSOs also collect valuable ethnobotanical information which captures the traditional uses of conserved plant varieties, local cultivation practices, and historical cultivation data. This holistic data collection enriches the overall understanding of PGR, including how different cultures have used and valued these resources over time, and aligns conservation practices with local needs and traditions.

While these non-profit organizations have demonstrated innovative approaches to data generation and management, the resources available for these activities can be limited. This affects the extent to which data can be systematically collected, curated, and stored. The lack of advanced technical infrastructure or expertise can make it challenging to implement comprehensive data management systems. As a result, while rich datasets are generated, they may not always be fully utilized or easily accessible.

3.1.4 Private sector

Over the years, the private sector has played a multifaceted role in PGR stewardship and use through various initiatives and collaborations. With the rise of biotechnology and growing importance of intellectual property rights (IPR) in agriculture, breeding companies have increasingly established their own genebanks. Recognizing the strategic necessity of having direct control over PGR, these companies began this initiative to ensure a reliable and consistent source of diverse germplasm, reducing dependence on external sources that could be subject to political, regulatory or logistical constraints (Engels *et al.*, 2024). Naturally, to protect their investments and maintain a competitive edge, these companies often restrict access to their genetic materials and associated data, limiting their availability to external researchers and public institutions.

Even with these restrictions, there are significant collaborations where the private sector has partnered with public institutions to facilitate broader utilization of PGR to ensure that the benefits of genetic diversity are shared across sectors. For instance, the French Network of Grapevine Repositories[41] (Réseau des Conservatoires de la Vigne en France), which holds one of the largest and most diverse collections of grapevine varieties in the world, represents a collaborative effort involving private nurseries, winegrowers and winemakers, public institutions such as the National Research Institute for Agriculture, Food and Environment, universities and other agricultural research centers, and key NGOs and biodiversity organizations. Another notable example is Pro-MAIS[42], a non-profit breeders' organization that aims to advance the conservation, study and genetic improvement of maize. In addition to these collaborations, private companies engage in initiatives to help regenerate PGR conserved in public institutions. For example, in CGN Netherlands[43], private seed companies such

[40]https://liberatediversity.org/

[41]https://bioweb.supagro.inra.fr/collections_vigne/Home.php?l=EN

[42]http://pro-mais.org/

[43]https://www.wur.nl/nl/onderzoek-resultaten/kennisonline-onderzoeksprojecten-lvvn/centre-for-genetic-resources-the-netherlands/about-cgn-1/partnerships.htm

as Enza Zaden and Rijk Zwaan support germplasm regeneration and multiplication. This collaboration not only helps maintain the viability and availability of these resources but also ensures that public genebanks can continue to serve as vital repositories for genetic diversity. Additionally, they play a significant role in generating and utilizing PGR-related data (Ebert *et al.,* 2023). Despite access restrictions to proprietary datasets, several mechanisms enable access to and use of private sector data. Seed companies and biotechnology firms frequently engage in collaborative research agreements and public-private partnerships with academic institutions, research organizations, and governmental bodies. These collaborations typically involve sharing specific datasets in exchange for research findings and advancements (Ebert *et al.,* 2023). For example, a private company may provide access to its genetic data for a particular crop to support a university's breeding program, with the understanding that any resulting innovations will be shared.

Some companies also participate in open innovation initiatives. These initiatives involve making specific datasets publicly available to encourage innovation and collaboration. Open innovation platforms may provide access to non-sensitive data while protecting proprietary information. Furthermore, they contribute to industry-wide data repositories or consortia aimed at addressing common challenges. These collaborative platforms allow multiple stakeholders to pool resources, share data, and develop collective solutions. One example is the International Wheat Genome Sequencing Consortium (IWGSC) [44], of which Syngenta, a leading global agribusiness company, is an active member. Syngenta has contributed genetic resources and expertise to help sequence the wheat genome. Their participation in the consortium has provided them access to the collective datasets and advanced genomic tools developed by the IWGSC.

At the regional level, the European PGRFA Evaluation Network (EVA) represents a strategic public-private partnership aimed at the advancement of systematic evaluation and utilization of PGR. EVA integrates the resources and expertise of national genebanks, research institutions, private breeding companies and farmers. It conducts collaborative projects, including participatory plant breeding, to generate standardized phenotypic and genotypic evaluation data for numerous crop accessions conserved in European genebanks. Currently, the network operates through six crop-specific networks focusing on both cereal and vegetable crops and involving 56 private breeding companies (from small cooperatives to large multinationals), 63 research institutions and 32 genebanks.

[44]https://www.wheatgenome.org/

3.2 Data Types

Figure 1 provides an overview of the various types of data generated from the study, conservation, and use of PGR. These data types differ significantly in scope and complexity, highlighting the multifaceted nature of this field. They range from intricate genetic sequences that detail the molecular structure of plants to broader phenotypic traits that describe their physical characteristics and responses to environmental conditions. Focusing this deliverable on specific descriptors, which are essential for systematically documenting, identifying, and evaluating PGRs, allows for the creation of a clear and practical framework that addresses the key requirements of conservationists, researchers, breeders, and other relevant stakeholders. This approach is intended to ensure that the most immediately valuable information to the users are captured, readily available and easily shared across PGR-focused efforts, thereby facilitating scientific research and promoting sustainable use of PGR.

It is important to recognize, however, that the comprehensive conservation, management and utilization of PGR require a more expansive array of data beyond the descriptors highlighted in this deliverable. Effective PGR stewardship, both *ex situ* and *in situ*, relies on additional data types such as legal and access information, material management and monitoring data (e.g. availability, viability and regeneration data, health and phytosanitary data, distribution and utilization data, *in situ* conservation plans, land use and management, etc.,), socio-agroecological data, and cultural and ethnobotanical knowledge.

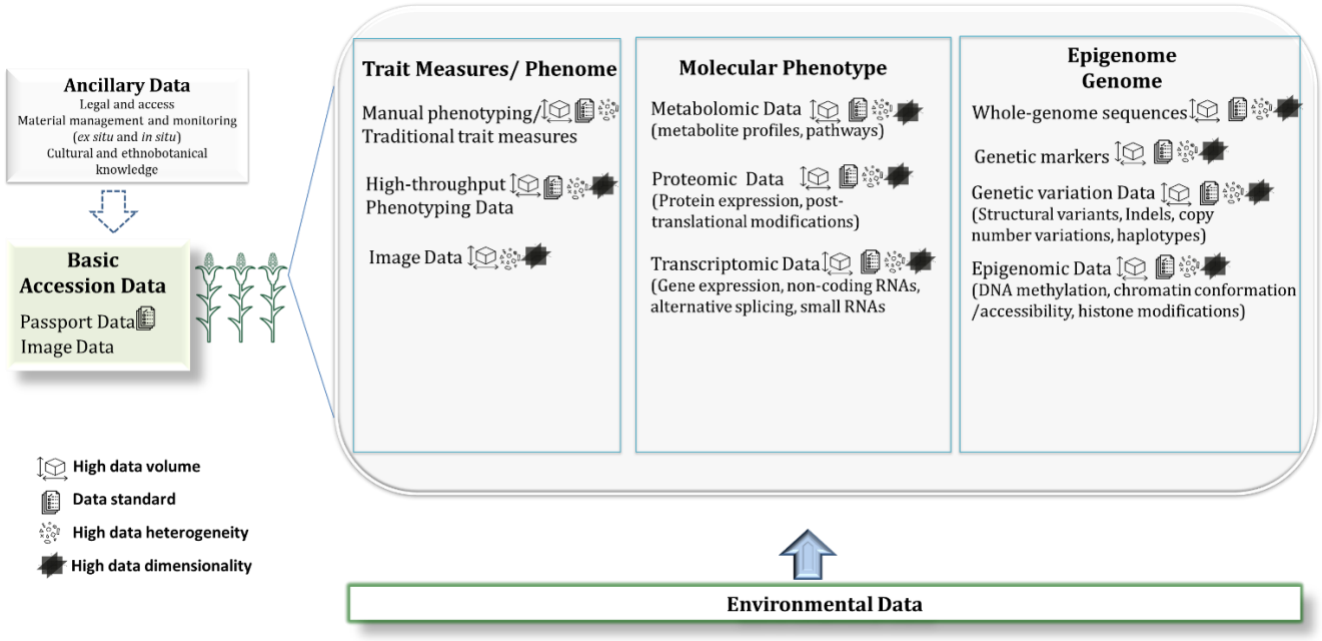

Figure 1. Types of data generated from the study, conservation, and use of PGR (i.e. characterization and evaluation).

3.2.1 Passport Data

Passport data are the essential baseline information collected about PGR at the time they are actively conserved, including details on the plant material's identity, PGR population description and provenance, and key characteristics (Engels & Visser, 2003; Maxted *et al.*, 2020). By capturing this critical metadata, passport information facilitates resource predictive description, effective tracking, and informed utilization of genetic diversity within genebanks and other conservation facilities (Gepts, 2006; Khoury *et al.,* 2010).

The evolution of passport data in PGR management has been a dynamic and progressive process, significantly advancing since its early conceptual stages in the mid-20th century. During this period, the establishment of genebanks and increasing recognition of the need for systematic conservation of genetic diversity highlighted the urgent need for enhance passport data quality, prompting the development of more advanced data recording practices (Frankel & Brown, 1984). Initially, passport data collection focused on a few simple descriptors, including accession number, origin, and species name. These rudimentary records were fundamental for the basic cataloguing and management of genetic collections. However, these early data recording practices were limited in scope and detail. They provided essential, yet minimal, information often insufficient for comprehensive genetic resource management, research applications and facilitating utilization. Specifically, the lack of detailed environmental, phenotypic, and ethnobotanical descriptive data hindered the full utilization of genetic materials for breeding programs and other scientific studies. As the importance of genetic diversity in agriculture and ecological research became more widely acknowledged, the scope of passport data began to expand. Several factors, including advancements in information technology and a growing recognition of the multifaceted value of genetic resources, drove this expansion. The introduction of computerized databases in the late 20th century enabled the storage and management of larger datasets, allowing for more comprehensive and detailed record-keeping.

A significant milestone in the standardization and enhancement of passport information was the development of the List of Multi-Crop Passport Descriptors (MCPD). Designed for use across multiple crops, the MCPD greatly facilitated the global exchange and comparison of passport data. The first version of MCPD was created in 1996 through a collaborative effort between Bioversity International[45] and the Food and Agriculture Organization (FAO). By 2001, a revised edition of the MCPD[46] was developed (Alercia *et al*., 2001), featuring 28 descriptors. This updated version provided comprehensive explanations of each descriptor's content, coding scheme, and suggested field names. The enhancements ensured greater consistency in the recording and reporting of passport information. In 2012, MCPD V.2[47] was introduced following extensive consultations with over 300 individuals from 187 institutions across 87 countries, reflecting the need for more detailed and inclusive data (Alercia *et al.,* 2012). While no major revisions were made, providing greater flexibility with existing descriptors and refining them for more detailed and specific data capture was essential. Improved definitions were introduced to reduce ambiguity, aligning fields with current international standards and enhancing compatibility with global databases. For example, the revision considered technological advancements, such as the use of global positioning system (GPS) tools, to allow for the precise recording of geographic coordinates in decimal degrees, aligning with international geographic standards and making it easier to integrate and share data globally. Additionally, information about the status of accessions within the Multilateral System (MLS) of Access and Benefit-Sharing (ABS) was included, ensuring compliance and facilitating benefit-sharing agreements. This update addressed the evolving documentation demands and the requirements set by international agreements, notably the ITPGRFA (Alercia *et al.,* 2012). In 2015, the MCPD V.2.1[48] introduced the Permanent Unique Identifier (PUID) to address the need for a global unique identifier or persistent identifier. This addition facilitated the integration of germplasm data by enabling the linkage required to identify accessions and other genotype entities across different information systems (Alercia *et al.,* 2015).

Characteristics of Passport Data

(1) Data size. The size of passport data is relatively small by modern standards. Each record typically contains basic descriptors like taxonomic identification, geographical coordinates, and collection details. On average, a single passport data record is around 1-2 kilobytes in size, including all relevant details. Therefore, even with large collections, the data size remains within a manageable range

[45]Currently known as the Alliance of Bioversity International and the International Center for Tropical Agriculture (CIAT); previously, the International Plant Genetic Resources Institute (IPGRI) (1991-2006

[46]https://alliancebioversityciat.org/publications-data/faoipgri-multi-crop-passport-descriptors-mcpd

[47]https://alliancebioversityciat.org/publications-data/faobioversity-multi-crop-passport-descriptors-v2-mcpd-v2-june-2012

[48]https://alliancebioversityciat.org/publications-data/faobioversity-multi-crop-passport-descriptors-v21-mcpd-v21-december-2015

Case in Point 1. What's in a Name? Taxonomic Ambiguity and Evolution: The Case for TaxonID

Recent advancements in taxonomy underscore the critical need for precise taxonomic identifiers (TaxonID). The shift from traditional morphological methods to advanced molecular techniques has revolutionized the field, making accurate classification of species, subspecies and varieties essential. These developments have led to the reclassification of numerous species, the discovery of cryptic species and a deeper understanding of evolutionary relationships (Maltsev and Erst, 2023, Pritchard *et al*., 2022; Thiele *et al*., 2021). By **integrating TaxonID into the MCPD**, these scientific advancements can be effectively documented, ensuring that the data remains current and maintains its scientific integrity. Accurate species identification is fundamental to advancing conservation efforts, enabling the development of precise and effective strategies tailored to protect specific species and populations most in need (Dempewolf *et al*., 2014). Integrating TaxonID will also enhance data interoperability across various databases and platforms, including those that handle a wide range of genetic data such as those within the INSDC, as well as broader biodiversity platforms like Global Biodiversity Information Facility (GBIF). This interoperability is particularly important for *in situ* conservation, where data from various sources, including field observations, genetic studies, and ecological surveys, need to be combined to create comprehensive biodiversity profiles.

It is worth noting that TaxonID has already been integrated and made mandatory in the Minimum Information about a Plant Phenotyping Experiments (MIAPPE) and genomic standards, such as the Minimum Information about any (x) Sequence (MIxS) standards. It is also required for data submission to INSDC repositories. This mandatory integration ensures the standardization and reliability of taxonomic data across various research and data management platforms, facilitating the seamless exchange and validation of taxonomic information.

To implement the inclusion of TaxonID, it is recommended to establish partnerships with taxonomic experts and broader consortia and utilize reputable taxonomic databases. The NCBI Taxonomy Database[49] provides comprehensive and authoritative taxonomic information and is widely used in genomic and biological research. GBIF's Taxonomic Backbone[50] offers a globally recognized standard for taxonomic data, integrating information from multiple sources to provide a consistent and authoritative list of species. The Catalogue of Life (CoL)[51] compiles taxonomic data from numerous sources, creating a unified and consistent taxonomic index with stable and widely accepted identifiers. The Integrated Taxonomic Information System (ITIS)[52] provides authoritative taxonomic information and unique identifiers for a wide range of organisms, supporting reliable data integration.

One effective strategy to integrate TaxonID into the MCPD involves utilizing an **attribute-value pair structure**. In this structure, each piece of information is represented as a pair, where the attribute denotes the type of information and the value denotes the actual data corresponding to that attribute. For instance, consider integrating TaxonID for Wheat (*Triticum aestivum*). When querying the NCBI Taxonomy Database, we retrieve a TaxonID of 4565, while querying CoL provides a TaxonID of 5944Q. These can be represented using attribute-value pairs: TaxonID_Source: NCBI and TaxonID_Value: 4565, and TaxonID_Source: CoL and TaxonID_Value: 5944Q. Here, "TaxonID_Source" indicates the database source, and "TaxonID_Value" provides the unique identifier assigned by that source. This structure allows for the seamless integration of multiple identifiers from different databases into the MCPD schema, enabling the representation of complex, multi-source taxonomic data in a straightforward manner. The attribute-value pairs can be easily queried, updated, and validated, ensuring that the MCPD records remain accurate and consistent across various taxonomy systems. Achieving broad consensus and acceptance of this recommendation, however, will require thorough discussion and careful consideration within the community.

[49] https://www.ncbi.nlm.nih.gov/taxonomy

[50] GBIF Backbone Taxonomy. Checklist dataset https://doi.org/10.15468/39omei

[51] https://www.catalogueoflife.org/

[52] National Museum of Natural History, Smithsonian Institution (2023). Integrated Taxonomic Information System (ITIS). Checklist dataset https://doi.org/10.5066/f7kh0kbk

(2) **Data Availability and Accessibility**. The availability of passport data varies widely among institutions. Many European GRC and organizations have made substantial progress in digitizing and making their passport data accessible online. However, smaller or regional institutions may lack the resources for comprehensive digitization efforts, limiting the availability of their data. Currently, EURISCO aggregates passport data for over 2 million accessions from 43 countries. Similarly, Genesys[53], the world's largest portal to information about crop diversity conserved in genebanks, provides extensive access to passport data. It is important to note, however, that data availability on these platforms is contingent upon submissions from each participating country. Legal and proprietary restrictions can also sometimes limit access, particularly for data governed by specific national or institutional policies.

(3) **Data Quality and Completeness.** The quality and completeness of passport data hinges on various factors from data collection methodologies to technological and infrastructural limitations. Despite the existence of international guidelines such as the MCPD, adherence to these standards is often inconsistent. The adoption of advanced data collection and management technologies is also uneven. In many countries there is limited access to modern tools such as GIS, mobile data collection apps, and digital databases, and the GRC staff may lack the skills to implement them effectively when they are available. This consequently hampers efforts to maintain high-quality and complete datasets. These underfunded institutions may likewise struggle with limited financial and human resource that affect their ability to collect, curate, and maintain comprehensive PGR passport data**.** In addition, human errors, which can arise from misinterpretation of data collection guidelines, transcription mistakes, or even typographical errors during data entry, can significantly impact data quality. Insufficient training or lack of experience may also lead to inaccurate or incomplete data entries. Furthermore, data completeness is frequently compromised by historical data gaps. Legacy data are often incomplete due to less rigorous data collection practices in the past. Retrospective data curation efforts, which involve revisiting old datasets (including field notes, collection logs), data cleaning, filling in missing information, and standardizing data according to current guidelines, are necessary to address these gaps, but they can be resource-intensive and technically challenging.

(4) **Data Complexity and Interoperability**. The complexity of passport data is relatively low compared to other PGR-associated datasets. It is primarily straightforward, consisting of standard fields describing basic information about a PGR accession. The main challenge arises when integrating passport data with other data types, such as environmental and multi-omics data, to gain comprehensive understanding of genotype-phenotype relationships, adaptive traits, and molecular mechanisms, among others. Accession numbers are intended to be unique identifiers for each PGR accession, which is crucial for the organization and management of passport data. Theoretically, these numbers should ensure that each accession can be distinctly tracked and referenced across various datasets. However, in practice, several factors often undermine this principle. Different institutions develop their internal numbering systems that might be unique within the institution but are not coordinated with other organizations. As a result, the same accession can be assigned different accession numbers in different databases. Even within a single institution, an accession may receive multiple identifiers due to separate departmental databases or research projects, compounded by the lack of standardized data protocols. The absence of a centralized, global registry for accession numbers exacerbates these issues, as there is no definitive reference to ensure each number is unique across all collections. These inconsistencies and issues in assigning accession numbers lead to data redundancy that can inflate estimates of genetic diversity and misinform conservation strategies. Additionally, integrating passport data with other PGR-associated datasets becomes highly problematic. Inconsistent identifiers prevent the effective linking of datasets, impeding comprehensive analyses and the extraction of meaningful insights.

[53]https://www.genesys-pgr.org/

Adopting PUID such as DOI, as stipulated in MCPD, offered a robust solution to these challenges. Unlike accession numbers, PUID provides a standardized, persistent means of identifying accessions, ensuring reliable referencing and data access over time. However, implementing PUID (e.g., DOI) comes with its own challenges. **Refer to Section 4 and Case in Point 3 for a more detailed discussion on unique identifiers.**

Meanwhile, legacy data, often recorded in inconsistent formats, further complicates integration efforts. Many legacy datasets were created before the establishment of standardized data protocols, resulting in significant variability in data formats, terminologies, and levels of detail. These datasets may differ in their metadata schemas, units of measurement, and descriptive terminologies, necessitating extensive efforts to reconcile and standardize the data before it can be integrated with other datasets. Standardizing legacy data involves converting it into compatible formats, harmonizing terminologies, and ensuring that data quality meets the requirements for integration.

3.2.2 Phenotypic Data

The ultimate goal of PGR conservation is not merely conservation *per se,* but the effective utilization of the conserved resource. Achieving such goal necessitates a rigorous process of characterization and evaluation (C&E). In genebanking parlance, characterization involves the systematic documentation of genetically controlled traits that are observable across different environments (e.g. flower colour). On the other hand, evaluation focuses on recording traits that only manifest under specific environmental conditions, such as resistance to biotic and abiotic stresses (Maxted et al., 2020). Phenotyping is a fundamental aspect of C&E. It refers to the detailed measurement and analysis of an organism's structural and functional qualities resulting from the complex interaction between its genotype and environment (Fasoula *et al.,* 2020). Phenotypic expression varies with the organism's developmental stage, with the same genetic makeup producing diverse phenotypes at various growth phases, from seedling to mature plant. Chemical modifications to DNA and histones, which do not alter the underlying DNA sequence but affect gene expression, also play a critical role in phenotypic variation (Pieruschka and Schurr, 2019). These epigenetic modifications, influenced by environmental factors and are sometimes heritable, add another layer of complexity to phenotypic expression (Dar *et al.,* 2022).

Phenotyping research has become increasingly interdisciplinary, drawing on fields such as genetics, agronomy, computer science, environmental science, bioinformatics and engineering (Pieruschka and Schurr, 2019). This interdisciplinary approach is necessary due to the diverse nature of phenotyping which involves: (1) multiple crops and varieties, each with its unique set of traits and genetic makeup: these crops can have hundreds to thousands of genotypes, leading to immense genetic diversity that needs to be captured and analysed; (2) experiments that are conducted for different contexts with different experimental designs and across varied experimental sites, including laboratories, greenhouses, and open fields: each of these settings presents different environmental conditions; (3) different methodologies and platforms, ranging from traditional manual measurements to high-throughput automated systems and (4) diverse datasets in terms of size, granularity, complexity, and dimensionality (Pieruschka and Schurr, 2019; Fasoula *et al.,* 2020; Volk *et al.,* 2021). Consequently, the lack of standardization in data structure, nomenclature, and standards across the different disciplines involved in plant phenotyping research complicates harmonization and data integration. Furthermore, due to the non-invasive nature of many experiments. researchers frequently modify experimental settings while trials are ongoing, leading to inconsistencies in data collection (Ugochukwu and Phillips, 2022). At the same time, technological advances, such as drones, automated imaging techniques, hyperspectral cameras, and machine learning algorithms, have significantly improved the ease of data collection, storage, and management (Rebetzke *et al.,* 2019; Fasoula *et al.,* 2020; Sheikh *et al.,* 2024). These technologies facilitate high-throughput phenotyping, capturing vast amounts of data quickly and with minimal human intervention. However, the resulting increase in data volume brings new challenges for data integration and meta-analysis. Managing and analysing these large, complex data sets require advanced computational tools and robust data management systems (Coppens *et al.,* 2017; Sheikh *et al.,* 2024).

In parallel, undertaking direct C&E recording *in situ* is much less straight forward under non-standardized natural or on-farm conditions. These environments lack the controlled settings of laboratories, greenhouses or other similar settings, resulting in variable and unpredictable conditions (e.g. weather patterns, land and crop management practices, pest pressures, broader ecological networks) that can significantly affect data collection. Determining who should perform the data collection is likewise a key consideration. Typically, this responsibility falls on *in situ* population managers (e.g., protected area managers, farmers, and other landowners), who are generally willing to participate in population management activities only if the additional resource costs are minimal. However, the process of collecting C&E data is resource-intensive and often requires skills and equipment that these individuals are unlikely to have. Given these constraints, these stakeholders may not be able to collect data comprehensively or as frequently as needed, which may consequently result in incomplete datasets that do not fully capture the range of phenotypic diversity present in the population. Data collection inconsistencies will also make it difficult to compile and compare information, analyse trends, identify patterns or make informed decisions across diverse ecosystems or within and across species. The logistical challenges of *in situ* phenotyping are further compounded by the need for long-term data collection to capture temporal dynamics and seasonal variations in plant traits. This necessitates sustained engagement and investment, which can be difficult to secure, particularly in resource-limited settings. Furthermore, integrating traditional/indigenous knowledge from local farmers and land managers can provide invaluable insights but requires careful documentation and validation to ensure accurate representation and utilization.

Nevertheless, the Minimum Information About Plant Phenotyping Experiments (MIAPPE)[54] emerged from a critical need to standardize metadata in the plant phenotyping domain, ensuring that data can be easily shared, integrated, and made discoverable and available in a machine-readable format. Initially, metadata documentation was handled independently by various phenotyping databases such as BreedBase[55] (Morales *et al.,* 2022), PIPPA[56], GnpIS[57] (Steinbach *et al.,* 2013; Pommier *et al.*, 2019), and the Plant Hybrid Information System (PHIS)[58] (Neveu *et al.,* 2018) each creating its own implicit, database-specific standards. This lack of standardization impeded data sharing and integration, highlighting the need for a unified approach (Krajewski *et al.,* 2015). In 2011, the European plant phenotyping community formally recognized these challenges and initiated discussions to develop a common framework. These discussions led to the formation of a consortium and the establishment of the MIAPPE working group, tasked with developing standardized guidelines for documenting plant phenotyping experiments (Krajewski *et al.,* 2015). The effort was supported by several significant EU funded projects, notably TransPLANT (2011-2015) and Elixir-Excelerate[59] (2015-2019) (Krajewski *et al.,* 2015).

The first version of MIAPPE (Ćwiek-Kupczyńska *et al.,* 2016) was released in 2016, marking a significant milestone. This initial version provided a structured approach to standardizing essential metadata categories such as study design, environmental conditions, and data acquisition methods. These guidelines were crucial for ensuring that phenotypic data could be accurately interpreted, reused, and integrated across different studies and platforms (Ćwiek-Kupczyńska *et al.,* 2016). In response to feedback from the research community and as the plant phenotyping community's needs became more sophisticated, MIAPPE version 1.1 (Papoutsoglou *et al.,* 2020) was released in 2019. This version provided better alignment with existing data standards (including MCPD), expanding the scope of the guidelines to include additional metadata elements and improving clarity and usability (Papoutsoglou *et al.,* 2020). The enhancements focused on making the guidelines more comprehensive and user-friendly, ensuring they could be easily integrated into existing data management workflows.

[54]https://www.miappe.org/
[55]https://breedbase.org/
[56]https://pippa.psb.ugent.be/
[57]https://urgi.versailles.inrae.fr/ephesis/ephesis/viewer.do;jsessionid=B90CC421E8CA21B2591E895AC6AEF87E#showForm
[58]http://www.phis.inra.fr/
[59]https://elixir-europe.org/about-us/how-funded/eu-projects/excelerate

Finally, MIAPPE version 1.1 was updated to align with the FAIR (Findable, Accessible, Interoperable, and Reusable) principles (Wilkinson *et al.* 2016) with the aim to enhance MIAPPE's utility for data sharing and integration, promoting a more open and collaborative approach to plant phenotyping research (Papoutsoglou *et al.,* 2020).

Characteristics of Phenotypic data:

(1) **Data size**- Phenotypic datasets are extensive owing to a multitude of variables: (1) the breadth of traits influenced by genetic, environmental and management factors (Hatfield and Walthall, 2015; Pieruschka and Schurr, 2019; Watt *et al.,* 2020); (2) diverse accessions across different plant species (Deng *et al.,* 2023); (3) continuous monitoring at high spatial and temporal resolution (Gill *et al.,* 2022; Zhang *et al.,* 2023); (4) use of advanced sensor, machine vision and automation technologies (Araus and Cairns, 2014; Pieruschka and Schurr, 2019; Yang *et al.,* 2020, Ninomiya. 2022); and (5) variety of experimental setups employed (Ninomiya, 2022; Papoutsoglou *et al.,* 2023). Traditionally, phenotyping has been conducted using classical, manual methods, which despite being tedious, labour-intensive and time-consuming (Watt *et al.,* 2020, Xiao *et al.,* 2022), have provided a foundational understanding of plant traits and their variability. This manual process produced smaller, less complex datasets typically recorded in spreadsheets or simple databases. Consequently, given the limited scale and resolution of manually collected data, this approach has resulted in a phenotyping bottleneck, limiting the functional analysis of key traits and impeding large-scale crop improvement (Smith *et al.,* 2021, Song *et al.,* 2021). Over the last two decades, the emergence of automated, high-throughput, and non-destructive phenotyping methods has revolutionized data collection. With cutting-edge imaging technologies, novel sensors in phenomobiles, gantry systems, and unmanned aerial vehicles (UAVs), phenotyping robots, and conveyor systems among others, these platforms have dramatically increased the scale, precision, and efficiency of phenotypic assessments (Furbank and Tester, 2011; Fahlgren *et al.,* 2015; Zhang and Zhang, 2018; Ninomiya *et al.,* 2019; Pieruschka and Schurr, 2019). These modern systems are capable of collecting petabytes of high-resolution data per growing season, providing fine-grained details of aboveground and belowground traits across various environmental conditions. Notably, imaging systems can acquire data in various spectral bands—visible, infrared, and hyperspectral—, while sensors monitor environmental parameters with high temporal resolution, often collecting data multiple times per minute. Machine vision technologies can process thousands of images daily. Consequently, the accumulation and subsequent analysis of extensive raw sensor data, processed datasets, and metadata, which detail experimental conditions and machine configurations, significantly contribute to the overall data volume.

(2) **Data heterogeneity, multidimensionality and complexity** - Phenotypic data is inherently complex and heterogeneous, reflecting the intricate nature of plant biology and diverse methods used to measure and analyse multi-factorial and multi-dimensional traits. These traits, including plant architecture, physiology, and performance-related characteristics, are observed across various biological scales (i.e., cell, tissue, organ, individual plant, plot, and field levels) and diverse spatial and temporal dimensions (Pieruschka and Schurr, 2019; Watt *et al.,* 2020; Papoutsoglou *et al.,* 2023). Phenotyping research can involve a limited number of accessions, or genetic populations observed through repeated measurements (e.g., diurnal, seasonal, or inter-annual phenotyping) or encompass multiple field locations, gathering extensive phenotypic data from both ground-based and aerial sensors (Watt *et al.,* 2020). Nevertheless, in theory, the number of possible phenotypic traits is almost limitless, as each trait represents a unique observational perspective on the plant phenotype (Li and He, 2024). These traits are quantified and characterized across a plethora of environments, ranging from controlled settings like growth chambers and greenhouses to expansive field-based phenotyping platforms. Each trait can be dissected and analysed at various levels, from molecular and cellular processes to whole-plant structure and physiology and interactions with

the ecosystem (e.g., diverse array of biotic and abiotic stresses that vary in type, space and time) (Araus and Cairns, 2014; Araus *et al.,* 2018; Smith *et al.,* 2021). Each scale provides different dimensions of information that need to be integrated for a comprehensive understanding of the phenotype. The temporal dynamics of plant development, encompassing ontogenetic processes and describing the sequential changes in plant traits from germination through maturity and senescence, add layers of complexity and dimensionality. Plants dynamically adjust their traits across different growth phases in response to the co-regulation of genetic and environmental cues (Sultan, 2000; Tao *et al.,* 2022; Li and He, 2024). The interaction between these factors leads to temporal variations and complex trait expressions that are often challenging to disentangle. Moreover, many phenotypic traits associated with fitness and performance, such as biomass, yield, responses to biotic and abiotic factors, and nutritional composition, are typically quantitative in nature and have multiple genetic determinants (i.e. quantitative trait loci (QTLs)) (Deng *et al*., 2023). The heterogeneity and complexity of phenotypic data is further compounded by methodological and measurement variability (Tardieu *et al.,* 2017). Different sensors and measurement techniques, each with its own set of advantages and limitations, can yield varying results, even when assessing the same trait. All these factors collectively contribute to the rich but challenging nature of phenotyping, underscoring the need for sophisticated analytical techniques, standardized protocols, trait correlation networks and integrated approaches that combine multiple data sources and advanced computational methods to accurately capture and understand the continuous, multifaceted and interpretive nature of phenotypic observations.

(3) **Data availability and accessibility.** Congruent with advancements in high-throughput phenotyping is the increasing challenge of managing burgeoning volume of datasets in ways that facilitate value extraction and ensure they are readily accessible and usable by different stakeholders. While varying environmental conditions can limit the direct reuse of phenotypic data, the long-term availability and accessibility of these datasets remain vital. Specific traits associated with certain genes, as reflected in current phenotypic data, can provide valuable insights across generations for the same genotypes. Despite a substantial volume of phenotypic data being available, its accessibility is often hindered by various barriers, including nonuniform data structure, nomenclature, and standards across different disciplines involved in plant phenotyping. Inadequate metadata documentation (e.g. descriptive, provenance, administrative and structural metadata) and annotation (e.g. trait definition using semantic standards, labelling and categorization) further exacerbates the situation. Yet, even meticulously standardized and well-described datasets, when confined solely to personal hard drives or local databases, remain inaccessible to the broader research community. To date, there is no single, universally recognized repository dedicated exclusively to plant phenotypic data. Nonetheless, there are various initiatives and databases that currently host phenotypic information. Aside from crop community (e.g. MaizeGDB, Legume Information System, Gramene, among others) and project-based databases (e.g. G2P-SOL, TRADITOM), EURISCO also hosts a vast array of phenotypic data as part of its extensive search catalogue for PGR in Europe. Advanced portals like GnpIS, BRIDGE portal (Konig *et al.*, 2020), and AgBase (McCarthy *et al.*, 2006) integrate phenotypic data with other data types, offering sophisticated tools for data analysis and accessibility. Moreover, collaborative platforms, networks and research infrastructures (RI), such as ELIXIR and EMPHASIS[60], have been pivotal in enhancing data accessibility. These platforms facilitate the exchange of datasets among researchers, supporting collaborative projects and promoting the dissemination of research findings. The effectiveness of these platforms, however, relies heavily on active participation and contribution from the research community. Without widespread engagement, even well-designed platforms can fall short in providing comprehensive accessibility. In recent years, significant efforts have been made to standardize phenotypic data management practices. The adoption of FAIR principles, adherence to standardized formats such as MIAPPE, and proper annotation using controlled vocabularies and ontologies have gradually improved the phenotypic data landscape.

[60]https://emphasis.plant-phenotyping.eu/

Case in Point 2. Understanding the Genetic Profiles of PGR Accessions

Effective PGR conservation and utilization hinge on a thorough understanding of their genetic profiles. The nuances inherent in these genetic profiles profoundly influence data collection, management, and curation processes. Different genetic profiles require specific data collection techniques to accurately capture the genetic diversity present in accessions. Consequently, a comprehensive grasp of these complexities is indispensable for GRC aiming to maintain and leverage the full spectrum of genetic diversity.

(1) **Heterozygous-Heterogeneous Accessions** are characterized by significant genetic variation both within individual plants and among the population. Multiple alleles are present at many loci within the genome, therefore each individual plant may carry different combinations of alleles, resulting in high levels of heterozygosity (presence of different alleles at a locus) within individuals and extensive allelic diversity across the population. Examples are landraces of outcrossing species (e.g. Maize) or populations that have not undergone controlled breeding and have adapted to local environments over generations. CWRs also often display significant genetic variability both within and between populations due to their adaptation to diverse environments and lack of selective breeding pressures.

(2) **Homozygous-Heterogeneous Accessions** consist of populations where each individual is genetically uniform but significant variation exists between individuals. Individuals are homozygous at most loci, but different individuals within the population may have different homozygous alleles. This results in low heterozygosity within individuals but high allelic diversity among the population. Ex. landraces of self-pollinating crops that are comprised of an assortment of different genotypes.

(3) **Heterozygous-Homogeneous Accessions** consist of individuals that are genetically similar, but each individual harbors heterozygosity at various loci. This results in a population where the genetic variation is within individuals rather than between them. Ex. hybrids and clonally maintained but originally outcrossing species (e.g. apples, grapes)

(4) **Homozygous-Homogeneous Accessions** are populations where individuals are genetically identical, with little to no genetic variation either within or between individuals. These accessions are often the result of self-pollination or vegetative propagation over multiple generations. Ex. Inbred lines and accessions derived from single seed descent (SSD).

While there is no specific descriptor in MCPD for these four genetic profiles**, information can be inferred from the descriptor, Biological Status (SAMPSTAT)**. Accurate SAMPSTAT information can guide the design of phenotypic C&E recording and facilitate targeted data collection strategies. For instance, accessions identified as wild relatives (which are likely to be heterozygous and heterogeneous) might require more detailed and repeated phenotypic measurements to account for within-population variability and can be sampled more extensively to capture genetic diversity.

Nonetheless, accurately representing within-accession heterogeneity (i.e. heterogeneous accessions) is a major challenge in GRC. The approach depends significantly on the application and the scientific question to be addressed. In GRC, where the conservation of maximum diversity is the primary goal, it is critical to adopt strategies that ensure the broadest possible genetic representation is preserved. **Should each unique genotype within an accession be meticulously recorded separately, or is it more practical to rely on composite data?** Both approaches offer distinct advantages and face notable drawbacks. Detailed genotype-specific records provide unparalleled precision but demand extensive storage and complex data collection and management protocols. On the other hand, composite data streamline information handling but risk oversimplifying or masking crucial genetic variations.

Moreover, the practical implications of using heterogeneous accessions present further complexities. **Should breeders and researchers prioritize mean characteristics of an accession, or should they delve into the full spectrum of traits available?** Contextual needs dictate optimal strategies: breeding programs may benefit from detailed assessments to pinpoint and select superior genotypes, whereas conservation efforts may emphasize conserving broader genetic diversity across populations.

To handle within-accession heterogeneity within MIAPPE (see table 5 for data element descriptions), the following approach may be used:

1. **Observation Unit Type and ID**: Utilize **"plant"** as the observation unit type, with each individual plant assigned a unique observation unit ID.
2. **Experimental Design:** Document **detailed subplot or block information** within the "Description of experimental design" section. This includes information on how plants are organized within the experimental setup, ensuring clarity on any spatial or environmental factors, if there any, influencing plant growth and development.
3. Observed Variables: Record **individual measurements or trait values** for each plant under the "Observed variables" section. This step ensures that data capture is granular, capturing the specific characteristics of each plant within the accession.

This approach enhances the granularity and specificity of data collection within MIAPPE, facilitating a more detailed understanding of the variability of multiple genotypes within a single accession.

3.2.2 Molecular Phenotypic Data

The concept of molecular phenotype refers to the extensive range of molecular-level characteristics that define the functional state of an organism. This includes in particular transcriptome, proteome, and metabolome data. The term "molecular phenotype" was introduced to encapsulate the spectrum of molecular expressions and interactions that characterize the physiological state of an organism, beyond the observable traits at the macroscopic level (de Vienne, 2022; Deng *et al.,* 2023).

Transcriptome data includes information about the complete set of RNA transcripts produced by the genome under specific conditions. This includes mRNA, rRNA, tRNA, and non-coding RNAs, reflecting the gene expression levels and regulatory mechanisms active within a cell or tissue at a given time. High-throughput techniques like RNA sequencing (RNA-seq) are employed to capture these data, providing comprehensive snapshots of gene activity and regulation in various contexts (Wang *et al.,* 2009). To standardize the reporting of microarray-based gene expression data, the Minimum Information About a Microarray Experiment (MIAME) guidelines were established by then the Microarray Gene Expression Data (MGED) Society[61] (now known as Functional Genomics Data Society (FGED)) (Brazma *et al.,* 2001). The rise of NGS, however, has introduced new complexities that MIAME was not originally designed to address. As NGS technologies have different data types and requirements, development of complementary guidelines such as the Minimum Information about a High-throughput Nucleotide Sequencing Experiment (MINSEQE) (Brazma *et al.,* 2012) was highly necessary.

Proteome data encompasses the full complement of proteins expressed by a genome, cell, tissue, or organism at a particular time, reflecting the functional state of the biological system. Proteomics, the study of the proteome, uses techniques such as mass spectrometry (MS) and two-dimensional gel electrophoresis (2D-GE) to identify and quantify proteins, study their modifications, and understand protein-protein interactions (Liu *et al.,* 2019). The Minimum Information About a Proteomics Experiment (MIAPE), developed by the Human Proteome Organization (HUPO) [62], provides guidelines for documenting and sharing proteomic data, ensuring that the data are sufficiently detailed to allow for reproducibility and comparative analysis (Taylor *et al.,* 2007).

Metabolome data comprises the complete set of small-molecule metabolites present within a biological sample, offering a direct readout of the biochemical activity and metabolic state of cells. Metabolomics captures these data using techniques like nuclear magnetic resonance (NMR) spectroscopy and mass spectrometry. Metabolomic analyses provide insights into the metabolic pathways and physiological responses of an organism to genetic and environmental changes (Vinay *et al.,* 2021). The Minimum Information About a Metabolomics Experiment (MIAMET) guidelines have been proposed to ensure consistency and detail in documenting metabolic experiments (Sumner *et al.,* 2007). As with other PGR-associated datasets, handling and integrating these diverse molecular phenotype data involves several complexities. Sophisticated computational tools and algorithms are required to handle large, complex datasets and extract meaningful biological insights (Lim *et al.,* 2022; Deng *et al.,* 2023). Additionally, standardizing data collection and reporting is crucial for effective data sharing and reuse, but achieving this consistency across different laboratories and studies can be challenging.

Characteristics of Molecular Phenotypic Data

(1) Data size. The huge volume of molecular phenotype datasets presents a fundamental challenge due to the extensive data generated by high-throughput technologies. Transcriptomic data typically involve the sequencing of millions of reads per sample, which translates into files ranging from gigabytes to terabytes depending on the depth of sequencing and the number of samples (Lim *et al.,* 2022). Proteomic data can be similarly large, as MS-based proteomics identifies and quantifies thousands of proteins and their post-translational modifications (PTMs) in a single experiment. Each MS run can produce files that are gigabytes in size, necessitating substantial computational and storage resources (Sun *et al.,* 2009). Metabolomic datasets, although generally smaller than transcriptomic and proteomic datasets, still involve significant data volumes when comprehensive profiling is conducted across numerous metabolites and samples (Vinay et al., 2021). These large data sizes necessitate advanced storage solutions, high-performance computing, and efficient data processing pipelines.

(2) Data Heterogeneity. The intrinsic diversity of molecular phenotype data arises from the varied nature of the molecules involved. Transcriptomic data include various RNA species, each with distinct roles in gene expression regulation and cellular function (Lim *et al.,* 2022). This diversity requires different sequencing approaches and data processing pipelines. Proteomic data add another layer of heterogeneity, as proteins exhibit a wide range of structural complexities, functional roles, and interactions, along with PTMs that further diversify the proteome (Sun *et al.,* 2009; Liu *et al.,* 2019). Metabolomic data are characterized by a vast array of small molecules with different chemical properties, metabolic pathways, and biological functions, requiring specialized analytical techniques such as NMR spectroscopy and MS (Vinay *et al.,* 2021). The heterogeneity of these datasets presents significant challenges for integration and requires sophisticated bioinformatics tools tailored to each data type's specific characteristics.

[61] https://www.fged.org/
[62] https://www.hupo.org/

(3) **Data Availability and Accessibility**. The reach of molecular phenotype data has been greatly enhanced by the establishment of public repositories. For transcriptomic data, repositories like the Gene Expression Omnibus (GEO)[63] (Clough & Barrett, 2016) and ArrayExpress[64] (Parkinson *et al.,* 2007) provide extensive archives of high-throughput gene expression data, promoting data sharing and reuse. Proteomic data are stored in repositories such as the Proteomics Identifications Database (PRIDE)[65], archiving protein and peptide identifications along with quantitative and PTM data (Hermjakob & Apweiler, 2014). Metabolomic data are available through platforms like the Metabolomics Workbench[66] (Sud *et al.,* 2016) and MetaboLights[67] (Haug *et al.,* 2013), which offer comprehensive repositories for metabolomics experiments. Despite these resources, the availability of data can still be limited by issues such as proprietary datasets, delayed data publication, inconsistent data submission practices, varying levels of data curation, and incomplete metadata.

(4) **Data Complexity and Dimensionality.** The complexity and dimensionality of molecular phenotypic data stems from the intricate biological systems they represent. Transcriptomic data capture dynamic changes in gene expression in response to various stimuli, reflecting complex regulatory networks and interactions (Lim *et al.,* 2022). measure the expression levels of tens of thousands of genes simultaneously, resulting in high-dimensional data matrices that require dimensionality reduction techniques such as principal component analysis (PCA) or t-distributed stochastic neighbour embedding (t-SNE) for effective visualization and interpretation. Proteomic data are complex due to the diversity of thousands of proteins, their modifications, functional roles, PTMs, and interactions within cellular pathways (Liu *et al.,* 2019; Basu *et al.,* 2022). Metabolomic data, likewise, can include hundreds to thousands of metabolites and metabolic pathways, each subject to regulation and environmental influences (Vinay *et al.,* 2021; Manickam *et al.,* 2023). The experimental settings, provenance and data acquisition pipelines, which include details about sample collection, preparation, and the specific technologies used for data acquisition introduce additional layers of complexity. It is worth noting that the accuracy and reproducibility of the results heavily depend on the statistical models and parameters employed during data processing and analysis. Proper documentation of these protocols, including the software and algorithms used, is therefore essential to ensure the reproducibility and reliability of the findings. High complexity and dimensionality of these data types necessitate advanced bioinformatics, computational and statistical methods for data analysis, including machine learning algorithms and network analysis tools that can handle multidimensional data and uncover underlying biological patterns (Vinay *et al.,* 2021; Lim *et al.,* 2022). Furthermore, visualizing high-dimensional data in a comprehensible form often requires dimensionality reduction techniques. Capturing complex biological information in a machine-actionable and standardized way is also crucial for automated data integration, sharing, and analysis.

3.2.3 Image Data

Image data are integral to documenting and understanding PGR, serving as a critical component throughout the stages of collecting, characterization, and evaluation of accessions. The integration of high-quality image data allows for the detailed capture of morphological traits and environmental interactions, which are essential for the accurate identification, analysis, and utilization of PGR. As digital imaging technologies advance, the ability to collect, analyse, and share visual data becomes increasingly important, providing a rich source of information that complements traditional genetic and phenotypic data.

[63] https://www.ncbi.nlm.nih.gov/geo/
[64] https://www.ebi.ac.uk/biostudies/arrayexpress
[65] https://www.ebi.ac.uk/pride/
[66] https://www.metabolomicsworkbench.org/
[67] https://www.ebi.ac.uk/metabolights

During the germplasm collection phase, collectors often photograph plant specimens in their natural habitats, providing a visual record invaluable for subsequent identification and analysis (Araus & Cairns, 2014). These images document the morphological characteristics of the plant, its environment, and any associated biotic interactions. This step ensures that the visual context of the collected specimens is preserved, aiding in the distinction between phenotypically similar accessions and assessing potential environmental adaptations. Additionally, photographs serve as a complement to herbarium specimens or seed files, aiding in identification and verification processes. As PGR moves from collection to characterization and evaluation, the role of image data becomes increasingly pronounced. Characterization involves the detailed description of plant traits, both qualitative and quantitative. For a single accession, multiple images can be captured, documenting different parts of the plant such as leaves, stems, flowers, fruits, and seeds. These images are taken across various environments, including controlled settings like laboratories and greenhouses, as well as in field conditions. High-resolution images can capture subtle differences in traits such as leaf shape, flower colour, seed morphology, and growth habits. Advanced imaging techniques, such as multispectral and hyperspectral imaging, reveal traits not visible to the naked eye (Fiorani & Schurr, 2013). During evaluation, when the performance of accessions under various environmental conditions and management practices are being assessed, numerous images to document plant development, phenology, and response to treatments over time are generated. Multiple images per accession are critical, capturing different stages of growth and development, as well as responses to various experimental setups.

For *in situ*-maintained populations, image data provides invaluable insights and tools that significantly enhance the study and conservation of CWRs, WFPs and on-farm LR. With high-resolution satellite imagery, researchers can map the precise locations of these plants in their natural habitats, allowing for the detailed tracking of their distribution and density. This spatial information is crucial for identifying regions of high genetic diversity and areas at risk due to environmental changes, such as climate shifts, urban expansion, and agricultural development. Ground-based imaging further enriches the data pool with images of individual plants or small populations, which allows for the close monitoring of growth rates, phenological stages, and micro-ecological dynamics within plant communities. Moreover, integrating image data with geographic information systems (GIS) and other analytical tools facilitates the creation of detailed, interactive maps and models. These models combine image data with other relevant datasets, such as climate information, soil types, and land use patterns, to provide comprehensive insights into habitat suitability and species distribution. Such integrative approaches are vital for planning conservation strategies, including the establishment of protected areas and the development of ecological corridors that connect fragmented populations and maintain genetic flow.

Currently, no specific minimum information standards exist solely for image data in PGR. However, general standards for biological research image data can be applied. The Taxonomic Databases Working Group (TDWG)[68] has drafted the Minimum Information about a Digital Specimen (MIDS). The Distributed System of Scientific Collections (DiSSCo), a pan-European RI for natural science collections, utilize DwC and Dublin Core as metadata standards along with image annotation and provenance data models, W3C Web Annotation Data Model and PROV-DM. Despite these guidelines, challenges persist in integrating image data into PGR documentation. A major issue is the lack of standardized imaging protocols across different missions. Variations in camera settings, lighting, and image resolution can affect data consistency and quality. Standardizing imaging protocols can help mitigate these issues. Another challenge is managing and storing large volumes of image data. High-resolution and time-series images accumulate quickly, necessitating robust data management systems for effective storage, retrieval, and analysis. Cloud-based storage and advanced image analysis software, including machine learning algorithms, can help but require significant investment and technical expertise (Fiorani & Schurr, 2013). For more information on image data standards, refer to Deliverable 1.1[69].

[68]Currently known as Biodiversity Information Standards. https://www.tdwg.org/community/cd/mids/

[69]https://www.grace-ri.eu/pro-grace/outputs/deliverables/standards-for-collecting-and-displaying-phenotypic-data-and-images

Characteristics of Image Data

(1) Data size. PGR image data files are typically large, especially when high-resolution images are used to capture fine morphological details. The file size of a single high-resolution image can range from several megabytes (MB) to gigabytes (GB), depending on the resolution and format. When capturing time-series images or multispectral/hyperspectral data, the total data volume can quickly escalate. This large data size necessitates robust storage solutions and efficient data management practices to ensure that the images can be stored, retrieved, and processed effectively.

(2) Data heterogeneity. The significant heterogeneity in PGR image data arises from the diversity of imaging techniques (e.g., RGB imaging, multispectral imaging, hyperspectral imaging, thermal imaging), the variety of plant parts captured (e.g., leaves, stems, flowers, seeds), and the different environments in which images are taken (e.g., laboratory, greenhouse, field). Additionally, images can vary in terms of resolution, format, and the specific traits being documented. This heterogeneity requires the development of standardized protocols and metadata to ensure that the data are comparable and interpretable across different studies and applications.

(3) **Data availability and accessibility.** Several factors influence the accessibility and availability of PGR image data, including data management infrastructure, user permissions, and the interoperability of databases. Many databases are not equipped to handle large volumes of image data, which can limit accessibility. Moreover, image data are often stored in proprietary formats or dispersed across different platforms, further complicating access. Ensuring the availability of comprehensive and high-quality image data requires coordinated efforts in data collection, adherence to standards and proper documentation of metadata. Initiatives to digitize existing herbarium specimens and field collections can significantly enhance the availability of historical image data. Improving accessibility involves developing integrated databases that support various image formats, implementing user-friendly interfaces, and establishing data-sharing protocols that facilitate easy access for researchers and breeders

(4) **Data complexity.** The complexity of PGR image data is attributed to the multifaceted nature of the images and the biological traits they capture. Images can contain information on various morphological and physiological traits that may require advanced image processing and analysis techniques to extract meaningful data. For example, analysing multispectral or hyperspectral images involves dealing with multiple layers of data corresponding to different wavelengths, which can be computationally intensive. Furthermore, integrating image data with other types of data, such as genomic or environmental data, adds another layer of complexity. Additionally, variability in imaging conditions, such as lighting and angle, requires standardization to ensure consistency across different datasets. Advanced analytical techniques, including machine learning and artificial intelligence, are often needed to handle the complexity and extract valuable insights from the data.

(5) **Data Dimensionality.** Dimensionality in PGR image data refers to the spatial, spectral, and temporal dimensions captured in the images. Spatial dimensionality involves the resolution and scale of the images, which determine the level of detail visible. Spectral dimensionality includes the range and number of wavelengths captured, as seen in multispectral and hyperspectral imaging. Temporal dimensionality involves capturing images over time to monitor changes and development in plant traits. High-dimensional image data provide a wealth of information but require sophisticated analytical tools and techniques to process and interpret.

3.2.2 Genomic Data

Genomic data involves the comprehensive sequencing of an organism's DNA, providing a complete representation of its genetic makeup. This data type includes various forms of genetic variation such as single nucleotide polymorphisms (SNPs), insertions and deletions (indels), structural variants (SVs), and copy number variations (CNVs). Epigenomic data, on the other hand, examines heritable changes in gene expression that do not involve alterations in the DNA sequence itself. This includes DNA methylation patterns, histone modifications, and chromatin accessibility.

High-throughput sequencing technologies such as Illumina, PacBio, and Oxford Nanopore have revolutionized the acquisition of genomic and epigenomic data, each bringing unique advantages and challenges to the field (Crossley *et al.,* 2020; *Hu et al.,* 2021). The vast array of data generated by these sequencing technologies necessitates standardized data formats and robust information standards to ensure data reproducibility and interoperability. For genomic data, common formats include FASTQ for raw sequence reads, containing nucleotide sequences and quality scores; SAM/BAM for aligned sequence data; and VCF for SNPs and other genomic variants (Deng *et al.,* 2023). Gene annotation data is frequently stored in formats like General Feature Format (GFF) or Gene Transfer Format (GTF) (Yandell & Ence, 2012). Meanwhile, epigenomic data is stored in formats tailored to specific types of information. For example, BED format represents regions of interest in the genome, such as peaks identified in ChIP-seq or ATAC-seq experiments (Quinlan & Hall, 2010). BigWig format is used for storing continuous data tracks, such as coverage or signal intensity, allowing efficient visualization and analysis (Pohl & Beato, 2014).

Adherence to established data standards like the Minimum Information about a Genome Sequence (MIGS) is crucial for ensuring comprehensive metadata capture (Field *et al.,* 2008). The Genomic Standards Consortium (GSC)[68] has developed the Minimum Information about any (x) Sequence (MIxS) standards, extending MIGS to various types of sequence data, including metagenomic sequences (MIMS), marker genes (MIMARKS), and environmental data (MIENS) (Field *et al.,* 2008). These standards provide structured formats for describing essential contextual information, ensuring that critical metadata is captured (Brazma *et al.,* 2001). In addition to MIGS and MIxS, MINSEQE offers standards for reporting high-throughput sequencing data, including details on experimental design, sample processing, and data analysis. **A more detailed discussion on the standards for collecting and displaying genetic data can be found in Deliverable 1.2**[69].

Despite these advancements, significant challenges and gaps persist in the field of PGR-associated genetic data standards. One major issue is the proliferation of numerous existing standards, which can result in confusion and inconsistency within the research community. Researchers often face difficulties in selecting the appropriate standard for their specific data type and study design, leading to variability in data reporting and metadata completeness (Field *et al.,* 2008; Sansone *et al.,* 2012). This inconsistency can undermine the reproducibility and comparability of research findings. Moreover, many existing standards are tailored for specific data types or experimental techniques, limiting their applicability across different genomic and epigenomic studies. This specialization can create fragmentation, making it challenging to integrate diverse datasets and complicating comparative analyses and meta-analyses (Brazma *et al.,* 2001; Wilkinson *et al.,* 2016).

Another critical issue is the inconsistent adoption and compliance with these standards across the research community. While some databases and journals mandate strict adherence to specific standards, others do not enforce such requirements, leading to significant variability in data quality and metadata availability (Field *et al.,* 2008; Sansone *et al.,* 2012; Wilkinson *et al.,* 2016).

[68]https://www.gensc.org//

[69]https://www.grace-ri.eu/pro-grace/outputs/deliverables/standards-for-collecting-and-displaying-genetic-data

This inconsistency undermines efforts to achieve interoperability and hampers the ability to effectively share and reuse data, thus diminishing the overall impact and utility of the collected information (Wilkinson *et al.,* 2016).

Moreover, the rapid evolution of sequencing technologies and analytical methods presents ongoing challenges for maintaining and updating these standards. As new methodologies emerge, they often fall outside the scope of existing standards, necessitating continuous efforts to revise and expand guidelines to keep pace with technological advancements. This dynamic landscape requires flexible and adaptive standards that can accommodate innovative approaches while ensuring consistency and reliability in data reporting.

Data quality and completeness remain paramount concerns. Sequencing errors, incomplete genome assemblies, and inaccuracies in gene annotation can significantly impede downstream analyses and applications (Yandell & Ence, 2012). High-quality reference genomes are currently available for only a limited number of plant species, predominantly those with high economic value (Goodwin *et al.,* 2016; Lewin *et al.,* 2018). There is an urgent need to sequence and assemble reference genomes for a broader range of plant species, including wild relatives and underutilized crops, to enhance the utility of genetic data in PGR. Addressing these gaps will improve the accuracy and comprehensiveness of genomic databases, facilitating more robust and reliable research outcomes.

Characteristics of Genetic Data

(1) **Data size**. High-throughput sequencing technologies can generate vast amounts of data; a single sequencing run can produce terabytes of information (Goodwin *et al.,* 2016). This immense size is due to the high resolution of the data, which includes detailed information about SNPs, indels, SVs, and epigenetic modifications. The substantial volume of data necessitates significant computational resources for storage, processing, and analysis, posing a challenge for many research institutions.

(2) **Data heterogeneity**. The heterogeneity of genomic and epigenomic data arises from the various types of sequencing technologies used, and the diverse experimental conditions applied. This heterogeneity results in datasets that vary significantly in structure and content (Yandell & Ence, 2012). For example, genomic data can include sequences from whole-genome sequencing, exome sequencing, or targeted sequencing, each providing different types of information. Similarly, epigenomic data may encompass DNA methylation patterns, histone modifications, and chromatin accessibility profiles, each captured using different assays such as bisulfite sequencing, ChIP-seq, or ATAC-seq (Laird, 2010). This diversity in data types adds layers of complexity to data integration and analysis.

(3) **Data availability and accessibility**. Genomic and epigenomic data are made available through several major repositories, ensuring that researchers worldwide can access this valuable information. The three primary nucleotide sequence databases are the DNA Data Bank of Japan (DDBJ)[51], the European Nucleotide Archive (ENA)[52], and the National Center for Biotechnology Information (NCBI)[53] GenBank in the United States. Together, these databases form INSDC[54], which facilitates the global exchange and accessibility of nucleotide sequence data (Kodama *et al.,* 2012). Researchers are generally required to submit their sequencing data to these repositories as part of the publication process, ensuring that underlying data are accessible for validation and further study (Cochrane *et al.,* 2011). These repositories provide critical infrastructure for data storage and retrieval, enhance transparency, reproducibility, and foster collaboration (Wilkinson *et al.,* 2016).

[51] https://www.ddbj.nig.ac.jp/index-e.html
[52] https://www.ebi.ac.uk/ena/browser/home
[53] https://www.ncbi.nlm.nih.gov/
[54] https://www.insdc.org/

This open-access approach supports scientific discovery by enabling the reuse of existing datasets, thereby accelerating research and innovation. The INSDC ensures long-term data preservation and provides tools for efficient data retrieval and analysis, making these datasets invaluable resources for the global scientific community (Cochrane *et al.,* 2011).

(4) **Data complexity**. The complexity of genomic and epigenomic data arises from the need to integrate multiple layers of diverse and dynamic biological information. Advanced computational methods are essential for managing and analysing the complex, multi-dimensional datasets generated from plant genomic and epigenomic studies. These methods rely heavily on the availability of high-quality, well-annotated data that conforms to established standards. Ensuring that data adhere to these standards facilitates the development and application of sophisticated bioinformatics tools and algorithms, enabling researchers to uncover meaningful patterns and relationships within the data (Cochrane *et al.,* 2011; Kodama *et al.*, 2012).

(5) Data Dimensionality. These data types are inherently multi-dimensional, involving numerous variables that need concurrent analysis. Each genome can be examined on multiple levels, such as sequence variation, gene expression, and epigenetic modifications, with each level representing a distinct dimension of the data (Kodama *et al.,* 2012).

3.3 PGR Information Management Challenges

3.3.1 Data Fragmentation and Inconsistencies in Data Collection Protocols and Formats

Data fragmentation is an unintended consequence of the decentralized approach to managing and characterizing PGR. It stems from the varying methodologies, standards, formats and tools utilized by GRC, research institutions, projects and breeding programs (Volk *et al.,* 2021; Deng *et al*., 2023). Inevitably, data fragmentation results in difficulties in coordinating information across multiple repositories, platforms, institutions and international borders (Halewood *et al.,* 2018a; Engels & Thormann, 2020). As a result, PGR-associated datasets become isolated and difficult to integrate, and subsequently create significant challenges for complementary *ex situ* and *in situ* conservation planning and implementation, comprehensive resource analysis and utilization (Halewood *et al.,* 2018b; Volk *et al.,* 2021).

Historically, PGR institutional databases were developed in silos, resulting in myriad information systems that differ widely in terms of data languages, formats, structures, and protocols. Such is still the case for the current development of *in situ* conservation data recording, analysis and storage (Maxted *et al.*, 2020). These differences arise from factors (i.e. organizational goals, stakeholder agendas, technological advancements, expertise levels, funding sources, and data management practices) that are inherent to these particular institutions. Consequently, researchers face significant challenges due to the dispersion of critical genetic resource data (Halewood *et al.,* 2018b). Essential information for crop improvement, such as traits for disease resistance or climate resilience, is often confined to multiple, disparate databases, making it difficult to access for comprehensive analyses (Volk *et al.,* 2021). For instance, phenotypic, genetic, and passport data for a single crop species might be distributed across various project-based, national, and international databases, each using unique identifiers, metadata standards, and access protocols. This lack of interoperability complicates the process of compiling and analysing datasets, thereby hindering research and breeding efforts and increasing the risk of redundant or contradictory information.

Inconsistent data collection, documentation and annotation protocols likewise exacerbate data fragmentation issues (Ćwiek-Kupczyńska et al., 2016; Andres-Hernandez et al., 2021; Lücking et al., 2022). For example, phenotypic data on a given set of accessions can vary significantly due to differences in measurement techniques, environmental conditions, and trait definitions (Pieruschka & Schurr, 2019). Similarly, genetic data might be sequenced using different technologies or analysed with disparate bioinformatics pipelines, resulting in non-comparable datasets (Deng et *al.,* 2023). These

inconsistencies emanate from differences in resource availability, expertise, and institutional priorities, which are vital factors to reckon with in PGR data management and characterization.

Geographical and institutional barriers further compound data fragmentation. Within individual countries, a national genebank, universities, research institutions and non-governmental organizations (NGOs) may hold extensive germplasm collections and associated datasets that have hitherto remained unshared due to administrative obstacles and insufficient coordination (Hammer *et al.,* 2003; Ramanatha Rao & Hodgkin, 2002). These barriers simultaneously create significant gaps and overlaps in PGR information and ultimately undermine efforts to establish comprehensive databases that support effective conservation and research initiatives.

These fragmentation and interoperability barriers exist even within the PGR community itself, specifically between those working on *ex situ* and *in situ* conservation applications, which even today are largely planned and managed in isolation (Maxted *et al.*, 2016). Obviously, we should be working in a more integrated manner to more effective conserve maximum diversity and make that diversity available to full range of end users. It would be timely to resolve the *in situ* / *ex situ* fragmentation and interoperability barriers now because having established the theoretical framework it is only recent that *in situ* conservation activities have begun to be implemented in practice and the novel setup could avoid unnecessary barriers. Now is the ideal time to commence the necessary *in situ* / *ex situ* informatics dialogue.

Data fragmentation, thus, negatively impacts PGR conservation and management in more ways than one. First, it hinders interoperability and complicates the process of compiling and analysing datasets, thereby hampering research and breeding efforts (Halewood *et al.,* 2018b). Second, it renders conservation efforts redundant and inefficient, leading to wastage of resources and the neglect of critical genetic resources that should have been given priority in the first place (Volk *et al.,* 2021). For reasons of their own, some genebanks often duplicate conservation of the same genetic material, leading to redundancy at the expense of other critical accessions. Through data paucity and mismanagement, multiple genebanks may thus unknowingly conserve numerous duplicates of widely grown crop varieties, while unique local varieties that are essential for biodiversity and adaptation, are insufficiently represented (Engels & Thormann, 2020; Ford-Lloyd *et al.,* 2011). This underrepresentation of critical genetic resources, while seemingly not an immediate cause for concern, will have dire consequences in the long term. It should also be remembered that unconserved or poorly conserved accessions / populations are more likely to suffer genetic erosion or go extinct and they by definition will remain unavailable for utilization (Maxted *et al.*, 2016).

3.3.2 Compromised and Incomplete Datasets

Ensuring data quality in PGR management is crucial for advancing conservation, research, and breeding programs. Among the determinants of data quality are accuracy, consistency, and reliability - all of which are influenced by human error, technological limitations, comprehensive metadata documentation, data integration practices, validation and verification processes, traceability, and data stewardship. Occasionally, however, several of these factors may work singly or jointly to compromise the quality of PGR datasets.

Human errors arising from manual data entry, subjective assessments, and inconsistent recording practices can introduce errors and inconsistencies. For example, different observers might record phenotypic traits differently, leading to variability in the data. While standardized training and detailed protocols can help mitigate these discrepancies, human error remains an inherent risk in data collection processes. Additionally, outdated or poorly calibrated equipment can yield poor readings and measurements, and consequently confound analysis. By adopting advanced technologies, such as

high-throughput sequencing, automated phenotyping platforms, and field guides for taxon identification, improve data accuracy and the potential for human error is significantly reduced. The hefty investments and maintenance expenses associated with these technologies, however, present problems for many institutions (Rife and Poland, 2014; Das *et al.,* 2022).

Comprehensive metadata documentation is critical but often inadequately addressed. Metadata provides the context needed to understand and correctly interpret data, including details about the conditions and methods used during data collection. Without comprehensive metadata, assessing the reliability and relevance of data becomes problematic. For instance, prevailing environmental conditions during data collection have to be taken into account since these may potentially skew the interpretation of phenotypic data (Wieczorek *et al.,* 2012). However, while comprehensive metadata improves data quality and enhances transparency and reproducibility (Robertson *et al.,* 2014), collecting detailed metadata is resource-intensive and time-consuming, requiring meticulous attention to detail and commitment to best practices (Wieczorek *et al.,* 2012).

The lack of built-in validation and verification mechanisms also predisposes a data management system to errors. Ensuring that data is accurate and complete at the point of entry is a significant hurdle. Verification, which involves cross-checking data against other reliable sources, is often neglected due to the additional effort and resources required. When, for instance, phenotypic data is not routinely compared with historical records or external datasets, errors persist undetected (Postman *et al.,* 2010). Like advanced technologies and comprehensive metadata, however, automated tools for data validation and verification require substantial resource expenditure (Joly *et al.,* 2014). Traceability issues also pose a challenge in maintaining data integrity and reliability. It involves keeping detailed records of data sources, including data acquisition methods, and any modifications. Initial plotting of distributional data may also help identify erroneous reported passport data, such a collection site in water bodies or in the wrong country. This practice is essential for tracking the history of data changes and identifying the origin of errors or inconsistencies. To date, however, many data management systems lack robust mechanisms to keep track of these details, making it difficult to ensure complete and reliable data. Implementing traceability requires comprehensive documentation practices and a commitment to maintaining detailed records throughout the data lifecycle (Wilkinson *et al.,* 2016).

Furthermore, early conservation efforts were often carried out without the systematic protocols and rigorous documentation standards that are now deemed crucial. As a result, many accessions collected during these formative years were recorded with scant information, or in some cases, no metadata at all. This dearth of detailed contextual information, such as precise collection locations, biological profiles, ecological and environmental context, severely limits the utility of historical datasets. Compounding this issue is the fact that many of these early accessions cannot be re-collected or validated in their natural settings today. In many instances, these genetic resources may no longer be extant in their original habitats due to several environmental and anthropogenic pressures. This loss underscores the critical need to develop and implement strategies for digitizing, standardizing and integrating legacy data (including field notes and collection logs) into current information systems to address documentation gaps and maximize the value of these surviving records. Each piece of information from these early collections represents a unique and potentially irreplaceable genetic resource.

Finally, effective data stewardship ensures that data remains accurate, consistent, and complete over time. It involves the careful and responsible management of data throughout its lifecycle- from collection and storage to sharing and archiving. It encompasses policies, procedures, and practices designed to ensure that data is handled ethically, legally, and efficiently.

3.3.3 Volume and Complexity

Today, data generation has become less of a bottleneck than managing the sheer volume and complexity of data produced by modern research techniques across various scientific fields (Fahlgren *et al.,* 2015; Goodwin *et al.,* 2016; Kersey *et al.,* 2018). This deluge of data, despite its potential to advance our understanding in biological sciences, creates substantial difficulties for current data management strategies.

As discussed in 3.2.2, automated phenotyping platforms can record hundreds of thousands of images and data points daily, translating to petabytes of data annually (Fahlgren *et al.,* 2015). The Phenovator system, for instance, can collect up to 100,000 images per day, resulting in over 36 million images per year that need to be processed and analysed (Flood *et al.,* 2016). Similarly, the European Plant Phenotyping Network (EPPN) generates massive datasets from multiple sites and crops, producing extensive phenotypic data that include millions of images and terabytes of sensor data annually, significantly contributing to our understanding of plant performance under various conditions (Tardieu *et al.,* 2017).

High-throughput sequencing technologies (detailed in 3.2.3) have dramatically increased the speed and scale at which genetic data can be collected, producing up to 1.8 terabases per run (Goodwin *et al.,* 2016). Sequencing a single plant genome can generate hundreds of gigabytes to several terabytes of raw data, necessitating substantial storage and processing capabilities. The 10,000 Wheat Genomes Project exemplifies this massive data generation effort by aiming to sequence and analyse the genomes of 10,000 wheat accessions from global genebanks. Each genome sequencing run within this project generated hundreds of gigabytes to terabytes of raw data, underscoring the vast scale and complexity involved in managing and analysing such extensive datasets (Appels *et al.,* 2018).

Furthermore, molecular phenotyping adds another dimension of complexity to PGR data. RNA-seq, for example, can produce gigabytes of data per sample, detailing the transcriptome with millions of reads that must be aligned and quantified, resulting in terabytes of data when scaled to hundreds or thousands of samples (Wang *et al.,* 2009). Likewise, mass spectrometry-based proteomics can generate extensive datasets, often comprising thousands of identified proteins, each with its own abundance and modification status, leading to complex data matrices that require advanced computational tools to analyse and interpret (Aebersold & Mann, 2016). Metabolomics studies also produce large datasets encompassing thousands of metabolites, each with unique properties (Benton *et al.,* 2015).

Each data type inevitably involves different scales and formats, from nucleotide sequences to protein interaction networks and metabolite profiles. A single multi-omics study can produce data in the range of terabytes, including tens of thousands of gene expression profiles, protein quantifications, and metabolite measurements (Chen *et al.,* 2012). Integrating these datasets requires advanced bioinformatics pipelines and computational infrastructure capable of handling high-dimensional and heterogeneous data (Kersey *et al.,* 2018).

3.3.4 Access and Availability

Tapping into the wealth of PGR data and metadata, much like finding the right pieces of a vast jigsaw puzzle, hinges not merely on data collection but on sophisticated frameworks that ensure its accessibility, availability, and integration. Addressing these facets involves tackling several interconnected challenges, spanning infrastructure, standards, policy frameworks and collaborative practices.

Researchers often face the daunting task of locating and retrieving PGR information dispersed across multiple repositories, including various online databases, remote datasets, and even files kept on hard disks or in lab notebooks. Recently, the adoption of FAIR data principles has gained traction. These principles aim to enhance the accessibility and utility of data by ensuring it is discoverable and usable by both humans and machines. Implementing FAIR principles involves establishing comprehensive metadata standards that ensure consistent data descriptions across repositories, facilitating easier searching and retrieval (Wilkinson *et al.,* 2016; Gullotta *et al.,* 2023). Persistent identifiers, such as Digital Object Identifiers (DOIs), have been increasingly adopted to improve data traceability and citation. These identifiers provide stable references to datasets and help maintain data integrity over time. The adoption of DOIs and similar persistent identifiers is part of a broader effort to standardize data management practices, making it easier for researchers to access and use genetic resources consistently (Gullotta *et al.,* 2023).

Significant advancements in interoperability protocols have also been made, enhancing the ability of different data systems to communicate effectively and ensuring seamless data exchange and integration across various platforms. Approaches like BrAPI (Breeding API), which provides a standardized way to access and share plant breeding data across multiple databases and software tools, are essential for linking disparate data sources and ensuring that researchers can leverage comprehensive datasets for their studies (Selby *et al.,* 2019).

Nevertheless, substantial challenges remain. Substantial investments in data storage solutions and infrastructure capable of handling large volumes and high velocities of data are still needed. Many institutions struggle to upgrade their systems to accommodate the increasing amounts of multi-omic data being generated, especially in developing countries where funding for advanced data management systems and training is often lacking. Some institutions continue to rely on legacy systems developed using now-obsolete technologies such as early versions of database management software. These systems were initially designed to handle the data volumes and types available at their creation and have preserved vast amounts of genetic information over the years. Migrating data from legacy systems to modern platforms, while ensuring data integrity and compatibility during the migration, is a complex process. This process is particularly challenging when data is stored in proprietary formats that need to be converted to modern, open standards. This migration is not only technically demanding but also time-consuming and costly, leading to uneven progress in data accessibility and utilization. It is also vital that the staff using the system are adequately trained in its usage, so they are able sustain use post-project or at least seek support if circumstances change. Moreover, despite the push towards standardized metadata, ensuring uniform data standards and practices across diverse platforms and institutions remains a significant struggle. The variation in resources, technical capabilities, and adherence to standards across different regions and organizations makes it challenging to implement a unified approach (Wilkinson *et al.,* 2016).

Political and institutional barriers further complicate the full accessibility and availability of PGR data. The Convention on Biological Diversity (CBD), its Nagoya Protocol, and the ITPGRFA link benefit-sharing obligations to the access and use of physical genetic material. However, they do not adequately address digital data, creating regulatory uncertainties that hinder data sharing and collaboration. This gap can slow down research progress by restricting access to crucial genetic information stored in digital formats (Volk *et al.,* 2021; Gullotta *et al.,* 2023) .

4. Assigning Persistent Unique Identifiers for PGR

To guarantee the accessibility and interoperability of PGR data, the use of Persistent Unique Identifiers (PUIDs) is identified as a critical component within the minimum information standards. PUIDs serve as a foundational element in ensuring that each genetic resource can be distinctly recognized and

accessed over time, thereby supporting the effective exchange and use of these resources across various platforms and databases (Manzella *et. al.*, 2022). The INSDC repositories have developed a comprehensive set of unique identifiers coherent across their archives. For instance, genomes are annotated with a consistent format, such as "GCA_XXXX," where "GCA" represents GenBank Complete Assembly. This standardized approach facilitates straightforward reference and comparison across different studies and databases. Genetic variants are identified using specific submitted SNP ID numbers ("ss#"), which provide detailed information about the taxon, reference genome, study, position, and reference/alternate alleles. The sequenced DNA samples are provided with a unique identifier, as outlined by Courtot *et al.* (2019), that follows a systematic format beginning with "SAM," followed by a letter (E, N, or D) indicating the original submission location (EMBL-EBI, NCBI, or DDBJ, respectively), and a subsequent code (A or G) denoting whether the sample is an assay sample or a group of samples, followed by a numeric component. This structured approach ensures consistent and clear identification of samples across different repositories.

Meanwhile, The ITPGRFA under the auspices of the Food and Agriculture Organization (FAO), explicitly promotes the use of Digital Object Identifiers (DOIs) as a specific form of PUID, to facilitate the unambiguous and permanent identification of PGR accessions (Alercia *et al.*, 2018; Weise *et al.*, 2020). DOIs are unique alphanumeric strings designed to persistently identify physical, digital, or abstract objects, providing a stable and reliable mechanism for data management and retrieval. DOIs provide a permanent link to a resource's digital location, ensuring that the associated data can be reliably accessed and cited. The DOI system's strength lies in its stability; even if the resource's URL or location changes, the DOI remains unchanged, with the resolution mechanism redirecting to the new location, ensuring continuous access to the resource (Paskin, 2006).

This persistence is crucial for accurately tracking and referencing data across various platforms and databases, fundamental to research, conservation, and breeding efforts. DOIs enable seamless integration and efficient information retrieval, facilitating exchange, collaboration, and cohesiveness among stakeholders and the broader PGR community. By linking genetic resources with comprehensive metadata, research findings, and ancillary data, DOIs ensure that detailed and relevant information is accessible for each accession. Interconnectedness allows stakeholders to leverage comprehensive data for informed decision-making, driving innovation and breakthroughs in various scientific domains. The robustness of DOIs ensures that data associated with PGR accessions remains accessible and useful over time. As digital resources evolve and locations change, the persistent nature of DOIs guarantees that researchers can continually locate and use the information they need, vital for longitudinal studies and maintaining the integrity of historical data. The adoption of DOIs in PGR accession enhances data discoverability and interoperability, integral to maintaining consistent and accurate records of PGR accessions, *in situ* / on-farm populations, and supporting the sustainable management and utilization of genetic resources globally. The use of DOIs fosters greater collaboration and cohesion within the PGR community, enabling researchers to make informed decisions and drive innovation across various scientific domains.

Case in Point 3. The Challenges of Making PUID, e.g. DOI, Use Mandatory

Despite the availability of infrastructure for PUID, specifically DOI, assignment through various initiatives, the use of DOIs remains strongly recommended rather than mandatory across the global PGR community. This raises critical questions about the pace of adoption and the future of standardized data management.

Although systems for DOI minting exist (e.g., use of the infrastructure of the ITPGRFA), their use is still limited in many institutions and genetic resource centers due to a lack of technical resources and expertise. Key challenges include:

- Many GRC use legacy data management systems that are not designed to handle modern identifier systems like DOIs. Integrating these systems can require significant technical upgrades and data migration efforts, necessitating meticulous planning to maintain data integrity, accuracy, and continuity.
- Ensuring that data across different accessions / populations is standardized and formatted correctly for DOI assignment can be a complex and resource-intensive process.
- GRC often operate with limited funding and resources. Implementing a new system for DOI assignment and management can be costly. Securing funds for such projects can be difficult when there are competing priorities.
- The process requires skilled personnel to manage the transition, maintain the system, and ensure data accuracy. It requires a certain level of technical expertise in information technology and digital resource management, which may not be readily available in all genebanks.
- Not all GRC may be fully aware of the benefits and processes involved in DOI implementation. The perceived complexity of implementing and managing a DOI system can also deter adoption.

The inconsistent adoption of DOIs across various GRC results in heterogeneous data management practices. This inconsistency generates data duplication and inefficiency, thereby undermining the potential advantages of a standardized DOI system. This issue raises a critical question: **how sustainable is the optional use of DOIs, and what are the potential consequences for global PGR data integration and interoperability if a mandatory DOI system is not implemented?**

Furthermore, assigning DOIs to *in situ* populations presents a complex set of challenges. *In situ* populations are subject to dynamic environmental conditions, natural evolutionary processes, and human activities. Consequently, these populations are often not static and can vary significantly over time and space, introducing unique difficulties in tracking, documenting, and maintaining accurate records. Comprehensive metadata describing the location, ecological context, and temporal aspects of the population should accompany each DOI.

Additionally, establishing accurate data linkages between *in situ* and *ex situ* populations is crucial. Implementing DOIs also requires clear agreements on data ownership, access rights, and benefit-sharing, for *in situ* data possibly involving indigenous or local communities. Given these complexities, the application of the DOI system to effectively manage *in situ* populations may need to be adapted, and novel protocols to maintain accurate and up-to-date records developed.

5. Utilizing Controlled Vocabularies and Ontologies for PGR Documentation

Ontologies, defined as systematic representation of a domain of knowledge where key concepts, entities, and the logical relationships among them are clearly established, are fundamental in enhancing the interoperability, precision, and reusability of scientific data (Smith *et al.*, 2007; Arnaud *et al.*, 2020). By providing a comprehensive, yet structured vocabulary for describing traits, phenotypes, environments, and genetic relationships, ontologies facilitate the consistent annotation of PGR data across various sources, data types, and platforms. Ontologies, owing to their capacity to facilitate the description of genetic resources in a standardized, machine-readable format, and promote understanding across different scientific domains, are therefore strongly recommended for use in PGR documentation.

The importance of ontology in the documentation and management of PGR data stems from the following reasons: Firstly, it promotes standardized terminologies that are essential for the efficient exchange of information. A common language allows researchers and practitioners from different fields to understand and utilize data without ambiguity (Arnaud *et al.*, 2020). Secondly, ontologies make possible comparative data analytics by providing a reference framework that aligns with recognized standards. This facilitates the identification of similarities and differences across datasets. Lastly, the integration of data from disparate sources becomes significantly more feasible with ontologies. By mapping diverse data elements into a unified vocabulary, ontologies support the aggregation, comparison, and synthesis of information across multiple studies and repositories (Deng *et al.,* 2023).

While this deliverable recommends relevant ontologies (viz. Crop Research Ontology, Crop Ontology (CO), Plant Trait Ontology (PTO), Plant Experimental Conditions Ontology (PECO), Environment Ontology (ENVO), among others) for the documentation of PGR, one must appreciate the nuances/intricacies associated with ontological use. A practitioner is thus expected to exercise due diligence to determine specific ontologies that are appropriate for particular datasets and situations. Stakeholders are encouraged to consult current literature, ontology repositories, and community guidelines to identify the most appropriate ontologies for their specific needs. Additionally, the use of Ontology Look Up Services (OLS) is recommended as a valuable tool for finding and selecting the most relevant and up-to-date ontologies to support the documentation and research of PGR (See Deliverable 4.1 for more information).

6. The Concept of Minimum Information Standards in Data-Driven Science

The rapid growth of data within scientific research has underscored the necessity of robust frameworks to guarantee data quality, reproducibility, and usability (Borgman, 2015). As a result, **Minimum Information Standards (MIS)** have been developed as guidelines for specifying critical details necessary for datasets, including metadata and methodological information, to be comprehensively utilized across various studies and applications (Taylor *et al.,* 2008; Sansone *et al.,* 2012). The term **"minimum"** in MIS **pertains to core information in data documentation**, which encompasses all requisite details that render a dataset self-sufficient for its primary objectives (Brazma *et al.,* 2001). For example, in genomic studies, core information includes data about sequencing methods, sample preparation, and data processing steps (Field *et al.,* 2009). These details ensure the replicability of the study's findings. It gives emphasis on meticulous documentation and comprehensive methodological descriptions, experimental conditions, and data processing workflows, thereby ensuring that datasets enable accurate replication of studies while also facilitating further analyses, validation, and building upon existing knowledge (Brazma *et al.,* 2001; Taylor *et al.,* 2008).

Ensuring a balanced approach to data documentation through MIS is essential in mitigating the risks associated with both excessive and insufficient documentation (Taylor *et al.,* 2008; Sansone *et al.,* 2012; Kuhn *et al.,* 2008; Brazma *et al.,* 2001). Excessive requirements can impede data collection, sharing, and adherence to standards, while inadequate documentation may compromise the utility of datasets. This strategic approach ensures that researchers can effectively manage data while meeting compliance standards and facilitating dissemination to pertinent stakeholders (Taylor *et al.,* 2008; Sansone *et al.,* 2012; Kuhn *et al.,* 2008; Brazma *et al.,* 2001). By providing checklists and standardized documentation practices, MIS help ensure that datasets are understandable and usable by others. This approach facilitates the immediate reuse of data and its longevity and relevance over time (Sansone *et al.,* 2012; Taylor *et al.,* 2008). Methodological transparency involves detailed descriptions of data collection and processing methods. This includes specifying tools, instruments, protocols, and software used, allowing other researchers to replicate the study or understand its limitations. Transparent methodologies also help identify potential sources of bias or error, enhancing the overall reliability of research findings (Wilkinson *et al.,* 2016; Taylor *et al.,* 2007).

The concept of MIS has evolved in response to the growing complexity of scientific research and data management needs. Early initiatives in high-throughput technologies laid the foundation for broader adoption across various domains. One of the first significant MIS initiatives was MIAME, developed by the microarray research community in 2001. MIAME provided guidelines for documenting microarray experiments and addressing the reproducibility crisis by specifying essential metadata and methodological details (Brazma *et al.,* 2001). The adoption of MIAME marked a turning point in how scientific data was documented and shared, setting a precedent for future standards.

Table 1. Examples of Minimum Information Standards

STANDARD	DOMAIN	KEY ELEMENTS	REFERENCE
MIAME	Microarray Experiments	Experimental design, sample information, data processing methods	Brazma *et al.,* 2001
MIAPE	Proteomics Experiments	Sample preparation, data acquisition, analysis procedures	Taylor *et al.,* 2007
MIxS	Genomic Sequences	Sequencing technology, environmental context, data processing	Yilmaz *et al.,* 2011
MIBBI	Biological/Biomedical	Integrates various MIS, comprehensive documentation framework	Taylor *et al.,* 2008
MINSEQE	High-Throughput Sequencing	Sequencing platforms, library construction, data processing	Yilmaz *et al.,* 2011
MIFLOWCYT	Flow Cytometry	Instrumentation, sample preparation, data analysis methods	Lee *et al.,* 2008
MIAPPE	Plant Phenotyping	Experimental design, environment details, data acquisition	Papoutsoglou *et al.,* 2017

Following MIAME, the proteomics community developed the Minimum Information About a Proteomics Experiment (MIAPE). MIAPE ensured that proteomics data was sufficiently detailed for reproducibility and reuse, covering sample preparation, data acquisition, and analysis methods (Taylor *et al.,* 2007). The detailed protocols required by MIAPE helped standardize proteomics research, making comparing and contrasting results from different studies easier. This standardization was

particularly important for large-scale proteomics projects, often involving collaborations across multiple laboratories. The success of MIAME and MIAPE led to the development of MIS for other fields, such as the MIxS for genomic data (Yilmaz *et al.,* 2011). These standards tailored the concept of minimum information to different scientific domains' specific needs and challenges. Each new MIS built upon the lessons learned from previous efforts, refining the balance between comprehensive documentation and practical implementation.

7. Rationale for Developing an Integrative Framework: Harmonizing Minimum Information Checklists through MI-PGR

The current state of scientific data management is characterized by a multitude of documents specifying the minimum information required when reporting various types of experimental data. These documents, developed independently, differ significantly in data formats, terminologies, levels of detail, scope and even the underlying conceptual frameworks. This heterogeneity creates considerable barriers to the integration of datasets adhering to disparate standards, thereby impeding comprehensive data analysis and utilization.

Independent standards, while ensuring thorough and context-specific data reporting, inadvertently create data silos that impede the seamless integration of information across different domains. This fragmentation is particularly problematic given the increasing emphasis on the secondary use of data, where existing datasets are reanalysed to generate new insights. The growing trend towards data-driven research—relying heavily on the ability to synthesize large, diverse datasets to uncover novel patterns and inform future investigations—further underscores the need for well-integrated data. Hence, the development of a comprehensive and integrative framework PGR-associated data standardization and management will be the ground for harmonizing the various minimum information checklists related to a given dataset. In the following discussion, 'standard' and 'standardization' refer specifically to the regularization of how data is captured, represented, annotated, and reported. This does not pertain to experimental best practices but focuses on three main areas: (i) minimum information checklists or guidelines, (ii) data formats and structures (syntax), and (iii) controlled vocabularies and ontologies (semantics).

- **Addressing Fragmentation in Data Standards**, ensuring consistency in data collection and reporting, thereby facilitating more seamless data integration and enhancing overall data utility.
- **Facilitating Efficient Data Integration and Interoperability.** Integrating data sets that adhere to different minimum information checklists is a labour-intensive and complex task. Researchers often need to merge data from multiple sources to gain comprehensive insights into PGR. However, the current fragmentation of standards requires substantial manual effort to reconcile and harmonize disparate data sets. An integrative framework would provide unified guidelines that ensure compatibility across different data types, simplifying the data integration process and making it more efficient and effective.
- **Enhancing the Value of Secondary Data Use.** The increasing recognition of the value of secondary data use is another compelling reason to harmonize minimum information standards. Secondary data use involves reanalysing existing data sets to generate new insights beyond the original scope of data collection. This practice has gained prominence with the rise of data-driven research approaches, which rely on large, diverse data sets to uncover patterns and generate new hypotheses. Effective secondary use of data depends on well-documented and standardized data. Harmonizing minimum information standards would ensure consistent and comprehensive data annotation, making data sets more accessible and useful for secondary analyses. This maximizes the potential of existing data, promoting innovation and discovery.
- **Supporting Data-Driven Research.** Data-driven research requires the integration of diverse data types and sources, which is particularly relevant for PGR. This research might involve combining environmental data, phenotypic traits, genetic sequences, and more to draw

comprehensive conclusions. The independent development of minimum information standards has led to a fragmented data ecosystem, impeding such integrative approaches. By harmonizing these standards into a unified framework, researchers will have a common foundation for data collection, reporting, and sharing. This supports the growing trend of data-driven investigations by ensuring that data from different sources can be readily integrated and analysed together.

8. Core Features of MI-PGR

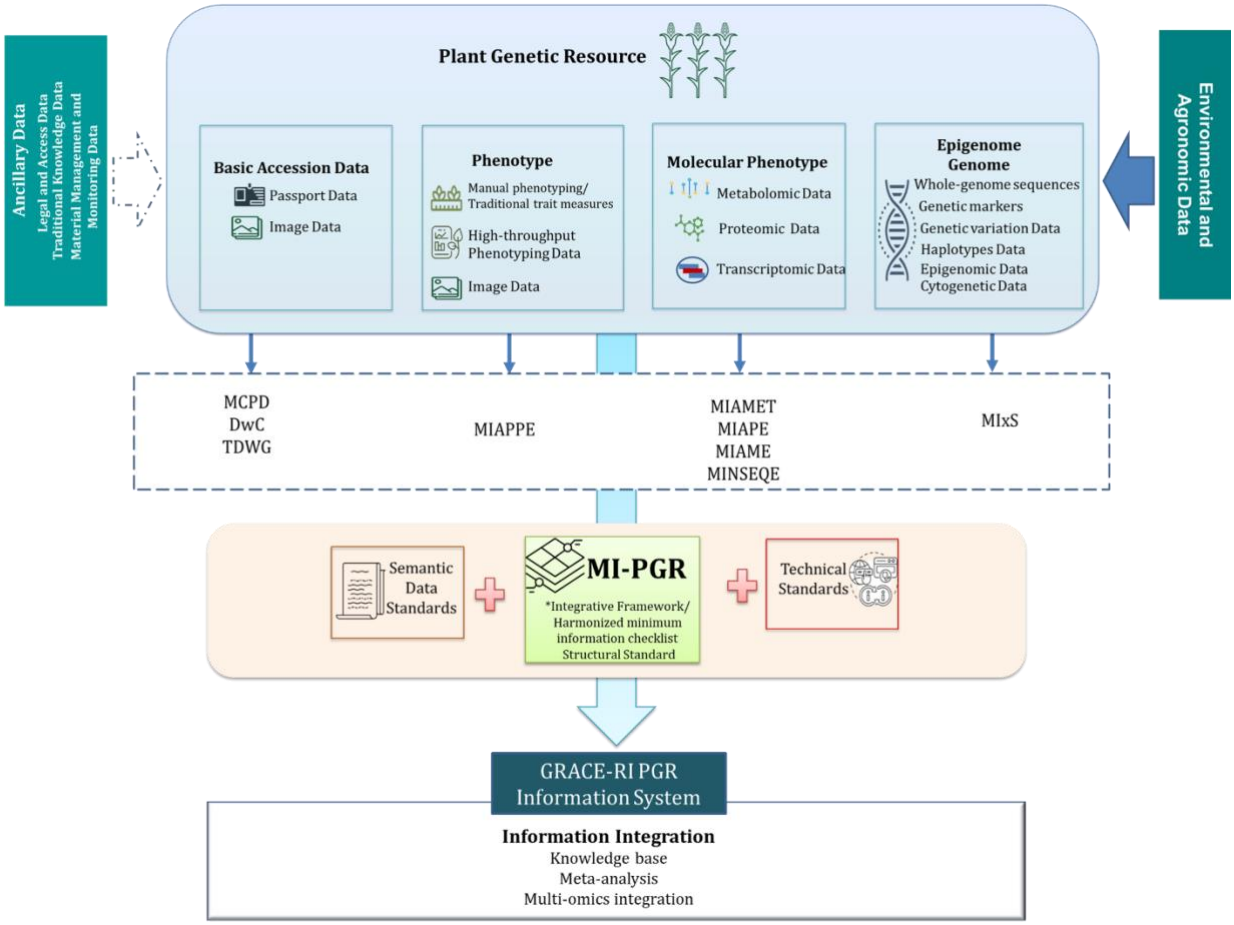

Figure 2. Conceptual framework of the proposed integrative framework or coherent minimum information checklist, MI-PGR.

Integrating existing documentation standards into a cohesive framework is a central feature of the proposed MI-PGR (Figure 2). This integration ensures that the framework builds upon the established minimum information/ data standards such as MCPD, DwC, MIAPPE, and MIxS while providing a unified approach to PGR documentation. This integrative approach addresses the fragmented nature of current documentation practices and paves the way for more comprehensive and standardized PGR data management.

Furthermore, this framework is designed to be scalable and adaptable, accommodating varying levels of data detail and complexity as presented in table 2. This scalability is achieved through the organization of information into hierarchical levels, each representing increasing degrees of detail and comprehensiveness. Institutions can start with basic identification data and progressively add more detailed information as their resources and needs evolve. This flexibility ensures that the framework can be implemented by a wide range of organizations, from small-scale genetic resource centers to large international research centers, each adapting the framework to fit their specific operational contexts and capacities.

Table 2. MI-PGR Levels of Minimum Information

Level	Information Level	Record Extent	Purpose	Data Handling Capabilities (EURISCO, Genesys & GRIN-Global)
1	Essential identification	Basic passport data (Mandatory); Inclusion of DOI is strongly recommended	To provide fundamental identification and traceability of PGR	All systems: Fully manages MCPD passport data. DOI inclusion is optional.
2	Detailed identification	Comprehensive passport data with initial image data	To provide essential, standardized details and visual records for precise identification and enhanced documentation of PGR, and facilitate global sharing and collaboration.	All systems: data comprehensiveness relies heavily on the submitting institute. GENESYS, GRIN: Image data handling is present but basic, with no advanced metadata standards.
3	Basic phenotypic traits	Basic morphological traits observable under standard conditions (following specific genebank protocols or international standards e.g. FAO, or specific crop consortia).	To document easily observable, stable and distinct characteristics of PGR for identification, conservation, cataloguing, and initial selection.	All systems: Capable of managing basic phenotypic traits, but not full MIAPPE-compliant
4	Detailed phenotypic evaluation traits and comprehensive image data	Detailed evaluation of phenotypic traits including agronomic performance under specific environmental conditions or stressors	To assess plant performance in diverse environments for breeding programs, agricultural improvement, understanding adaptability and resilience to environmental factors or stressors.	All systems: Capable of managing detailed phenotypic evaluation traits but not full MIAPPE-compliant GENESYS, GRIN: Limited image data handling
5	Molecular phenotype	Transcriptomic data: Reflecting gene expression patterns related to biochemical pathways and physiological processes. Proteomic data: Including proteins that regulate biochemical pathways and physiological responses. Metabolomic data: Detailing metabolites associated with	To enhance the utility of PGR for breeding programs aimed at nutritional improvement, disease resistance, or stress tolerance by providing detailed molecular phenotype information i.e.	All Systems: Do not manage molecular phenotype data (as of August 2024), and no submission mechanisms in place for such datasets.

Level	Information Level	Record Extent	Purpose	Data Handling Capabilities (EURISCO, Genesys & GRIN-Global)
		biochemical traits and physiological conditions.	transcriptomes, proteomes, and metabolomes.	
6	Genotypic information	Whole Genome Sequences; Genetic Variation Data: Information on SNPs, indels, and structural variants within the genome; Genetic markers used for genetic linkage and association studies. Epigenome: data on epigenetic modifications and their effects on gene expression.	To provide a molecular basis for the characterization, evaluation, and conservation of PGR, enhancing the understanding of genetic diversity and aiding in precise selection for breeding programs.	GRIN-Global: Minimal Genotypic Data, but not fully compliant with MIxS

Note on the Structure of Levels 4, 5, and 6:

While the proposed MI-PGR framework progresses sequentially from Levels 1 to 6 (Figure 3), it is important to note that Levels 4 (Detailed phenotypic evaluation traits and comprehensive image data), 5 (Molecular Phenotype), and 6 (Genotypic Information) are not strictly hierarchical in their application or importance. These levels represent specialized areas of data collection and analysis that are complementary rather than sequential.

The design of these levels acknowledges that advancements in one area may occur independently of others, and improvements or updates to data in one level can be made without necessitating changes across all levels. Thus, while Levels 1 to 3 lay the foundation for PGR accession identification, Levels 4 through 6 offer advanced, specialized data dimensions that contribute to a comprehensive understanding and utilization of PGR collections. This tiered approach facilitates a modular yet integrated system of information that supports diverse needs and applications in PGR conservation, research, breeding, and utilization.

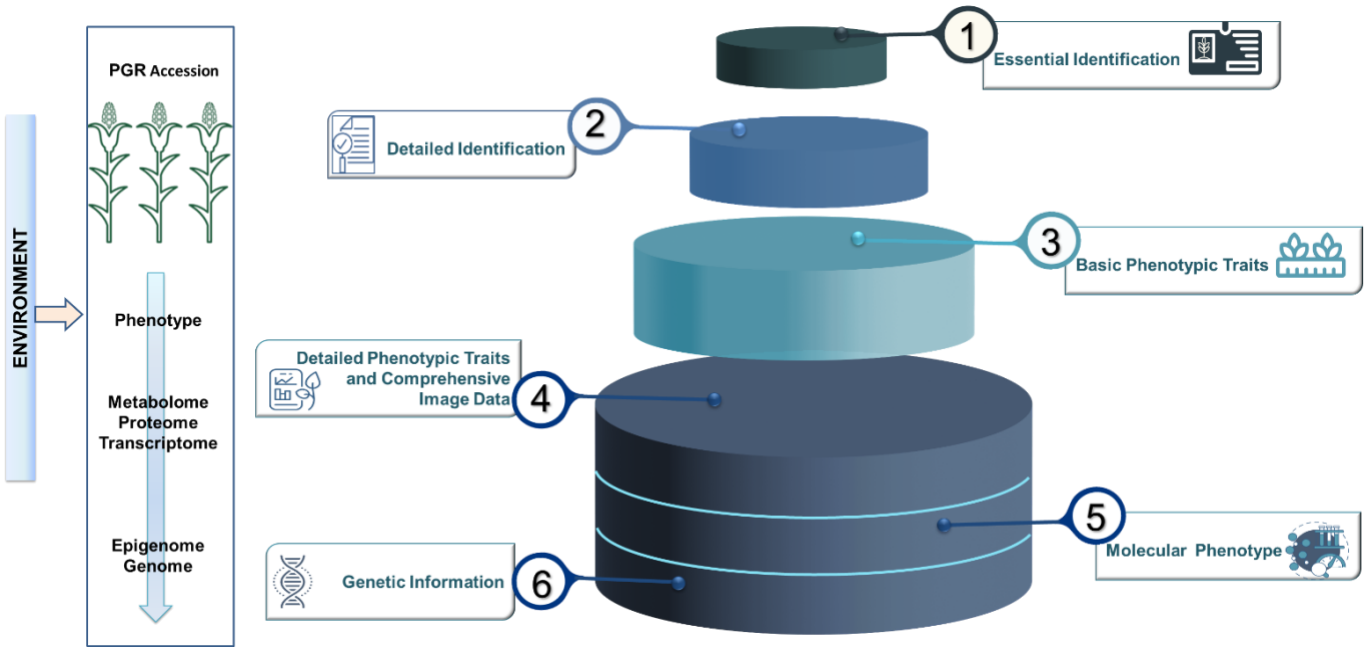

Figure 3. MI-PGR Levels of Minimum Information

Case in Point 4. Sexual System as an Important Descriptor? Significance and Considerations

For certain crop species, **sexual systems can vary significantly between wild and cultivated varieties, or even among different cultivars** (VanBuren *et al.*, 2015; Chavez-Pesqueira and Nuñez-Farfán, 2017; Dey *et al.*, 2023). Understanding and recording these differences have significant implications for various aspects of biodiversity, agriculture and food security, including:

1. **Optimizing breeding strategies**. E.g., enhancement of fruit set, pollination efficiency, genetic gain, seed production and overall yield; selection of compatible parents and development of hybrids that can exploit heterosis (Charlesworth, 2006; Dey *et al.*, 2023).
2. **Conserving genetic diversity**. Sexual systems influence gene flow between populations and the random changes in allele frequencies (genetic drift). Knowledge of these systems helps in managing genetic diversity within crop populations, ensuring that both male and female plants (or various sexual forms) are conserved, maintaining a broad genetic base and mitigating the risks of inbreeding and genetic erosion (Barrett and Harder, 2017).
3. **Supporting ecosystem health and *in situ* conservation.** Accurately recording sexual systems is crucial for elucidating plant-pollinator interactions, seed dispersal mechanisms and overall ecosystem dynamics. This understanding, in turn, informs ecological studies, aids in the conservation of pollinator species, and is integral for the survival and reproduction of diverse CWR, WFPs and other plant species conserved *in situ* (Kearns and Inouye *et al.*, 1997).
4. **Enhancing crop management.** It informs planting and pollination strategies to maximize yield and quality (Ghazoul, 2005; Boopalakrishnan *et al.*, 2021).
5. **Facilitating genetic mapping**. Understanding sexual systems can aid in mapping genes related to sex determination and reproductive traits, providing valuable data for genomic studies (Ming *et al.*, 2011; Dey *et al.*, 2023).

However, there are notable challenges and dilemmas associated with recording this information, owing to the variability and instability of this trait. In some species, sex expression is not static, but rather highly dynamic, influenced by epigenetic, environmental, and physiological factors (Dey *et al.*, 2023; Lou *et al.*, 2023). As a result, the stability of sex expression can vary considerably among different accessions of the same species. For example, in cucumbers (*Cucumis sativus*), which exhibit diverse sexual systems including monoecy, gynoecy, andromonoecy, and trimonoecy, sex expression is highly plastic and influenced by multiple factors such as temperature, photoperiodism, hormonal signals and other environmental stressors that can induce epigenetic modifications (Xinxin *et al.*, 2015; Boualem *et al.*, 2015; Dey *et al.*, 2023; Lou *et al.*, 2023). With such, several specific challenges include:

1. As mentioned above, **consistency and reliability of sexual system of an accession may vary across different environments or growth stages**. Recording a single sexual system descriptor may not accurately represent the accession's reproductive strategy, leading to inaccurate or misleading data.
2. **Ambiguity** in interpreting and categorizing highly variable sex expression can lead to inconsistencies in data recording, reducing the descriptor's reliability.
3. Data collection is **complex and cannot be captured in a single observation**. Continuous monitoring and repeated observations are necessary to capture the full range of sex expression variability
4. There may be a **lack of standardized protocols** for recording sexual systems, leading to variations in how different researchers and institutions document this information.

Given the complexities involved, it is essential to carefully weigh the benefits and drawbacks of including sexual systems as a descriptor. A strategic approach might be necessary, prioritizing the recording of sexual systems for species where this information has significant implications for breeding, conservation, and crop management. Conversely, for species where sexual systems do not vary significantly at the species and accession level and remain stable across different environments, recording this information may be less critical.

8.1 MI-PGR Level 1: Essential Identification

Level 1 (Essential Identification) introduces a universal entry point for the identification of a PGR *ex situ* accession or *in situ*-maintained population (i.e. CWR, WFP and LR). At this foundational level, the emphasis is on the establishment of a minimum but mandatory set of data that guarantees the basic traceability and recognition of each accession across institutions and global platforms. This level ensures that any PGR can be uniquely and consistently identified in compliance with the initial steps in PGR conservation, management and utilization. Additionally, the inclusion of PUID at this level is strongly recommended. This recommendation acknowledges the significant value of PUID in enhancing the traceability and accessibility of PGR. Despite this, the mandatory status of DOIs is deferred, recognizing that not all accessions and *in situ*-maintained populations are currently assigned a PUID, often due to logistical, historical, or technical constraints (See Case in Point 3).

Data elements are systematically divided into two distinct tables, each addressing one of the main PGR conservation strategies ((1) *Ex situ* (2) *In situ* CWR/WFP and on-farm LR). Table 3.1, aligning with the FAO/Bioversity Multi-Crop Passport Descriptors (MCPD) version 2.1, provides the essential data elements for identifying accessions within genetic resource centers and other repositories where PGR collections are conserved *ex situ*. In parallel, Table 3.2 outlines the data elements for documenting PGR populations conserved *in situ* and on-farm based on the Descriptors for Crop Wild Relatives (CWRI v.1) and proposed EURISCO descriptors for *in situ* CWR and on-farm LR[70]. Both tables provide a framework of the minimum mandatory and strongly recommended information, referred to as MI-PGR Level 1.

A new data element, TAXONID (highlighted in blue) (see Case in Point 1), is being proposed for comprehensive discussion and consideration. For INSTCODE, MNGINSTCODE and LIAISONCODE, the use of Research Organization Registry (https://ror.org/),[71] a community-led registry of open persistent identifiers for research organizations, is also being proposed.

Table 3.1 MI-PGR Level 1 Data Elements and Mappings (*Ex situ* PGR accessions)

Descriptor	Description	Example	Crosswalk Equivalents		
			EURISCO	MCPD v2.1	Darwin Core
0. ‡Persistent unique identifier (PUID)	Any persistent, unique identifier assigned to the accession so it can be unambiguously referenced at the global level and the information associated with it harvested through automated means. Report one PUID for each accession; The Secretariat of the International Treaty on PGR is facilitating the assignment of a persistent unique identifier (PUID), in the form of a DOI, to PGR at the accession level. (http://www.planttreaty.org/doi).	10.18730/1PGAP	PUID	PUID	GlobalUniqueIdentifier

[70] https://www.ecpgr.org/fileadmin/templates/ecpgr.org/upload/EURISCO/Third_meeting_of_the_EURISCO_AC_July_2021/D2.5_EURISCO_in_situ_extension_concept.pdf

[71] ROR is a global, community-led registry of open persistent identifiers for research organizations. The registry currently includes globally unique persistent identifiers and associated metadata for more than 105,000 research organizations. ROR IDs are specifically designed to be implemented in any system that captures institutional affiliations and to enable a richer networked research infrastructure.

1. ★Institute code (INSTCODE)	FAO WIEWS code of the institute where the *ex situ* accession is maintained. The codes consist of the 3-letter ISO 3166 country code of the country where the institute is located plus a number. The current set of institute codes is available from **http://www.fao.org/wiews.** **§The use of Research Organization Registry (https://ror.org/) is being proposed**	PHL001	INSTCODE	INSTCODE	institutionCode
2. ★Accession number (ACCENUMB)	This number serves as a unique identifier for accessions within a genebank, and is assigned when a sample is entered into the genebank collection.	IRGC 4	ACCENUMB	ACCENUMB	catalogNumber
3. ★Genus (GENUS)	Genus name for taxon. Initial uppercase letter required.	*Oryza*	GENUS	GENUS	genus
4. ★Species (SPECIES)	Specific epithet portion of the scientific name in lowercase letters. Following abbreviation is allowed: 'sp.'	*sativa*	SPECIES	SPECIES	specificEpithet
5. §Taxon ID (TAXONID)	A unique identifier for the taxon, as assigned by a taxonomic database or authority, e.g.: https://www.ncbi.nlm.nih.gov/taxonomy GBIF Backbone Taxonomy. https://doi.org/10.15468/39omei https://www.catalogueoflife.org/ Integrated Taxonomic Information System (ITIS). https://doi.org/10.5066/f7kh0kbk Proposed strategy: attribute-value pair structure (TaxonID_Source + TaxonID_Value)	NCBI4565			taxonID

★Mandatory; ‡Strongly recommended; §Open question

Table 3.2 MI-PGR Level 1 Data Elements and Mappings (*In situ*- maintained populations)

Descriptor	Description	Example	Crosswalk Equivalents		
			EURISCO	CWRI v.1	Darwin Core
0. ‡Persistent unique identifier (PUID)	Any persistent, unique identifier assigned to the accession so it can be unambiguously referenced at the global level and the information associated with it harvested through automated means. Report one PUID for each accession. The Secretariat of the International Treaty on PGR is facilitating the assignment of a persistent unique identifier (PUID), in the form of a DOI, to PGR at the accession level. (http://www.planttreaty.org/doi). **NOTE:** This descriptor should be assigned only to those CWR populations that are considered as long-term available sources of germplasm (e.g. the population is being monitored and potentially available under the terms of the MLS).	10.18730/1PGAP	PUID	PUID	GlobalUniqueIdentifier
1. ★Managing Institute code (MNGINSTCODE)	FAO WIEWS code of the institution responsible for, and/or organization that manages the *in situ* PGR population (e.g. protected area authority, nature reserve manager, national park manager, private landowner/farmer, etc.). The codes consist of the three-letter ISO 3166 country code		INSTCODE		institutionCode

	of the country where the institute is located plus a number. The current set of institute codes is available from http://www.fao.org/wiews. • If new institute codes are required, they can be generated online by FAO NFPs: (https://www.fao.org/cgrfa/overview/national-focal-point/en) or they can be requested from: WIEWS@fao.org. • In case no FAO WIEWS code of the institution responsible for, and/or organization that manages the CWR population is available and cannot be generated, the code ('DUMMY') can be used. • For institutes that no longer exist, or that were not assigned a FAO WIEWS institute code, please provide full details in the descriptors MNGINSTNAME and LIAISONNAME, respectively. [§]The use of Research Organization Registry (https://ror.org/) is being proposed				
2. ★Managing institute, legal entity or individual name (MNGINSTNAME)	Name of the institute, legal entity, or individual managing the population (e.g. protected area authority, nature reserve manager, national park manager, private owner, etc.,). **Note**: **This descriptor should be used only if MNGINSTCODE is not available**	Fauna and Wild Services, Ministry of the Interior 1453, Nicosia	INSTNAME	MNGINSTNAME	

3. Country of occurrence (ORIGCTY)	Country where the CWR population was observed or inventoried. Use the Three-letter ISO 3166-1 code of the country where the site is located.		ORIGCTY	ORIGCTY	
4. ★Observation date [YYYYMMDD] (OBSDATE)	The most recent date the population was observed, where YYYY is the year, MM is the month and DD is the day. Missing data (MM or DD) should be indicated with hyphens or '00' [double zero].	19610327	ACQDATE	OBSDATE	
5. ★Liason institute code (LIAISONCODE)	FAO WIEWS code of the institution that can liaise between the organization managing the CWR population and the interested user. §The use of Research Organization Registry (https://ror.org/) is being proposed	DEU440	LIAISONCODE		
6. Liaison institute name (LIAISONNAME)	Name, and brief address, of the institution that can liaise between the organization managing the CWR population and the interested user. **Note: This descriptor should be used only if LIAISONCODE is not available**		LIAISONNAME		
7. ★Population identifier (POPID)	A unique identifier (sequential number or code) assigned to a population **(a group of individuals of the same species that are found in a specific, contiguous geographic area, exhibit genetic similarity, and interact within a shared ecological setting). The managing institute is responsible for assigning a unique population identifier to each distinct**	PSRR2931		POPID	occurrenceID

	population. If the managing institute does not provide an identifier, the Liaison Institute (identified by LIASONCODE) will assign the POPID. Note: This description deviates from CWRI v.1				
8. ★Genus (GENUS)	Genus name for taxon. Initial uppercase letter required.	*Medicago*		genus	NameBotanical/GenusOrMonomial
9. ★Species (SPECIES)	Specific epithet portion of the scientific name in lowercase letters. Following abbreviation is allowed: 'sp.'	*monspeliaca*		specificEpithet	Species
10. §Taxon ID (TAXONID)	A unique identifier for the taxon, as assigned by a taxonomic database or authority, e.g.: https://www.ncbi.nlm.nih.gov/taxonomy GBIF Backbone Taxonomy. https://doi.org/10.15468/39omei https://www.catalogueoflife.org/ Integrated Taxonomic Information System (ITIS). https://doi.org/10.5066/f7kh0kbk Proposed strategy: attribute-value pair structure (TaxonID_Source + TaxonID_Value)	NCBI4565			taxonID

★Mandatory; ‡Strongly recommended; §Open question

8.2 MI-PGR Level 2: Detailed Identification

Level 2 (Detailed Identification), detailed within Tables 4.1 and 4.2, extends beyond the fundamental data captured in Level 1 by integrating a broader range of descriptors in line with MCPD V.2.1, CWRI v.1. The inclusion of image data (proposed descriptors) at this stage introduces a visual dimension to the passport data and aids in physical recognition and facilitating initial comparative analysis. It is highly recommended that all applicable data elements within this level are filled in to ensure the acquisition of a complete and informative dataset. The comprehensive passport data supports precise classification, streamlined search and retrieval, and assists in informed decision-making across research, breeding, and conservation initiatives.

For COLLCODE, BREDCODE and DONORCODE, the use of Research Organization Registry (https://ror.org/) is being proposed. For CROPNAME, the use of the AGROVOC Multilingual Thesaurus is strongly recommended.

For *in situ*-maintained populations, an additional descriptor being proposed is Conservation Status (CONSTATUS) (highlighted in blue). The existing SITEPROT descriptor indicates whether a site is under any legal or official legislation, while CONSACTION is particular to the IUCN scheme for conservation actions in place. However, both do not specify the type of conservation management applied. The proposed descriptor fills this gap by detailing the specific conservation environment and strategy. This distinction is crucial as it provides a comprehensive view of conservation efforts, applicable to CWR, WFP and LR.

Table 4.1 MI-PGR Level 2 Data Elements and Mappings (*Ex situ* PGR accessions)

(Note: This level requires at least the completion of data requirements of Level 1).

Descriptor	Description	Example	Crosswalk Equivalents		
			EURISCO	MCPD V2.1	Darwin Core
6. Collecting number (COLLNUMB)	Original identifier assigned by the collector(s) of the sample, normally composed of the name or initials of the collector(s) followed by a number (e.g. FM9909"). This identifier is essential for identifying duplicates held in different collections.	FA90-110	COLLNUMB	COLLNUMB	recordNumber
7. Collecting institute Code	FAO WIEWS code of the institute collecting the sample. If the holding	PHL001	COLLCODE	COLLCODE	collectingInstituteCode

Descriptor	Description	Example	Crosswalk Equivalents		
			EURISCO	MCPD V2.1	Darwin Core
(COLLCODE)	institute has collected the material, the collecting institute code (COLLCODE) should be the same as the holding institute code (INSTCODE). Follows INSTCODE standard. [§]The use of Research Organization Registry (https://ror.org/) is being proposed				
7.1 Collecting institute name (COLLNAME)	Name of the institute collecting the sample. **Note: This descriptor should be used only if COLLCODE cannot be filled because the FAO WIEWS code for this institute is not available.** Multiple values are separated by a semicolon without space.		COLLNAME	COLLNAME	
7.2 Collecting Institute Address (COLLINSTADDRESS)	Address of the institute collecting the sample. **Note: This descriptor should be used only if COLLCODE cannot be filled since the FAO WIEWS code for this institute is not available.** Multiple values are separated by a semicolon without space.		COLLINSTADDRESS	COLLINSTADDRESS	

Descriptor	Description	Example	Crosswalk Equivalents		
			EURISCO	MCPD V2.1	Darwin Core
7.2 Collecting mission identifier (COLLMISSID)	Identifier of the collecting mission used by the Collecting Institute (4 or 4.1)	CIATFOR-0512	COLLMISSID	COLLMISSID	
8. Subtaxon **(SUBTAXA)**	Subtaxa can be used to store any additional taxonomic identifier. Following abbreviations are allowed: "subsp." (for subspecies); "convar." (for convariety); "var." (for variety); "f." (for form).	subsp. japonica	SUBTAXA	SUBTAXA	infraspecificEpithet
9. Common Crop Name (CROPNAME)	Name of the crop in colloquial language, preferably English (i.e.' malting barley', 'cauliflower', or 'white cabbage') **The use of the AGROVOC Multilingual Thesaurus is strongly recommended.**	Rice	CROPNAME	CROPNAME	vernacularName
10. Accession Name (ACCENAME)	Either a registered or other formal designation given to the accession. First letter uppercase. Multiple names separated with semicolon without space.	Munji Sufaid	ACCENAME	ACCENAME	breedingIdentifier
11. Acquisition Date [YYYYMMDD] (ACQDATE)	Date on which the accession entered the collection where YYYY is the year, MM is the month and DD is the day. Missing data (MM or DD) should be indicated with hyphens or "00" (double zero).	19610327	ACQDATE	ACQDATE	acquisitionDate
12. Country of Origin (ORIGCTY)	3-letter ISO 3166-1 code of the country in which the sample was originally collected	MYS	ORIGCTY	ORIGCTY	countryCode

Note: Descriptors 13 to 16 below should be completed accordingly only if the accession was “collected”.

Descriptor	Description	Example	Crosswalk Equivalents		
			EURISCO	MCPD V2.1	Darwin Core
13. Location of Collecting Site (COLLSITE)	Location information below the country level that describes where the accession was collected.	7km east of Wageningen in the province of Gelderland	COLLSITE	COLLSITE	locality
14. Geographical coordinates	Latitude and longitude in decimal degree format with a precision of four decimal places corresponds to approximately 10 m at the Equator and describes the point-radius representation of the location.				
14.1 Latitude of collecting site (Decimal degrees format) (DECLATITUDE)	Latitude expressed in decimal degrees. Positive values are North of the Equator; negative values are South of the Equator.	-44.6975	DECLATITUDE	DECLATITUDE	decimalLatitude
14.2 Longitude of collecting site (Decimal degrees format) (DECLONGITUDE)	Longitude expressed in decimal degrees. Positive values are East of the Greenwich Meridian; negative values are West of the Greenwich Meridian.	+120.9123	DECLONGITUDE	DECLONGITUDE	decimalLongitude
14.3 Coordinate Datum (COORDATUM)	The geodetic datum or spatial reference system upon which the coordinates given in decimal latitude and decimal longitude are based (e.g. WGS84, ETRS89, NAD83). The GPS uses the WGS84 datum	WGS84	COORDATUM	COORDATUM	geodeticDatum
14.4 Georeferencing Method (GEOREFMETH)	The georeferencing method used (GPS, determined from map, gazetteer, or estimated using software). Leave the		GEOREFMETH	GEOREFMETH	georeferenceProtocol

Descriptor	Description	Example	Crosswalk Equivalents		
			EURISCO	MCPD V2.1	Darwin Core
	value empty if georeferencing method is not known				
15. Elevation of collecting site [meters above sea level (masl)] (ELEVATION)	Elevation of collecting site expressed in meters above sea level. Negative values are allowed.	800	ELEVATION	ELEVATION	minimumElevationInMeters
16. Collecting date of sample (COLLDATE)	Collecting date of the sample as YYYYMMDD. Missing data (MM or DD) should be indicated with hyphens. Leading zeros are required.	19900826	COLLDATE	COLLDATE	eventDate
17. Breeding Institute Code (BREDCODE)	FAO WIEWS code of the institute breeding the sample. If the holding institute has collected the material, the breeding institute code (BREDCODE) should be the same as the holding institute code (INSTCODE). Follows INSTCODE standard. [§]The use of Research Organization Registry (https://ror.org/) is being proposed	IND008	BREDCODE	BREDCODE	breedingInstituteID
17.1 Breeding institute name (BREDNAME)	Name of the institute (or person) that bred the material. **Note: This descriptor should be used only if BREDCODE cannot be filled**	Institute of Plant Breeding	BREDNAME	BREDNAME	

Descriptor	Description	Example	Crosswalk Equivalents		
			EURISCO	MCPD V2.1	Darwin Core
	because the FAO WIEWS code for this institute is not available. Multiple names are separated by a semicolon without space.				
18. Biological Status of Accession (SAMPSTAT)	The coding scheme proposed can be used at 3 different levels of detail: either by using the general codes (in boldface) such as 100, 200, 300, 400 or by using the more specific codes such as 110, 120 etc. **100) Wild** 110) Natural 120) Semi-natural/wild 130) Semi-natural/sown **200) Weedy** **300) Traditional cultivar/landrace** **400) Breeding/research material** 410) Breeder's line 411) Synthetic population 412) Hybrid 413) Founder stock/base population 414) Inbred line (parent of hybrid cultivar) 415) Segregating population 416) Clonal selection	300	SAMPSTAT	SAMPSTAT	biologicalStatus

Descriptor	Description	Example	Crosswalk Equivalents		
			EURISCO	MCPD V2.1	Darwin Core
	420) Genetic stock 421) Mutant (e.g. induced/insertion mutants, tilling populations) 422) Cytogenetic stocks (e.g. chromosome addition/substitution, aneuploids, amphiploids) 423) Other genetic stocks (e.g. mapping populations) **500) Advanced or improved cultivar** (conventional breeding methods) **600) GMO (by genetic engineering)** **999) Other** (Elaborate in REMARKS field)				
19. Ancestral Data (ANCEST)	Information about pedigree or other description of ancestral information (e.g. parent variety in case of mutant or selection).	A pedigree 'Hanna/7*Atlas//Turk/8*Atlas' or a description 'mutation found in Hanna', 'selection from Irene' or 'cross involving amongst others Hanna and Irene	ANCEST	ANCEST	ancestralData; purdyPedigree
20. Collecting / Acquisition Source (COLLSRC)	The coding scheme proposed can be used at 2 different levels of detail: either by using the general codes (in boldface)	21	COLLSRC	COLLSRC	acquisitionSource

Descriptor	Description	Example	Crosswalk Equivalents		
			EURISCO	MCPD V2.1	Darwin Core
	such as 10, 20, 30, 40 or by using the more specific codes such as 11, 12 etc. **10) Wild habitat** 11) Forest or woodland 12) Shrubland 13) Grassland 14) Desert or tundra 15) Aquatic habitat **20) Farm or cultivated habitat** 21) Field 22) Orchard 23) Backyard, kitchen or home garden (urban, peri-urban or rural) 24) Fallow land 25) Pasture 26) Farm store 27) Threshing floor 28) Park **30) Market or shop** **40) Institute, Experimental station, Research organization, Genebank** **50) Seed company** **60) Weedy, disturbed or ruderal habitat** 61) Roadside 62) Field margin				

Descriptor	Description	Example	Crosswalk Equivalents		
			EURISCO	MCPD V2.1	Darwin Core
	99) Other (Elaborate in REMARKS field)				
21. Donor Institute Code (DONORCODE)	FAO WIEWS code of the donor institute. §The use of Research Organization Registry (https://ror.org/) is being proposed		DONORCODE	DONORCODE	donorInstituteID
21.1 Donor Institute Name (DONORNAME)	Name of the donor institute (or person). **Note: This descriptor should be only used of DONORCODE cannot be filled because the FAO WIEWS code for this institute is not available.**				
22. Donor accession number (DONORNUMB)	Number assigned to an accession by the donor.		DONORNUMB	DONORNUMB	donorsIdentifier
23. Other identification (numbers) associated with the accession (OTHERNUMB)	Any other identification (numbers) known to exist in other collections for this accession. Use the following system: INSTCODE:ACCENUMB;INSTCODE:ACCENUMB;… INSTCODE and ACCENUMB follow the standard described above and are separated by a colon. Pairs of INSTCODE and ACCENUMB are separated by a semicolon without space. When the institute is not known, the number should be preceded by a colon.		OTHERNUMB	OTHERNUMB	otherCatalogNumbers

Descriptor	Description	Example	Crosswalk Equivalents		
			EURISCO	MCPD V2.1	Darwin Core
24. Location of safety duplicates (DUPLSITE)	FAO WIEWS code of the institute(s) where a safety duplicate of the accession is maintained. Follow INSTCODE standard.	NOR051	DUPLSITE	DUPLSITE	safetyDuplicationInstituteID
25. Type of germplasm storage (STORAGE)	If germplasm is maintained under different types of storage, multiple choices are allowed, separated by a semicolon (e.g. 20;30). (Refer to FAO/IPGRI Genebank Standards 1994 for details on storage type.) **10) Seed collection** 11) Short term 12) Medium term 13) Long term **20) Field collection** **30) In vitro collection** **40) Cryopreserved collection** **50) DNA collection** **99) Other** (elaborate in REMARKS field)	13	STORAGE	STORAGE	storageCondition

Descriptor	Description	Example	Crosswalk Equivalents		
			EURISCO	MCPD V2.1	Darwin Core
26. MLS status of the material (MLSSTAT)	The status of the accession with regards to the Multilateral System of Access and Benefit-Sharing (MLS) of the International Treaty, if available. 0: not available under the MLS 1: Available under the MLS		MLSSTAT	MLSSTAT	
27. Remarks (REMARKS)	The remarks field is used to add notes or to elaborate on descriptors with value 99 or 999 (=Other). Prefix remarks with the field name they refer to and a colon. Separate remarks referring to different fields are separated by semicolons without space.		REMARKS	REMARKS	occurenceRemarks
INITIAL IMAGE DATA **Each accession may include multiple images, with each image assigned a unique image ID that combines the accession number with the image number (using a Concatenated Identifier System). For each image, the following descriptors are mandatory.**					
28. ★Image ID (IMAGEID)	Unique identifier for each image; structured to indicate relationships and ensure each image can be individually tracked and managed. (ex. ACCENUMB_image number)	IRGC 4_IMG001			
29. ★Image Type (IMAGETYPE)	Specifies if the image is “Primary”, “Detail”, or “Contextual” (i.e Primary image represents the main view; Detail images focus on specific traits; Contextual images provide environmental or setup context.)	Primary			

Descriptor	Description	Example	Crosswalk Equivalents		
			EURISCO	MCPD V2.1	Darwin Core
30. ★Date of Capture (IMAGEDATE)	Exact date when the image was taken.	20230415		eventDate	hasDateCreated
31. ★Material Type (MATYPE)	Type of specimen/ Nature of the plant material 10) Seeds 20) Leaves 30) Stems 40) Roots 50) Flowers 51) Male Flower 52) Female Flower 53) Hermaphroditic 54) Inflorescences 60) Fruits 61) Unripe 62) Ripe 63) Dried 70) Whole Plant 71) Seedling 72) Mature Plant 80) Plant tissue culture 90) Planting material 91) Bulbs 92) Rhizomes	20		preparations	preparationsText

Descriptor	Description	Example	Crosswalk Equivalents		
			EURISCO	MCPD V2.1	Darwin Core
	93) Corms 94) Stolons 95) Tubers **99) Other** (elaborate in IMAGEDESCRIPT field)				
32. ★Record Creator (RECREATOR)	Person/entity responsible for collection/observation for Origin data tracking and source attribution (Recommendation: ORCID if known)	ORCID: 0000-0002-1825-0097			recordedBy
33. ★Description of Image Content (IMAGEDESCRIPT)	Brief description of what the image portrays.	Lanceolate leaf, 15 cm length, 4 cm width, with acuminate apex, cuneate base, and serrated margins.			Description

★Mandatory; ‡Strongly recommended (note: for image data, applicable only if images are available); §Open question

Table 4.2 MI-PGR Level 2 Data Elements and Mappings (*In situ*-maintained PGR population)

(Note: This level requires at least the completion of data requirements of Level 1).

Descriptor	Description	Example	Crosswalk Equivalents		
			EURISCO	CWR v1	Darwin Core
11. Subtaxon (SUBTAXA)	Subtaxa can be used to store any additional taxonomic identifier. Following abbreviations are allowed: "subsp." (for	subsp. japonica	SUBTAXA	SUBTAXA	infraspecificEpithet

Descriptor	Description	Example	Crosswalk Equivalents		
			EURISCO	CWR v1	Darwin Core
	subspecies); "convar." (for convariety); "var." (for variety); "f." (for form).				
12. Location of occurrence site (OCCURSITE)	Location information below the country level where the population sample was observed. This might include the distance in km and direction from the nearest town, village or map grid reference point (e.g. 7km east of Wageningen in the province of Gelderland)		COLLSITE	OCCURSITE	locationRemarks
13. Latitude of collecting site (Decimal degrees format) (DECLATITUDE)	Latitude expressed in decimal degrees. Positive values are North of the Equator; negative values are South of the Equator.	-44.6975	DECLATITUDE	DECLATITUDE	decimalLatiitude
14. Longitude of collecting site (Decimal degrees format) (DECLONGITUDE)	Longitude expressed in decimal degrees. Positive values are East of the Greenwich Meridian; negative values are West of the Greenwich Meridian.	+120.9123	DECLONGITUDE	DECLONGITUDE	decimalLongitude
15. Coordinate Datum (COORDATUM)	The geodetic datum or spatial reference system upon which the coordinates given in decimal latitude and decimal longitude are based (e.g. WGS84, ETRS89, NAD83). The GPS uses the WGS84 datum	WGS84	COORDATUM	COORDATUM	geodeticDatum
16. Elevation of site [meters above sea level (masl)] (ELEVATION)	Elevation of site expressed in meters above sea level. Negative values are allowed.	800	ELEVATION	ELEVATION	minimumElevationIn Meters

Descriptor	Description	Example	Crosswalk Equivalents		
			EURISCO	CWR v1	Darwin Core
17. Status of occurrence site (POPSRC)	Habitat of the occurrence site of the population (s). The coding scheme can be applied either by using the general codes (in boldface) or the more specific codes. Multiple values are separated by a semicolon without space. **10)Wild** 11) Forest or woodland 12) Shrubland 13) Grassland 14) Desert or Tundra 15) Aquatic Habitat **20) Farm or cultivated area** 21) Field 22) Orchard 23) Backyard, kitchen or home garden 24) Fallow land 25) Pasture 26) Park **60) Weedy, disturbed or ruderal habitat** 61) Roadside 62) Field margin **99) Others** (Elaborate in REMARKS field)		POPSRC	POPSRC	
18. Site Protection (SITEPROT)	Indicate whether the site is under any legal or official legislation. Follow IUCN Guidelines available at		**SITEPROT**	**SITEPROT**	protectedArea

Descriptor	Description	Example	Crosswalk Equivalents		
			EURISCO	CWR v1	Darwin Core
	https://www.iucn.org/theme/protected-areas/about/protected-area-categories 0) not protected 1) strict nature reserves 2) wilderness area 3) national park 4) natural monument or treasure 5) habitat/ species management area 6) protected landscape/seascape 7) protected area with sustainable use of natural resources 8) other effective conservation measures (OECM)				
19. Conservation actions in place (CONSACTION)	Indication whether conservation actions related to the population are in place. Use the IUCN classification scheme for conservation actions in place. 0) no conservation actions 1) Monitoring and planning 2) Land/water protection and management 3) Species management 4) Education and legislation 5) Other (Elaborate in REMARKS field)		CONSACTION	CONSACTION	conservationStatus
20. Conservation Status (CONSTATUS)	The type of conservation environment or strategy in place for the population. 0) No active conservation	2			

Descriptor	Description	Example	Crosswalk Equivalents		
			EURISCO	CWR v1	Darwin Core
	1) *In situ*- genetic reserve 2) *In situ*-protected area 3) *In situ*- managed wild 4) On-farm 5) Other (Elaborate in REMARKS field)				
21. Biological Status of Accession (SAMPSTAT)	The coding scheme proposed can be used at 3 different levels of detail: either by using the general codes (in boldface) such as 100, 200, 300 or by using the more specific codes such as 110, 120 etc. **100) Wild** 110) Natural 120) Semi-natural/wild 130) Semi-natural/sown **200) Weedy** **300) Traditional cultivar/landrace** **999) Other** (Elaborate in REMARKS field)	300	SAMPSTAT	SAMPSTAT	biologicalStatus
22. Code of the institute or herbarium holding *ex situ* samples **22.1 Institute code (INSTCODE)** **22.2 Index Herbariorum code**	FAO WIEWS institute code or Index Herbariorum code of the institute where the *ex situ* accession/herbarium specimen is maintained, or both. http://sweetgum.nybg.org/science/ih/			INSTCODE HERBCODE	

Descriptor	Description	Example	Crosswalk Equivalents		
			EURISCO	CWR v1	Darwin Core
(HERBCODE)					
23. Name of the institute or individual holding *ex situ* samples	Name of the institute, legal entity, herbarium, or individual where collected population samples are held (e.g., local or national genebank, herbarium or landowner). If the Managing institute holds the material, the holding institute name should be the same as the Managing institute. Note: This descriptor should be only used of INSTCODE and HERBCODE cannot be filled.			INSTNAME	acquisitionSource
23.1 Address of the holding organization or individual (INSTADDRESS)				INSTADDRESS	
24. Accession/specimen identifier	This is the unique identifier for accessions or specimens collected (e.g., genebank, herbarium, etc.) and is assigned when a sample/specimen is entered into the collection.				
24.1 *Ex situ* accession PUID (PUID)				ACCEDOI	

Descriptor	Description	Example	Crosswalk Equivalents		
			EURISCO	CWR v1	Darwin Core
24.2 *Ex situ* accession number (ACCENUMB) **24.3 Herbarium specimen number (SPECNUMB)**				ACCENUMBER SPECNUMB	
25. MLS status of the material (MLSSTAT)	The status of the material with regards to the Multilateral System of Access and Benefit-sharing of the International Treaty, if available. 0) Not available under the MLS 1) Available under the MLS		MLSSTAT	MLSSTAT	
26. Links to associated information (URL) (LINKS)	URL linking to additional data about the population.	http://gbis.ipk-gatersleben.de/gbis_i/detail.jsf?akzessionId=31805	ACCEURL	LINKS	
27. Remarks (REMARKS)	The remarks field is used to add notes or to elaborate on descriptors with value 99 or 999 (=Other). Prefix remarks with the field name they refer to and a colon. Separate remarks referring to different fields are separated by semicolons without space.		REMARKS	REMARKS	occurenceRemarks

Descriptor	Description	Example	Crosswalk Equivalents		
			EURISCO	CWR v1	Darwin Core
INITIAL IMAGE DATA. Each population may include multiple images, with each image assigned a unique image ID that combines the POPID with the image number (using a Concatenated Identifier System). For each image, please record the following descriptors					
28. ★Image ID (IMAGEID)	Unique identifier for each image; structured to indicate relationships and ensure each image can be individually tracked and managed. (ex. POPID_image number)	PSRR2931_IMG001			
29. ★Image Type (IMAGETYPE)	Specifies if the image is "Primary", "Detail", or "Contextual" (i.e Primary image represents the main view; Detail images focus on specific traits; Contextual images provide environmental or setup context.)	Primary			
30. ★Date of Capture (IMAGEDATE)	Exact date when the image was taken.	20230415		eventDate	hasDateCreated
31. ★Material Type (MATYPE)	Type of specimen/ Nature of the plant material 10) Seeds 20) Leaves 30) Stems 40) Roots 50) Flowers 51) Male Flower 52) Female Flower 53) Hermaphroditic 54) Inflorescences	20		preparations	preparationsText

Descriptor	Description	Example	Crosswalk Equivalents		
			EURISCO	CWR v1	Darwin Core
	60) Fruits 61) Unripe 62) Ripe 63) Dried 70) Whole Plant 71) Seedling 72) Mature Plant 80) Plant tissue culture 90) Planting material 91) Bulbs 92) Rhizomes 93) Corms 94) Stolons 95) Tubers 99) Other (elaborate in IMAGEDESCRIPT field)				
32. *Record Creator (RECREATOR)	Person/entity responsible for collection/observation for Origin data tracking and source attribution (Recommendation: ORCID if known)	ORCID: 0000-0002-1825-0097		recordedBy	
33. *Description of Image Content (IMAGEDESCRIPT)	Brief description of what the image portrays.	Lanceolate leaf, 15 cm length, 4 cm width, with acuminate apex, cuneate base, and serrated margins.		description	

★Mandatory; ‡Strongly recommended (note: for image data, applicable only if images are available); §Open question

8.3 MI-PGR Level 3: Basic phenotypic characteristics

Level 3 (Table 5) focuses on capturing specific data elements about the physical and observable characteristics of a PGR accession under standardized conditions (i.e. **traditional morphological characterization** following specific genebank protocols, international standards by FAO or specific crop consortia). The structured framework of data elements, including detailed descriptions for documenting the phenotypic profile of each accession at this level, is derived from MIAPPE v1.1. This approach represents a simplified version of MIAPPE, designed to balance comprehensiveness with practicality. It emphasizes pragmatic data collection, particularly in light of the varying capacities of different institutions. By focusing on essential and observable characteristics, this framework ensures thorough and consistent documentation, guaranteeing reproducibility and traceability while remaining mindful of resource constraints and institutional limitations.

Table 5. MI-PGR Level 3 Data Elements (Note: This level requires at least the completion of data requirements of MI-PGR Level 1 (referred to in MIAPPE as Biological material).

Descriptor	Description	MIAPPE Equivalent	Recommendations (relevant standards, ontologies and formats)
STUDY			
1. Study unique ID (STUDYID)	Unique identifier comprising the name or identifier of the institution and the identifier assigned to the study/ morphological characterization activity	Study unique ID	Unique identifier
2. Study title (STUDYTITLE)	Human-readable text summarizing the study/morphological characterization activity	Study title	Free text
3. Study description (STUDYDESCRIP)	Human-readable text describing the study/ morphological characterization activity	Study description	Free text
4. Start date (STARTDATE)	Date when the characterization activity commenced	Start date of study	YYYYMMDD (ACQDATE)
5. End date (ENDDATE)	Date when the characterization activity ended	End date of study	
6. Contact Institution Code (CHARACINSTCODE)	FAO WIEWS code of the institute responsible for the study. §The use of Research Organization Registry (https://ror.org/) is being proposed	Contact Institution	FAO WIEWS code

6.1 Contact Institution Name and Address (CHARACINSTNAME)	Name and Address of the institute responsible for the study. **Note: This descriptor should be used only if CHARACINSTCODE cannot be filled**. Multiple values are separated by a semicolon without space.		Research Organization Registry (https://ror.org/) Free text
7. Geographic location (country) (CHARACTCOUNTRY)	Code of the country where the study/characterization was carried out	Geographic location (country)	ISO 3166-1
8. Experimental site name (CHARACTSITE)	The name of the natural site, experimental field, greenhouse, etc. where characterization was carried out	Experimental site name	Free text
9. Geographic location (latitude) (DECLATITUDE)	Latitude of the study/characterization site in degrees, in decimal format.	Geographic location (latitude)	DECLATITUDE (MCPD) ISO 6709
10. Geographic location (longitude) (DECLONGITUDE)	Longitude of the study/ characterization site in degrees, in decimal format.	Geographic location (longitude)	DECLONGITUDE (MCPD) ISO 6709
11. Geographic location (altitude) (ELEVATION)	Altitude of the experimental site, provided in meters (m).	Geographic location (altitude)	ELEVATION (MCPD)
12. Description of the experimental design (EXPDESCRIPT)	Short description of the experimental design, possibly including statistical design. In specific cases, e.g. legacy datasets or data computed from several studies, the experimental design can be "unknown"/"NA", "aggregated/reduced data", or simply 'none'.	Description of the experimental design	Free text
13. Experimental design (EXPDESIGN)	Type of experimental design of the study	Type of experimental design	Crop Ontology
14. Growth Facility/ Growth environment type (GROWTHFAC)	Type of growth facility or environments in which the characterization was carried out (e.g., field, greenhouse)	Type of growth facility	XEML Environment Ontology, Crop Ontology
DATA FILE	A file or digital object holding observation data recorded during characterization, typically in tabular form. Multiple data files may be provided per study, and each file can include observations for several observation units and several observed variables.		

15. Data file link	Link to the data file (or digital object) in a database or in a persistent institutional repository	Data file link	
16. Data file description	Description of the format of the data file. May be a standard file format name, or a description of organization of the data in a tabular file.	Data file description	
17. Data file version	The version of the dataset (the actual data)	Data file version	
BIOLOGICAL MATERIAL (Complete data requirements of at least MI-PGR Level 1)			
ENVIRONMENT	Environment parameters that were kept constant throughout the study and did not change between observation units		
18. Temperature Range (TEMPRANGE)	The range of temperatures to which the PGR material is exposed during the study.	Environment parameters	WMO guidelines
19. Precipitation (PRECIP)	The amount of rainfall received in the study area (i.e field environment), measured in millimeters.	Environment parameters	
20. Soil description (SOILDSCRPT)	Description of the soil in field experiments or where accessions are planted and evaluated. This may include texture, pH, stoniness, drainage, and organic matter content.	Environment parameters	FAO/IPGRI Environment descriptors
21. Topography (TGRPHY)	Describes the physical configuration of the landscape where characterization is carried out, including features such as elevation, slope, and landforms		FAO/IPGRI Environment descriptors
22. Water availability (WTRAVLBL)	The primary source and extent of water accessible to plants in their growing environment.		FAO/IPGRI Environment descriptors
OBSERVATION UNIT (Synonym: Experimental Unit)	Observation units are objects that are subject to instances of observation and measurement. An observation unit comprises one or more plants, and/or their environment. There can be pure environment observation units with no plants. Synonym: Experimental unit.		
23. Observation unit ID (OBSRVID)	Identifier used to identify the observation unit in data files containing the values observed or measured on that unit: must be locally unique.	Observation unit ID	Unique identifier
24. Observation unit type (OBSRTYPE)	Type of observation unit in textual form, usually one of the following: block, sub-block, plot, sub-plot, pot, plant	Observation unit type	
SAMPLE			

25. Sample ID (SAMPID)	Unique identifier assigned to the plant material sample being characterized. Essential for ensuring that data on various traits are accurately linked to the correct sample across different observation or measurement times and methods.	Sample ID	Unique Identifier
26. Plant structure development stage (PLANTDEVSTAGE)	The stage in the life of a plant structure during which the sample was taken	Plant structure development stage	Plant Ontology
27. Plant anatomical entity (PLANTPART)	A description of the plant part (e.g. leaf) or the plant product (e.g. resin) from which the sample was taken	Plant anatomical entity	
28. Trait Data Collection date (TRAITCOLLDATE)	The date and time when the trait data were recorded	Collection date	
OBSERVED VARIABLE	An observed variable describes how a measurement has been made. It typically takes the form of a measured characteristic of the observation unit (plant or environmental trait), associated to the method and unit of measurement. Multiple variables with the same combination of trait, method and scale can be used in association with different plant parts (leaf 1, leaf 2), when this distinction is necessary for observations referring to different parts of the same observation unit.		
29. Variable ID (VARIABLEID)	Code used to identify the variable in the data file. We recommend using a variable definition from the Crop Ontology where possible. Otherwise, the Crop Ontology naming convention is recommended: <trait abbreviation>_<method abbreviation>_<scale abbreviation>). A variable ID must be unique within a given investigation.	Variable ID	
30. Variable name (VARIABLENAME)	Name of the observed variable	Variable name	
31. Variable accession number (VARACCNUM)	Accession number of the variable in the Crop Ontology	Variable accession number	Crop Ontology
32. Trait (TRAIT)	Name of the plant trait under observation	Trait	

33. Trait accession number (TRAITACCNUM)	Accession number of the trait in a suitable controlled vocabulary.	Trait accession number	Crop Ontology, Plant Trait Ontology, XML Environment Ontology
34. Method (METHOD)	Name of the method of observation	Method	
35. Method accession number (METHACCNUM)	Accession number of the method in a suitable controlled vocabulary.	Method accession number	Crop Ontology, Plant Trait Ontology, XML Environment Ontology
36. Method description (METHODESCRIPT)	Textual description of the method, which may extend a method defined in an external reference with specific parameters	Method description	
37. Reference associated to the method (METHREF)	URI/DOI of reference describing the method.	Reference associated to the method	
38. Scale (SCALE)	Name of the scale associated with the variable	Scale	
39. Scale accession number	Accession number of the scale in a suitable controlled vocabulary	Scale accession number	Crop Ontology
40. Time scale (TIMESCALE)	Name of the scale or unit of time with which observations of this type were recorded in the data file (for time series studies).		

Image Data: Each accession may include multiple images, with each image assigned a unique image ID that combines the study ID_accession number with the image number (using a Concatenated Identifier System). For each image, the following descriptors will be used.				
Checklist Section	**Descriptor**	**Description**	**Format/Example**	**Recommended Ontologies/Notes**
Image Acquisition	**0. Image ID (IMAGEID)**	Unique identifier for each image; structured to indicate relationships and ensure each image can be individually tracked and managed. (ex. STUDYID_ACCENUMB_image number)	IR546_IRGC 4_IMG001	

	1. Image Type (IMAGETYPE)	Specifies if the image is "Primary", "Detail", or "Contextual" (i.e Primary image represents the main view; Detail images focus on specific traits; Contextual images provide environmental or setup context.)	Primary	
	2. Date of Capture (IMAGEDATE)	Exact date when the image was taken.	20240218	YYYYMMDD
	3. Image Format (IMAGEFORMAT)	File format of the image (e.g., JPEG, PNG).	JPEG	Specify acceptable formats and any restrictions on file size or dimensions.
	4. Resolution (IMAGERES)	Resolution of the image in PPI.	300 PPI	PPI (Pixels per inch)
	5. ‡Description of Image Content (IMAGEDESCRIPT)	Brief description of what the image portrays.	Image of *Oryza sativa* at flowering stage, focusing on panicle structure.	Plant Trait Ontology, Plant Ontology
	6. Location of Image Capture (IMAGELOC)	Specific location where the image was taken.	IRRI Greenhouse, Philippines	Crop Research Ontology, Environment Ontology, ENVO
Image Annotation	**7. Annotated Phenotypic Trait (OBSERVEDTRAIT)**	Traits observed in the image.	Panicle type: Loose Grain color: Golden	Plant Trait Ontology, Crop Ontology
	8. Stage of Plant Development (PLANTDEVSTAGE)	Development stage of the plant.	Flowering stage	Plant Ontology, Crop Research Ontology
	9. Anatomical Part Captured (PLANTPART)	Organ or part of the plant shown in the image.	Panicle	Plant Ontology, Crop Ontology

	10. Annotation Notes (NOTES)	Additional notes about the annotation.		Provide context or any specific observations not captured by ontologies.
Image Processing	**11. Software Used (IMAGESOFT)**	Name of the software used for processing the image.	Adobe Photoshop 2024.1	Include version number used if relevant for reproducibility.
	12. Processing actions (IMAGEPROCESS)	Specific actions (if applicable) performed on the image.	Cropped	Detailing modifications aids in understanding the alterations and maintaining data integrity.
	13. Processing parameters (IMAGEPARA)	Parameters for each action taken.	Crop: 10% from top	Allows precise replication of processing steps for scientific verification and reproducibility.
	14. Record Creator (RECREATOR)	Person/entity responsible for collection/observation for Origin data tracking and source attribution (Recommendation: ORCID if known)		
Image Utilization	**15. Usage Rights (IMAGERIGHTS)**	Copyright and usage permissions for the image.	CC-BY-SA	Specify licensing (e.g., CC-BY, Public Domain), and any restrictions.
	16. Storage Location (IMAGESTORE)	Where the image is digitally stored.		Include details such as server, cloud service, or physical media
	17. Link to Associated Study/Assay (IMAGELINK)	Direct link to related studies or assays.	doi:10.1000/journal.pone.0153000	Unique identifier links for reference

8.4 MI-PGR Level 4: Detailed Phenotypic Evaluation Traits and Comprehensive Image Data

Level 4 involves detailed documentation of phenotypic traits observed under various experimental and environmental conditions or stressors. **For thorough and extensive documentation at this level, it is strongly recommended to use the full scope MIAPPE v1.1.** https://github.com/MIAPPE/MIAPPE/tree/master/MIAPPE_Checklist-Data-Model-v1.1.

However, with practical constraints in mind, Level 4 is designed to be less exhaustive than MIAPPE while still providing a framework to capture phenotypic traits under various conditions. Moreover, the inclusion of image data at this level is driven by the reliance of many institutions on conventional methods for assessing and measuring plant traits, rather than fully adopting image-based phenotyping. As a result, raw images collected through traditional methods remain a crucial component of the documentation process.

Table 6. MI-PGR Level 4 Data Elements (Note: This level requires at least the completion of data requirements of MI-PGR Level 1 (referred to in MIAPPE as Biological material).

Descriptor	Description	MIAPPE Equivalent	Recommendations (relevant standards, ontologies and formats)
INVESTIGATION	Investigations are research programmes with defined aims. They can exist at various scales (for example, they could encompass a grant-funded programme of work, the various components comprising a peer-reviewed publication, or a single experiment).		
1. Investigation unique ID (INVESTID)	Identifier comprising the unique name of the institution/database hosting the submission of the investigation data, and the accession number of the investigation in that institution.	Investigation unique ID	
2. Investigation title (INVESTITLE)	Human-readable string summarising the investigation.	Investigation title	
3. Investigation description (INVESTDESCRIPT)	Human-readable text describing the investigation in more detail.	Investigation description	
STUDY	A study (or experiment) comprises a series of assays (or measurements) of one or more types, undertaken to answer a particular biological question.		
4. Study unique ID (STUDYID)	Unique identifier comprising the name or identifier of the institution and the identifier assigned to the study/ morphological characterization activity	Study unique ID	Unique identifier

5. Study title (STUDYTITLE)	Human-readable text summarizing the study/morphological characterization activity	Study title	Free text
6. Study description (STUDYDESCRIP)	Human-readable text describing the study/ morphological characterization activity	Study description	Free text
7. Start date (STARTDATE)	Date when the characterization activity commenced	Start date of study	YYYYMMDD (ACQDATE)
8. End date (ENDDATE)	Date when the characterization activity ended	End date of study	
9. Contact Institution Code (CHARACINSTCODE) **9.1 Contact Institution Name and Address (CHARACINSTNAME)**	FAO WIEWS code of the institute responsible for the study. §The use of Research Organization Registry (https://ror.org/) is being proposed Name and Address of the institute responsible for the study. **Note: This descriptor should be used only if CHARACINSTCODE cannot be filled**. Multiple values are separated by a semicolon without space.	Contact Institution	FAO WIEWS code Research Organization Registry (https://ror.org/) Free text
10. Geographic location (country) (CHARACTCOUNTRY)	Code of the country where the study/characterization was carried out	Geographic location (country)	ISO 3166-1
11. Experimental site name (CHARACTSITE)	The name of the natural site, experimental field, greenhouse, etc. where characterization was carried out	Experimental site name	Free text
12. Geographic location (latitude) (DECLATITUDE)	Latitude of the study/characterization site in degrees, in decimal format.	Geographic location (latitude)	DECLATITUDE (MCPD) ISO 6709
13. Geographic location (longitude) (DECLONGITUDE)	Longitude of the study/ characterization site in degrees, in decimal format.	Geographic location (longitude)	DECLONGITUDE (MCPD) ISO 6709
14. Geographic location (altitude) (ELEVATION)	Altitude of the experimental site, provided in meters (m).	Geographic location (altitude)	ELEVATION (MCPD)

15. ‡Description of the experimental design (EXPDESCRIPT)	Short description of the experimental design, possibly including statistical design. In specific cases, e.g. legacy datasets or data computed from several studies, the experimental design can be "unknown"/"NA", "aggregated/reduced data", or simply 'none'.	Description of the experimental design	Free text
16. Experimental design (EXPDESIGN)	Type of experimental design of the study	Type of experimental design	Crop Ontology
17. ★Growth Facility/ Growth environment type (GROWTHFAC)	Type of growth facility or environments in which the characterization was carried out (e.g., field, greenhouse)	Type of growth facility	XEML Environment Ontology, Crop Ontology
18. Cultural practices (CULTPRACTICES)	General description of the cultural practices of the study.	Cultural practices	Free text
PERSON	A human involved in the investigation or specifically any of its studies.		
19. Person name (CONTNAME)	The name of the main contact person, either their full name or the name used in scientific publications, who is responsible for the data (e.g. collection, curation, management, and any inquiries related to it) **Note: Description deviates from MIAPPE**	Person name	Name
20. Person email (CONTEMAIL)	The electronic mail address of the person.	Person email	Address
21. Person ID (CONTID)	An identifier for the data submitter. If that submitter is an individual, ORCID identifiers are recommended.	Person ID	Unique identifier
DATA FILE	A file or digital object holding observation data recorded during characterization, typically in tabular form. Multiple data files may be provided per study, and each file can include observations for several observation units and several observed variables.		
41. Data file link	Link to the data file (or digital object) in a database or in a persistent institutional repository	Data file link	
42. Data file description	Description of the format of the data file. May be a standard file format name, or a description of organization of the data in a tabular file.	Data file description	
43. Data file version	The version of the dataset (the actual data)	Data file version	
★BIOLOGICAL MATERIAL (Complete data requirements of at least MI-PGR Level 1)			

ENVIRONMENT	Environmental parameters that were kept constant throughout the study and did not change between observation units or assays. Environment characteristics that vary over time, i.e. environmental variables, should be recorded as Observed Variables (see below).		
22. Environment parameter	Name of the environment parameter constant within the experiment.	Environment parameters	Free text (see Appendix II)
23. Environment parameter value	Value of the environment parameter (defined above) constant within the experiment.	Environment parameter vakue	Free text
EXPERIMENTAL FACTOR	The object of a study is to ascertain the impact of one or more factors on the biological material. Thus, a factor is, by definition a condition that varies between observation units, which may be biotic (pest, disease interaction) or abiotic (treatment and cultural practice) in nature. Depending on the level of the data, an experimental factor can be either "what is the factor applied to the plant" (i.e. Unwatered), or the "environmental characterisation" (i.e. if no rain on unwatered plant: Drought ; if rain on unwatered plant: Irrigated)		
24. Experimental Factor type (EXPFCTRTYPE)	Name/Acronym of the experimental factor	Experimental Factor type	Free text
25. Experimental Factor description (EXPFCTRDESCR)	Free text description of the experimental factor. This includes all relevant treatments planification and protocol planned for all the plants targeted by a given experimental factor.	Experimental Factor description	Free text
26. Experimental Factor values (EXPFCTRVALS)	List of possible values for the factor (i.e. treatments)	Experimental Factor values	Free text
OBSERVATION UNIT	Observation units are objects that are subject to instances of observation and measurement. An observation unit comprises one or more plants, and/or their environment. There can be pure environment observation units with no plants. Synonym: Experimental unit.		
27. Observation unit ID (OBSUNITID)	Identifier used to identify the observation unit in data files containing the values observed or measured on that unit: must be locally unique.	Observation unit ID	Unique identifier
28. Observation unit type (OBSUNITTYPE)	Type of observation unit in textual form, usually one of the following: block, sub-block, plot, sub-plot, pot, plant	Observation unit type	Free text

29. Observation unit factor value (OBSUNITFCTRVAL)	List of values for each factor applied to the observation unit.	Observation unit factor value	Free text
SAMPLE	A sample is a portion of plant tissue harvested, non-harvested or extracted from an observation unit for the purpose of sub-plant observations and/or molecular studies. A sample must be used when there is a physical sample that needs to be stored and traced. Otherwise, observations made at the sub-plant level should be recorded as plant level observations using the observed variables to characterize the object of the observation (e.g. Berry sugar content, Fruit weight, Grain Protein content, Leaf 1 width, Leaf 2 width, Leaf 2 length).		
30. Sample ID (SAMPID)	Unique identifier assigned to the plant material sample being characterized. Essential for ensuring that data on various traits are accurately linked to the correct sample across different observation or measurement times and methods.	Sample ID	Unique Identifier
31. Plant structure development stage (PLANTDEVSTAGE)	The stage in the life of a plant structure during which the sample was taken	Plant structure development stage	Plant Ontology
32. Plant anatomical entity (PLANTPART)	A description of the plant part (e.g. leaf) or the plant product (e.g. resin) from which the sample was taken	Plant anatomical entity	
33. Trait Data Collection date (TRAITCOLLDATE)	The date and time when the trait data were recorded	Collection date	
OBSERVED VARIABLE	An observed variable describes how a measurement has been made. It typically takes the form of a measured characteristic of the observation unit (plant or environmental trait), associated to the method and unit of measurement. Multiple variables with the same combination of trait, method and scale can be used in association with different plant parts (leaf 1, leaf 2), when this distinction is necessary for observations referring to different parts of the same observation unit.		
34. Variable ID (VARIABLEID)	Code used to identify the variable in the data file. We recommend using a variable definition from the Crop Ontology where possible. Otherwise, the Crop Ontology naming convention is recommended: <trait abbreviation>_<method abbreviation>_<scale abbreviation>). A variable ID must be unique within a given investigation.	Variable ID	

35. Variable name (VARIABLENAME)	Name of the observed variable	Variable name	
36. ‡Variable accession number	Accession number of the variable in the Crop Ontology	**Variable accession number**	Crop Ontology
37. Trait (TRAIT)	Name of the plant trait under observation	Trait	
38. Trait accession number	Accession number of the trait in a suitable controlled vocabulary.	Trait accession number	Crop Ontology, Plant Trait Ontology, XML Environment Ontology
39. Method (METHOD)	Name of the method of observation	Method	
40. Method accession number	Accession number of the method in a suitable controlled vocabulary.	Method accession number	Crop Ontology, Plant Trait Ontology, XML Environment Ontology
41. Method description (METHODESCRIPT)	Textual description of the method, which may extend a method defined in an external reference with specific parameters	Method description	
42. Reference associated to the method	URI/DOI of reference describing the method.	Reference associated to the method	
43. Scale (SCALE)	Name of the scale associated with the variable	Scale	
44. Scale accession number	Accession number of the scale in a suitable controlled vocabulary	Scale accession number	Crop Ontology
45. Time scale	Name of the scale or unit of time with which observations of this type were recorded in the data file (for time series studies).		

Image Data: Each accession may include multiple images, with each image assigned a unique image ID that combines the study ID, accession number with the image number (using a Concatenated Identifier System). For each image, the following descriptors will be used.

Checklist Section	Descriptor	Description	Format/Example	Recommended Ontologies/Notes
Image Acquisition	**46. Image ID (IMAGEID)**	Unique identifier for each image; structured to indicate relationships and ensure each image can be individually tracked and managed. (ex. STUDYID_ACCENUMB_image number)	IRGC 4_IMG001	Unique identifier
	47. Image Type (IMAGETYPE)	Specifies if the image is "Primary", "Detail", or "Contextual" (i.e Primary image represents the main view; Detail images focus on specific traits; Contextual images provide environmental or setup context.)	Primary	
	48. Date of Capture (IMAGEDATE)	Exact date when the image was taken.	20240218	YYYYMMDD
	49. Stage of Study/ Experiment	The specific phase or period in the overall study/experiment timeline during which image was captured (e.g. "Germination," "Vegetative Stage," "Flowering," etc.)	Vegetative stage	Free text
	50. Image Format (IMAGEFORMAT)	File format of the image (e.g., JPEG, PNG).	JPEG	Specify acceptable formats and any restrictions on file size or dimensions.
	51. Resolution (IMAGERES)	Resolution of the image in PPI.	300 PPI	PPI (Pixels per inch)
	52. Description of Image Content (IMAGEDESCRIPT)	Brief description of what the image portrays.	Image of *Oryza sativa* at flowering stage, focusing on panicle structure.	Plant Trait Ontology, Plant Ontology
	53. Lighting conditions	Description of the lighting conditions during image capture	Natural light, Artificial light	Free text

	54. Location of Image Capture (IMAGELOC)	Specific location where the image was taken.	IRRI Greenhouse, Philippines	Crop Research Ontology, Environment Ontology, ENVO
Image Annotation	**55. Annotated Phenotypic Trait (OBSERVEDTRAIT)**	Traits observed in the image.	Panicle type: Loose Grain color: Golden	Plant Trait Ontology, Crop Ontology
	56. Stage of Plant Development (PLANTDEVSTAGE)	Development stage of the plant.	Flowering stage	Plant Ontology, Crop Research Ontology
	57. Anatomical Part Captured (PLANTPART)	Organ or part of the plant shown in the image.	Panicle	Plant Ontology, Crop Ontology
	58. Annotation Notes (NOTES)	Additional notes about the annotation.		Provide context or any specific observations not captured by ontologies.
Image Processing	**59. Software Used (IMAGESOFT)**	Name of the software used for processing the image.	Adobe Photoshop 2024.1	Include version number used if relevant for reproducibility.
	60. Processing actions (IMAGEPROCESS)	Specific actions (if applicable) performed on the image.	Cropped	Detailing modifications aids in understanding the alterations and maintaining data integrity.
	61. Processing parameters (IMAGEPARA)	Parameters for each action taken.	Crop: 10% from top	Allows precise replication of processing steps for scientific verification and reproducibility.
	62. Record Creator (RECREATOR)	Person/entity responsible for collection/observation for Origin data tracking and source attribution		

		(Recommendation: ORCID if known)		
Image Utilization	**63. Usage Rights (IMAGERIGHTS)**	Copyright and usage permissions for the image.	CC-BY-SA	Specify licensing (e.g., CC-BY, Public Domain), and any restrictions.
	64. Storage Location (IMAGESTORE)	Where the image is digitally stored.		Include details such as server, cloud service, or physical media
	65. Link to Associated Study/Assay (IMAGELINK)	Direct link to related studies or assays.	doi:10.1000/journal.pone.0153000	Unique identifier links for reference
	66. Link to Phenotypic data (PHENOLINK)	Direct link to detailed phenotypic data associated with the image.		

★Mandatory; ‡Strongly recommended

8.4 MI-PGR Level 5: Molecular Phenotype

Molecular phenotype encompasses a broad range of high-throughput biological data types that provide detailed insights into the functional state of cells, tissues, and organisms. This level includes data elements related to gene expression, protein profiles, and metabolite composition. Table 7.1 outlines the common metadata requirements for these molecular phenotype data types. Subsequent tables (7.2 to 7.4) detail the specific metadata elements for metabolomic, proteomic, and transcriptomic data, respectively, which were built upon MIAMET, MIAME, MIAPE and MINSEQE.

Table 7.1 Common metadata requirements/ set of core descriptors for molecular phenotype data types

Descriptor	Description	Example	Crosswalk Equivalents	Recommendations/ Notes
INVESTIGATION				
1. Investigation title **(INVESTITLE)**	Name of the project within which the study/ experiment was organized	Metabolomic Profiling of Tomato Varieties	MIAPPE: Investigation title	
2. Investigation description **(INVESTDESCRIPT)**	Human-readable text describing the investigation/ project in detail		MIAPPE: Investigation description	
STUDY				
3. Study unique ID	A unique identifier composed of a prefix denoting the institution conducting the study/experiment, followed by a sequential or random numerical code.	ETHZ-2023-004	MIAPPE: Study unique ID	Unique identifier
4. Study title	Human-readable text summarizing the study	Stress Response Metabolomics in Tomatoes	MIAPPE: Study title	Free text
5. Study description	Human-readable text describing the study in detail	metabolite profiling of various tomato cultivars to determine factors contributing to their stress response	MIAPPE: Study description	Free text

		using GC-MS and LC-MS		
6. Contact institution	Name and address of the institution responsible for the study	ETH Zurich, Institute of Agricultural Sciences, Switzerland		Use of FAO WIEWS code or ROR for institute identifier

Biological material (The following data fields are required to be completed in addition to the mandatory Level 1 data requirements)				
7. **Collection Date** [YYYYMMDD] (COLLDATE)	Date when the sample for analysis was collected.	20230702		ISO 8601.
8. **Location of collecting site** (COLLSITE) 8.1 Latitude of collecting site (Decimal degrees format) (DECLATITUDE) 8.2 Longitude of collecting site (Decimal degrees format) (DECLONGITUDE)	Geographic location where the sample for analysis was collected. Latitude expressed in decimal degrees. Positive values are North of the Equator; negative values are South of the Equator. Longitude expressed in decimal degrees. Positive values are East of the Greenwich Meridian; negative values are West of the Greenwich Meridian.	Zurich, Switzerland 47.3769 8.5417		
9. Developmental Stage	The growth stage of the plant at the time of sample collection.			Plant Ontology

10. Sample Provider Information (PROVINFO)	Information about the donor of the sample, including name and institution.	Elizabeth Jones, ETH Zurich		Use FAOWIEWS code to get institute code or use ROR For individuals, use ORCID. If none, indicate full name and name of the institution
ENVIRONMENT				
Broad-scale environmental context	Specifies the primary environmental system from which the sample or specimen was collected. It should have a broad, spatially coarse description of the environmental context, helping to understand the general nature of the sampling location (e.g. major ecosystems or biomes, desert, rainforest, etc)		MIxS: env_broad_scale	use of subclasses of EnvO's biome class: http://purl.obolibrary.org/obo/ENVO_00000428 is highly recommended
Local environmental context	Specifies the particular entities or conditions present in the immediate surroundings of the sample or specimen that are believed to have significant causal influences on it.		MIxS: env_local_scale	ENVO
Environmental medium	The environmental material(s) immediately surrounding the sample or specimen at the time of sampling		MIxS: env_medium	ENVO
DATA FILE				
Data File Link	Link to the data file (or digital object) (Raw data and processed datasets) in a database or in a persistent institutional repository			

Data File Description	Description of the format of the data file. May be a standard file format name, or a description of organization of the data in a tabular file.			
Data File Version	The version of the dataset (the actual data)			
LINKS				
9. Data repository identifier	A unique identifier assigned to a dataset when it is deposited in a public repository. This identifier provides a permanent link to the dataset, ensuring that it can be easily accessed, cited, and referenced			Identifiers from NCBI GEO, EMBL-EBI etc.

Table 7.2 Set of core descriptors for metabolomic data

Descriptor	Description	Example	Crosswalk Equivalents	Recommendations/ Notes
1. Tissue type	Biological material (tissue or organ) used for metabolite extraction.	Leaf tissue	MSI: Sample type	Plant ontology
2. Sample Preparation	Details on how samples were prepared (extraction method, solvents used, storage conditions).	Methanol extraction at -80°C	MSI: Sample Preparation	
3. Chromatography Details	Description of the chromatography method, if applicable (e.g., column, solvent system).		MSI: Chromatography	
4. Data type	Type of data generated from the instrument.	GC-MS, LC-MS		Standardize data in formats like mzML or NetCDF.
5. Metabolite Quantification	Method for quantifying metabolites (e.g., absolute or relative quantification).		MSI: Quantification	Specify the method (e.g., peak areas, internal standards).

6. Calibration Method	Instrument calibration method used (internal or external standards).		MSI: Calibration	Include reference standards or external calibration methods.
7. Quality Control (QC)	Measures taken for quality control (e.g., blanks, pooled samples).		MSI: Quality Control	Include details of QC samples, pooled standards, and blanks.

Table 7.3 Set of core descriptors for proteomic data

Descriptor	Description	Example	Crosswalk Equivalents	Recommendations/ Notes
1. Tissue type	Description of the tissue or organ used for protein extraction.	Leaf tissue	MIAPE: Sample Type	Plant ontology
2. Protein Extraction Protocol	Detailed procedure for extracting proteins from the sample.	Protein extraction using RIPA buffer	MIAPE: Sample Processing	Include details about buffer composition and extraction conditions.
3. Enzymatic Digestion	Method used for digesting proteins (e.g., trypsin digestion).	Trypsin digestion at 37°C overnight	MIAPE: Sample Processing	Specify enzyme, temperature, and incubation times.
4. Instrument Details	Mass spectrometer used for proteomics analysis, including manufacturer and model.		MIAPE: Instrument	
5. Data Type	Type of proteomics data generated	MS/MS spectra, mzML format	MIAPE: Data Type	
6. Peptide Identification	Software and algorithm used for identifying peptides and proteins from MS data.		MIAPE: Peptide Identification	Specify database used (e.g., UniProt, NCBI) and search parameters.

7. Protein Quantification	Method used for quantifying proteins		MIAPE: Quantification	Specify quantification method, software, and normalization steps.
8. Calibration Details	Calibration methods for the instrument.		MIAPE: Calibration	Include details of internal or external standards.
9. Quality Control (QC)	Measures taken for quality contro		MIAPE: Quality Control	Describe QC measures such as pooled standards, blank injections

Table 7.4 Set of core descriptors for transcriptomic data

Descriptor	**Description**	**Example**	**Crosswalk Equivalents**	**Recommendations/ Notes**
2. Tissue type	Biological material (tissue or organ) used for used for RNA extraction.	Leaf tissue	MIAME: Tissue	Plant ontology
2. RNA Extraction Protocol	Method used for RNA extraction, including reagents and enzymes used.	TRIzol extraction method	MIAME: RNA Preparation	
3. Library Preparation Method	Method used to prepare RNA libraries for sequencing (e.g., poly-A selection, ribo-depletion).	Poly-A selection for mRNA-Seq	MINSEQE: Library Preparation	
4. Sequencing Platform	Platform and sequencing technology used (e.g., Illumina, PacBio).	Illumina HiSeq 4000	MINSEQE: Platform	
5. Read Length	Length of reads generated during sequencing.		MINSEQE: Read Length	Specify single-end or paired-end sequencing
6. Read Depth	Sequencing depth, typically measured in number of reads or coverage.		MINSEQE: Sequencing Depth	

7. Read Alignment Method	Software and algorithm used for aligning RNA-seq reads to a reference genome.	HISAT2, reference genome: IWGSC RefSeq v1.0	MINSEQE: Read Alignment	Provide tool name, version, and reference genome.
8. Gene Annotation Method	Method used for annotating genes from aligned reads.		MINSEQE: Annotation	Use standard gene annotation databases

8.4 MI-PGR Level 6: Genetic Data

Table 8 outlines the proposed data elements necessary for describing genetic data (Level 6), including genetic markers and genomic sequencing, associated with PGR accessions. The structure of this level follows the framework of MIAPPE. The checklist is organized into sections covering Investigation, Study, Person, Biological Material, Environment, and specific parts dedicated to Genotyping/Genetic Markers and Genomic Sequences. Data elements and terminologies for metadata are aligned with MIAPPE and MCPD wherever applicable. The required data elements specific to a dataset type are derived from the GnpIS-specific standard for genetic marker data, and MIxS (i.e. MIGS, MIMS, MIMARKS, MISAG, MIMAG) for genome sequencing data. Any deviations from the definitions provided by these existing standards are explicitly indicated.

Table 8. Set of core descriptors for genetic data

Data element	**Description**	**Example**	**Equivalents**		**Notes/ Recommendations**
			MIxS	**GnpIS**	
INVESTIGATION					
1. Investigation title	Name of the project within which the genetic study/ sequencing/ experiment was organized **NOTE**: This description has minor deviations from MIAPPE and MIxS	Wheat Genome Sequencing Project	project name	ProjectAcronym	Free text

2. Investigation description	Human-readable text describing the investigation/ project in detail				
STUDY					
3. Study unique ID	A unique identifier composed of a prefix denoting the institution conducting the genetic study/experiment, followed by a sequential or random numerical code. **NOTE**: This metadata field is not captured in either MIxS or GnpIS.	CIMMYT-2023-005			Unique identifier
4. Study title	Human-readable text summarizing the study	Drought Tolerance in Wheat Varieties		ExperimentName	Free text
5. Study description	Human-readable text describing the study in detail	Investigation into the genetic basis of drought tolerance in various wheat cultivars using whole-genome sequencing.		Description	Free text
6. Study type	Type of genetic investigation based on its primary focus or methodology, often referring to the type of genotyping experiment conducted **NOTE**: This description deviates from GnpIS	Whole Genome Sequencing		ExperimentType	See Sequence Ontology (SO) and Ontology for Biomedical Investigations (OBI)

7. Contact institution	Name and address of the institution responsible for the study	International Maize and Wheat Improvement Center (CIMMYT), Mexico		Institution	Use of FAO WIEWS code or ROR for institute identifier
PERSON					
8. Person name	The name of the main contact person, either their full name or the name used in scientific publications, who is responsible for the data (e.g. collection, curation, management, and any inquiries related to it) **NOTE**: This description has minor deviations from GnpIS	Lopez, Jennifer		LastName, FirstName	Free text
9. Person email	The electronic mail address of the person.	jennifer.lopez@cimmyt.org		Email	Email address
10. Person ID	An identifier for the data submitter. If that submitter is an individual, ORCID identifiers are recommended.	0000-0002-1825-0098			ORCID identifiers are recommended.
BIOLOGICAL MATERIAL (The following data fields are required to be completed in addition to the mandatory Level 1 data requirements)					
11. Taxon ID	NCBI taxon id of the sample.	4565	samp_taxon_id		GSC recommends the use of NCBI taxonomy

12. Geographical location	The geographical origin of the sample as defined by the country or sea name followed by specific region name.	Mexico: Sonora	geo_loc_name		Country or sea names should be chosen from the INSDC country list (http://insdc.org/country.html), or the GAZ ontology (http://purl.bioontology.org/ontology/GAZ)
13. Geographical location (latitude)	The geographical origin of the sample as defined by the latitude	29.072967	lat_lon		Decimal degrees format
14. Geographical location (longitude)	The geographical origin of the sample as defined by the longitude	-110.955919			
15. Collection date	The time of sampling, either as an instance (single point in time) or interval. In case no exact time is available, the date/time can be right truncated i.e. all of these are valid times: 2008-01-23T19:23:10+00:00; 2008-01-23T19:23:10; 2008-01-23	2023-04-15T00:00:00Z	collection_date		ISO 8601
ENVIRONMENT					

16. Broad-scale environmental context	Specifies the primary environmental system from which the sample or specimen was collected. It should have a broad, spatially coarse description of the environmental context, helping to understand the general nature of the sampling location (e.g. major ecosystems or biomes, desert, rainforest, etc)	Temperate grassland biome	env_broad_scale		use of subclasses of EnvO's biome class: http://purl.obolibrary.org/obo/ENVO_00000428 is highly recommended
17. Local environmental context	Specifies the particular entities or conditions present in the immediate surroundings of the sample or specimen that are believed to have significant causal influences on it.	Irrigated wheat field with surrounding maize crops	env_local_scale		Use of EnVO is highly recommended
18. Environmental medium	Report the environmental material(s) immediately surrounding the sample or specimen at the time of sampling	Sandy loam soil with moderate organic content	env_medium		Use of subclasses of 'environmental material' (http://purl.obolibrary.org/obo/ENVO_00010483) is recommended
GENOTYPING/GENETIC MARKERS (data fields specific to genetic marker data)					
19. Genotype value type	The specific format or representation of genotype data used in the results of a genetic study: allelic dose, allelic frequency, phased genotyping, genotype, IUPAC	Allelic frequency		GenotypeValueType	

	NOTE: This description has minor deviations from GnpIS				
20. Protocol name	Name of the protocol used for genotyping	Illumina HiSeq 2500		ProtocolName	Free text
21. Hardware name	Name of the hardware used for the genotyping	Illumina HiSeq 2500			
22. Marker ID	unique identifier assigned to a specific genetic marker (e.g. SNP or QTL); typically includes an institution or project prefix combined with a unique numerical or alphanumeric sequence **NOTE**: This data field is not captured in GnpIS.	CIMMYT-12345			If the genetic markers are already known and catalogued, existing Marker IDs from established databases like dbSNP, Ensembl, or the International HapMap Project can be used
23. Marker type	Type of the marker used in the study	SNP		MarkerType	See GO and SO for ontology-based terms
24. Marker name	Name of the marker used in this experiment and originally detected as the polymorphism	SNP-AX-123456789			
25. Reference genome name	Name of the reference genome	IWGSC RefSeq v1.0			

26. Chromosome name	Name of the chromosome where the marker was found/positioned	3B		ChromosomeName	Mandatory if the marker is positioned
27. Contig name	Name of the chromosome where the marker was found/positioned	ctg12345		ContigName	Mandatory if the marker is positioned
28. Marker sequence name	Name of the sequence where the marker was found/positioned	Seq123456		MarkerSequenceName	Mandatory if there is a sequence for this marker
29. Marker sequence	Sequence where the marker was found/positioned	ATCGTACGTA...		MarkerSequence	Mandatory if there is a sequence for this marker
30. Variation	Variation of the polymorphism	A/T			Mandatory for SNP Discovery
31. 5' Flanking	Sequence flanking the variation on it's 5'	ATCGTACG...			Mandatory for SNP Discovery
32. 3'Flanking	Sequence flanking the variation on it's 3'	...TACGTAGC			Mandatory for SNP Discovery
GENOMIC SEQUENCES (data fields specific to gene sequencing data)					
33. Sequence ID	Unique identifier for the genetic sequence **NOTE**: This data field is not captured in MIxS	CIMMYT-SEQ-2023-001			

34. Sequencing method	Sequencing machine used. Where possible the term should be taken from the OBI list of DNA sequencers	Illumina HiSeq 2500	seq_meth		OBI list of DNA sequencers (http://purl.obolibrary.org/obo/OBI_0400103) is strongly recommended
35. Nucleic acid extraction	A link to a literature reference, electronic resource or a standard operating procedure (SOP), that describes the material separation to recover the nucleic acid fraction from a sample	doi:10.1002/pmic.200800562	nucl_acid_ext		PMID, DOI or URL
36. Nucleic acid amplification	A link to a literature reference, electronic resource or a standard operating procedure (SOP), that describes the enzymatic amplification (PCR, TMA, NASBA) of specific nucleic acids	doi:10.1002/pmic.200800563	nucl_acid_amp		PMID, DOI or URL
37. Read Length	Indicates the length of each read produced during sequencing	150 bp			
38. Depth of Coverage	Reflects how many times each base of the genome has been sequenced.	30x			
39. Assembly name	Name/version of the assembly provided by the submitter that is used in the genome browsers and in the community	IWGSC RefSeq v1.0	assembly_name		name and version of assembly

40. Assembly software	Tool(s) used for assembly, including version number and parameters	SPAdes v3.13.0	assembly_software		name and version of software, parameters used
41. Annotations	Provide functional information about the genomic sequences, such as gene locations, predicted functions, and regulatory elements.	Genes: 107374			Use standardized formats (e.g., GFF3, BED)
LINKS					
42. Data repository identifier	A unique identifier assigned to a dataset when it is deposited in a public repository. This identifier provides a permanent link to the dataset, ensuring that it can be easily accessed, cited, and referenced **NOTE**: This metadata field is not captured in either MIxS or GnpIS.	PRJNA123456			

9. Way Forward

While great strides have been made in technology and PGR data collection, significant challenges remain in managing, integrating, and using this data effectively. Large volumes of data from robust technologies tend to overwhelm and obfuscate trends and relationships that can otherwise be easily ascertained from PGR data, thereby limiting their utility and potential. Moreover, these technologies generate a mosaic of data types resulting in integration and interoperability problems.

This deliverable presents an overview of the current PGR data landscape and underscores issues ushered in by massive influxes of diverse and rich datasets during recent years. The influx of vast phenotypic, environmental, and genetic data has the potential to transform PGR conservation and utilization. However, to fully realize this potential, it is essential to make PGR-associated data more findable, accessible, interoperable, and reusable (FAIR) for all stakeholders. To this end, this deliverable introduces an integrative framework, named MI-PGR, that proposes a coherent and harmonized set of guidelines for data collection, representation, annotation, and reporting to better describe and understand a PGR accession.

MI-PGR draws from an array of established standards. Given the multifaceted and interdisciplinary nature of PGR data, it seeks to reduce fragmentation and improve consistency across diverse datasets, promote interoperability, and enhance the utility of data. This will allow various actors to comprehensively understand a PGR accession, thereby allowing them to comprehensively understand a PGR accession, utilize existing PGR-associated datasets for their intended purposes and consequently support data-driven research initiatives for the conservation and judicious use of PGR.

This deliverable represents the foundational step intended to stimulate ongoing discussions and collaborations among different stakeholders. As data standards typically evolve from initial proposals, subsequent steps will involve refining these guidelines to meet the specific data needs of the PGR conservation and use community.

Actionable steps going forward

- The next phase involves **deepening stakeholder engagement** to refine the MI-PGR. This includes consultations and active engagement with conservationists, researchers, breeders, data managers, policymakers, and other relevant actors to facilitate discussions and gain deeper insights into the challenges and opportunities in PGR data collection and management. It is important to engage in ongoing discussions about the inclusion of additional data elements such as taxon ID, ploidy, and sex to provide a comprehensive and accurate description of PGR accessions. Furthermore, **establishing a dedicated working group** comprised of a diverse and expert team will be essential. This group will lead the continuous development and implementation of MI-PGR, ensuring it meets the complex and varied needs of all stakeholders involved in PGR activities.
- **Collaboration and interactions with existing research infrastructures (RI), consortia and relevant communities** is another critical step. Leveraging the expertise of RIs such as EMPHASIS, ELIXIR, and DiSSCo and communities involved in data standardization will ensure that the MI-PGR complements existing data standards, thereby eliminating redundancies and facilitating seamless integration with established systems. This collaboration will help the PGR community work towards similar goals, ensuring coherent and sustainable progress in data management practices.
- **Validation studies as proof of concept** are essential to substantiate the efficacy and practical applicability of MI-PGR. These studies will rigorously test the framework in real-world settings, gauging its robustness and utility across diverse contexts. The findings will subsequently refine the framework by addressing data elements, descriptions, structure, and the use of controlled vocabularies and ontologies, ultimately creating a formalized and machine-readable description of PGR accessions. This process ensures the framework meets stakeholder needs and remains effective across various real-world scenarios and institutional capacities. Key areas to be assessed include:

- **Data Quality and Completeness**. This involves evaluating the framework's ability to capture comprehensive and accurate data across different settings. This involves assessing precision, accuracy, consistency, and identifying any gaps or missing information.
- **Standardization and Consistency**. Ensuring the framework promotes uniform data collection and reporting practices by verifying adherence to established guidelines and consistent replication across different sites and studies.
- **Scalability.** Assessing the framework's applicability to varying scales of data management, from small-scale projects to large national or international initiatives, including evaluating the framework's flexibility in accommodating varying institutional capabilities and resources, ensuring its utility across a spectrum of organizational scales.
- **Usability and Practicality**. Gathering feedback on the ease of use and practicality of the guidelines and tools to identify any challenges or areas for improvement in the framework's design and implementation, particularly in institutions with differing levels of data management capabilities.
- **Compliance with FAIR Principles**. Evaluating the metadata standards and overall data documentation practices to ensure compliance with FAIR principles

- The effective implementation of the MI-PGR requires the **development and adaptation of comprehensive digital tools and resources to support stakeholders** across different capacities in data collection and management. Key components include standardized templates for data collection and reporting. Complementing these templates are a range of instructional materials, including video tutorials, interactive courses, and written guides. These resources focus on essential topics utilizing templates effectively and adhering to guidelines. They aim to ensure that users of all skill levels can comprehensively understand and implement the data standard, offering practical insights and demonstrations.
- **Developing and implementing targeted communication strategies** is essential to raise awareness and emphasize the importance of utilizing data standards effectively. Many stakeholders are currently unaware of these standards and concerned that adherence may limit their scientific autonomy. It's crucial to highlight the benefits of standards in improving data quality, facilitating data exchange, and ensuring reproducible research outcomes. Demonstrating how adherence to standards has contributed to advancements in genetic research, conservation efforts, and agricultural practices will help alleviate concerns and promote broader adoption among scientists and stakeholders alike.
- **Implementing MI-PGR within EURISCO and future GRACE-RI** represents a crucial step towards enhancing the standardization and interoperability of PGR data across Europe. However, achieving this goal requires **a systematic and strategic approach tailored to EURISCO's diverse network of member institutions**. Serving as the central hub for 43 countries and over 400 institutions, EURISCO plays a pivotal role in coordinating PGR information management. Nevertheless, this initiative faces inherent challenges, including aligning heterogeneous datasets, addressing varying institutional capabilities in data stewardship, and ensuring widespread adoption and adherence to these standards.
- **Ensuring the reliability and utility of data related to PGR requires the development and implementation of quality management systems (QMS)**, alongside data standards. The effectiveness of this framework or any data standards for that matter is heavily influenced by the rigor of the procedure used in data collection. Implementing robust QMS that encompass regular audits, thorough staff training and strict adherence to standardized protocols is imperative. These systems ensure that data collection and evaluation processes are conducted consistently and precisely, thereby minimizing variability and enhancing data reliability. Likewise, best practices in experimental procedures and use of scientifically validated methodologies are essential for maintaining high data quality and integrity.

- **Developing strategies for better capturing data for PGR conserved *in situ* is crucial.** *In situ* conservation is inherently complex and, while equally important as *ex situ* efforts, has often received less attention. While initiatives like the CWR Project under the Crop Trust, as well as various regional efforts under projects like the PGR Secure have made significant progress in establishing guidelines for data collection in natural habitats, methodologies for characterization and integrating environmental data, more work is needed to fully capitalize on these efforts. To better support *in situ* conservation, it is essential to understand its unique dynamics, and establish flexible, pragmatic yet robust systems that can be tailored to different environments while maintaining a core set of standards. Regular consultations and collaborative workshops involving various stakeholders, with feedback incorporated from all levels of implementation, are paramount to ensure that these strategies are not only relevant and scientifically rigorous but also address the practical realities faced by those on the ground.

- **Advocating for the use of persistent and unique identifiers such as DOIs** to ensure that PGR data, which is widely dispersed across organizational and international boundaries, can be reliably referenced and accessed over time. While adopting DOIs can be challenging for many institutions, it is a necessary step towards achieving greater data interoperability and reproducibility. Persistent identifiers will eventually facilitate better integration of PGR data into broader scientific research and policy-making efforts.

Ultimately, high-quality, well-managed, FAIR data is crucial for transforming extensive datasets into actionable insights. By addressing current data practice gaps and enhancing stakeholder collaboration, we can also contribute to bridging the gap between PGR conservation and utilization, ensuring our efforts today support future generations.

References

Aebersold R, Mann M. (2016). Mass-spectrometric exploration of proteome structure and function. Nature. 537(7620):347-55. doi: 10.1038/nature19949. PMID: 27629641.

Alam, O., Purugganan, M.D. 2024. Domestication and the evolution of crops: variable syndromes, complex genetic architectures, and ecological entanglements, The Plant Cell, 36 (5): 1227–1241, https://doi.org/10.1093/plcell/koae013

Alercia, A., Diulgheroff, S., Mackay, M. (2015) FAO/Bioversity Multi-crop Passport Descriptors V. 2.1 [MCPD V. 2.1]-December 2015. Bioversity International. [WWW document] URL https://cgspace.cgiar.org/handle/10568/69166

Alercia, A., Diulgheroff, S., Metz, T., (2001). FAO/IPGRI Multi-Crop Passport Descriptors [MCPD]. Pp. 1-4. Food and Agriculture Organization of the United Nations and International Plant Genetic Resources Institute, Rome, Italy.

Alercia, A., López, F., Marsella, M., Cerutti, A.L. (2021) Descriptors for Crop Wild Relatives conserved *in situ* (CWRI v.1). Rome, FAO on behalf of the International Treaty on Plant Genetic Resources for Food and Agriculture. https://doi.org/10.4060/cb3256en

Alercia, A., López, F.M., Sackville Hamilton, N.R. and Marsella, M. (2018) Digital object identifiers for food crops—Descriptors and guidelines of the global information system. Food and Agriculture Organization of the United Nations, Rome.

Ali J, Jewel ZA, Mahender A, Anandan A, Hernandez J, Li Z. (2018) Molecular Genetics and Breeding for Nutrient Use Efficiency in Rice. Int J Mol Sci. https://doi.org/10.3390/ijms19061762.

Andrés-Hernández, L., Halimi, R.A., Mauleon, R., Mayes, S., Baten, A., King, G.J. (2021) Challenges for FAIR-compliant description and comparison of crop phenotype data with standardized controlled vocabularies. Database (Oxford). https://doi.org/10.1093/database/baab028

Araus, J.L. and Cairns, J.E. (2014) Field High-Throughput Phenotyping: The New Crop Breeding Frontier. Trends in Plant Science, 19, 52-61. https://doi.org/10.1016/j.tplants.2013.09.008.

Araus, J.L., Kefauver, S.C., Zaman-Allah, M., Olsen, M.S., Cairns, J.E. (2018). Translating high-throughput phenotyping into genetic gain. Trends Plant Sci. 23(5):451–466. https://doi.org/10.1016/j.tplants.2018.02.001.

Arend, D., Junker, A., Scholz, U., Schuler, D., Wylie, J., Lange, M. (2016). PGP repository: a plant phenomics and genomics data publication infrastructure. Database (Oxford), Volume 2016, baw033, https://doi.org/10.1093/database/baw033

Barrett, S. C. H., & Harder, L. D. (2017). The ecology of mating and its evolutionary consequences in seed plants. Annual Review of Ecology, Evolution, and Systematics, 48, 135-157. https://doi.org/10.1146/annurev-ecolsys-110316-022922.

Basu P, Kruse CPS, Luesse DR, Wyatt SE. (2022) Plant Proteomic Data Acquisition and Data Analyses: Lessons from Spaceflight. Methods Mol Biol. ;2368:199-214. doi: 10.1007/978-1-0716-1677-2_13. PMID: 34647257.

Bioversity International. (2007). Guidelines for the development of crop descriptor lists. Bioversity Technical Bulletin Series. Bioversity International, Rome, Italy. xii+72p.

Boualem, A., Troadec, C., Camps, C., Lemhemdi, A., Morin, H., Sari, M.-A., *et al.,* (2015). A cucurbit androecy gene reveals how unisexual flowers develop and dioecy emerges. Science 350 (6261), 688–691. https://doi.org/10.1126/science.aac8370

Brazma, A., Hingamp, P., Quackenbush, J. *et al.,* (2001) Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. Nat Genet 29, 365–371. https://doi.org/10.1038/ng1201-365

Brink, M. and van Hintum, T. (2020) Genebank Operation in the Arena of Access and Benefit-Sharing Policies. Front. Plant Sci. 10:1712. https://doi.org/10.3389/fpls.2019.01712

Brozynska, M., Furtado, A., Henry, R. J. 2016. Genomics of crop wild relatives: expanding the gene pool for crop improvement. Plant Biotechnol. J. 14, 1070–1085.https://doi.org/10.1111/pbi.12454

Canella, M., Ardenghi, N.M.G., Müller, J.V. *et al.* (2022). An updated checklist of plant agrobiodiversity of Northern Italy. Genet Resour Crop Evol 69, 2159–2178. https://doi.org/10.1007/s10722-022-01365-y

Castañeda-Álvarez, N.P., Vincent, H.A., Kell, S.P., Eastwood, R.J., Maxted, N., (2011). Ecogeographic surveys. In: Guarino, L. Ramanatha Rao, V. and Goldberg, E. (eds.), *Collecting Plant Genetic Diversity:*

Technical Guidelines. 2011 update. Bioversity International, Rome, Italy. Available online: http://cropgenebank.sgrp.cgiar.org/index.php?option=com_content&view=article&id=679

CBD, (1992). Convention on Biological Diversity: Text and Annexes. pp. 1-34. Secretariat of the Convention on Biological Diversity, Montreal, Canada.

Charlesworth, D. (2006). Evolution of plant breeding systems. Current Biology, 16(17), R726-R735. https://doi.org/10.1016/j.cub.2006.07.068.

Choi, HK. (2019). Translational genomics and multi-omics integrated approaches as a useful strategy for crop breeding. Genes Genom 41, 133–146. https://doi.org/10.1007/s13258-018-0751-8

Clough E, Barrett T. 2016. The Gene Expression Omnibus Database. Methods Mol Biol. 2016;1418:93-110. doi: 10.1007/978-1-4939-3578-9_5. PMID: 27008011; PMCID: PMC4944384.

Cochrane, G., Karsch-Mizrachi, I., Nakamura, Y. 2011. International Nucleotide Sequence Database Collaboration. The International Nucleotide Sequence Database Collaboration. Nucleic Acids Res. 39(Database issue):D15-8. doi: 10.1093/nar/gkq1150. Epub 2010 Nov 23. PMID: 21106499; PMCID: PMC3013722.

Crossley B.M., Bai J., Glaser A. *et al.,* (2020) Guidelines for Sanger sequencing and molecular assay monitoring. J. Vet. Diagn. Invest., 32, 767–775.

Curry, H.A. (2023). Data, Duplication, and Decentralisation: Gene Bank Management in the 1980s and 1990s. In: Williamson, H.F., Leonelli, S. (eds) Towards Responsible Plant Data Linkage: Data Challenges for Agricultural Research and Development. Springer, Cham. https://doi.org/10.1007/978-3-031-13276-6_9

Ćwiek-Kupczyńska, H., Altmann, T., Arend, D. *et al.,* (2016) Measures for interoperability of phenotypic data: minimum information requirements and formatting. Plant Methods 12, 44. https://doi.org/10.1186/s13007-016-0144-4

Dar FA, Mushtaq NU, Saleem S, Rehman RU, Dar TUH, Hakeem KR. (2022) Role of Epigenetics in Modulating Phenotypic Plasticity against Abiotic Stresses in Plants. Int J Genomics. https://doi.org/10.1155/2022/1092894.

Das, S., Massey-Reed, S.R., Mahuika, J., et al., 2022. A High-Throughput Phenotyping Pipeline for Rapid Evaluation of Morphological and Physiological Crop Traits Across Large Fields. IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium, Kuala Lumpur, Malaysia, 2022, pp. 7783-7786, doi: 10.1109/IGARSS46834.2022.9884530.

Dempewolf, H., Baute, G., Anderson, J., Kilian, B., Smith, C. and Guarino, L. (2017). Past and Future Use of Wild Relatives in Crop Breeding. Crop Science, 57: 1070-1082. https://doi.org/10.2135/cropsci2016.10.0885

Dempewolf, H., Eastwood, R. J., Guarino, L., Khoury, C. K., Müller, J. V., & Toll, J. (2014). Adapting Agriculture to Climate Change: A Global Initiative to Collect, Conserve, and Use Crop Wild Relatives. Agroecology and Sustainable Food Systems, 38(4), 369–377. https://doi.org/10.1080/21683565.2013.870629

Deng CH, Naithani S, Kumari S, Cobo-Simón I, Quezada-Rodríguez EH, Skrabisova M, Gladman N, Correll MJ, Sikiru AB, Afuwape OO, Marrano A, Rebollo I, Zhang W, Jung S. (2023) Genotype and phenotype data standardization, utilization and integration in the big data era for agricultural sciences. Database (Oxford). https://doi.org/10.1093/database/baad088.

de Vienne, D. (2022). What is a phenotype? History and new developments of the concept. Genetica 150, 153–158. https://doi.org/10.1007/s10709-021-00134-6

Dey, S. S., Sagar, V., Kujur, S. N., Pradeep, K. N., Munshi, A. D., Pandey, S., & Behera, T. K. (2023). Cucumber: Breeding and Genomics. Vegetable Science, 50:208-220.

Dong, Y., Duan, S., Xia, Q., Liang Z. *et al.* (2023). Dual domestications and origin of traits in grapevine evolution. Science. 379: 892-901. https://doi.org/10.1126/science.add8655

Ebert, A.W,, Engels, J.M.M., Schafleitner, R., Hintum, T.v., Mwila, G. (2023) Critical Review of the Increasing Complexity of Access and Benefit-Sharing Policies of Genetic Resources for Genebank Curators and Plant Breeders–A Public and Private Sector Perspective. Plants. 12(16):2992. https://doi.org/10.3390/plants12162992

ECPGR European Genebank Managers Network. (2024). Genebank Managers Network Available et: https://www.ecpgr.org/about/genebank-managers-network#:~:text=In%202023%2C%20the%20Steering%20Committee,leading%20and%20managing%20a%20genebank. (accessed 02.08.2024).

ECPGR. (2021). Plant Genetic Resources Strategy for Europe. European Cooperative Programme for Plant Genetic Resources, Rome, Italy.

Endresen, D.T.F., Knüpffer, H. (2012) The Darwin Core Extension for Genebanks Opens up New Opportunities for Sharing Genebank Datasets. Biodiversity Informatics 8 (1). https://doi.org/10.17161/bi.v8i1.4095

Engels, J.M., Visser, L. (2003). A Guide to Effective Management of Germplasm Collections; International Plant Genetic Resources Institute: Rome, Italy.

Engels, J.M., Maggioni, L. (2010) Managing germplasm in a virtual European genebank (AEGIS) through networking. In Theorien der Lebendsammlung: Pflanzen, Mikroben und Tiere als Biofakte in Genbanken. (Lebenswissenschaften im Dialog, Band 25); Karafyllis, N.C., Ed.; Verlag Karl Alber: Freiburg, Germany, pp. 169–197.

Engels JMM, Ebert AW, van Hintum T. (2024). Collaboration between Private and Public Genebanks in Conserving and Using Plant Genetic Resources. Plants (Basel). 13(2):247. https://doi.org/doi:10.3390/plants13020247.

Engels JMM, Ebert AW. (2021) A Critical Review of the Current Global *Ex situ* Conservation System for Plant Agrobiodiversity. II. Strengths and Weaknesses of the Current System and Recommendations for Its Improvement. Plants. 10(9):1904. https://doi.org/10.3390/plants10091904

EURISCO, (2015). *European Search Catalogue for Plant Genetic Resources (EURISCO)*. Available online: https://eurisco.ipk-gatersleben.de/

EURISCO, (2022). Descriptors for uploading *in situ* CWR passport data to EURISCO. Available at: https://www.ecpgr.org/fileadmin/templates/ecpgr.org/upload/WORKING_GROUPS/WILD_SPECIES/EURISCO_in_situ_CWR_descriptors.pdf. (accessed 02.08.2024).

Fahlgren, N., Gehan, M.A., Baxter, I. (2015). Lights, camera, action: High-throughput plant phenotyping is ready for a close-up. Curr. Opin. Plant Biol. 24:93–99. https://doi.org/10.1016/j.pbi.2015.02.006

FAO, (1998). *State of the World's Plant Genetic Resources for Food and Agriculture*. Food and Agriculture Organization of the United Nations, Rome, Italy. Available online: www.fao.org/agriculture/crops/thematic-sitemap/theme/seeds-pgr/sow/en/ (Accessed 06.03.2015).

FAO, (2010). Second report on the State of the World's Plant Genetic Resources for Food and Agriculture. Food and Agriculture Organization of the United Nations, Rome, Italy. Available online: http://www.fao.org/agriculture/seed/sow2/en/ [Accessed 25 July 2013].

FAO (2022) WIEWS: Ex Situ (SDG 2.5.1) – Overview. Rome. Available at https://www.fao.org/wiews/data/ex-situ-sdg-251/overvew/en (Accessed 06 May 2024)

Fasoula DA, Ioannides IM, Omirou M. (2020) Phenotyping and Plant Breeding: Overcoming the Barriers. Front Plant Sci. https://doi.org/10.3389/fpls.2019.01713.

Field, D., Garrity, G., Gray, T. *et al.,* (2008) The minimum information about a genome sequence (MIGS) specification. Nat Biotechnol 26, 541–547. https://doi.org/10.1038/nbt1360

Fiorani, F. and Schurr, U. (2013) Future Scenarios for Plant Phenotyping. Annual Review of Plant Biology, 64, 267-291. https://doi.org/10.1146/annurev-arplant-050312-120137

Frankel, O.H., Bennett, E. 1970. Genetic Resources in Plants—Their Exploration and Conservation; IBP Handbook No. 11; Blackwell Scientific Publications: Oxford, UK; 554p.

Furbank, R.T., Jimenez-Berni, J.A., George-Jaeggli, B., Potgieter, A.B., Deery, D.M. (2019). Field crop phenomics: enabling breeding for radiation use efficiency and biomass in cereal crops. New Phytol 223. https://doi.org/10.1111/nph.15817

Germeier, C.U. and Unger, S. (2019). Modeling Crop Genetic Resources Phenotyping Information Systems. Front. Plant Sci. Volume 10. https://doi.org/10.3389/fpls.2019.00728

Gepts, P. (2006). Plant Genetic Resources Conservation and Utilization: The Accomplishments and Future of a Societal Insurance Policy. Crop Sci. 46, 2278–2292

Ghaffar, M., Schüler, D., König, P., Arend, D., Junker, A., Scholz, U., Lange, M. (2019). Programmatic Access to FAIRified Digital Plant Genetic Resources. Journal of Integrative Bioinformatics, vol. 16, no. 4, pp. 20190060. https://doi.org/10.1515/jib-2019-0060

Ghazoul, J. (2005). Pollen and seed dispersal among dispersed plants. Biological Reviews, 80(3), 413-443. https://doi.org/10.1017/S1464793105006731.

Gill, T., Gill, S.K., Saini, D.K., Chopra, Y., de Koff, J.P., Sandhu, K.S. (2022). A Comprehensive Review of High Throughput Phenotyping and Machine Learning for Plant Stress Phenotyping. Phenomics. 2(3):156-183. doi: 10.1007/s43657-022-00048-z. PMID: 36939773; PMCID: PMC9590503.

Giller, K.E., Delaune, T., Silva, J.V. *et al.,* (2021). The future of farming: Who will produce our food?. Food Sec. 13, 1073–1099. https://doi.org/10.1007/s12571-021-01184-6

Goldringer, I., Prouin, C., Rousset, M., Galic, N., Bonnin, I., (2006). Rapid differentiation of experimental populations of wheat for heading time in response to local climatic conditions. Annals of Botany, 98(4), 805–817. https://doi.org/10.1093/aob/mcl160

Gollin, D. (2020) Conserving genetic resources for agriculture: economic implications of emerging science. Food Sec. 12, 919–927. https://doi.org/10.1007/s12571-020-01035-w

Goodwin, S., McPherson, J.D., McCombie, W.R. (2016) Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. 17(6):333-51. doi: 10.1038/nrg.2016.49. PMID: 27184599; PMCID: PMC10373632.

Gullotta, G., Engels, J.M.M., Halewood, M. (2023) What Plant Genetic Resources for Food and Agriculture Are Available under the Plant Treaty and Where Is This Information? Plants (Basel). 12(23):3944. https://doi.org/10.3390/plants12233944.

Halewood, M., Lopez Noriega, I., Ellis, D., Roa, C., Rouard, M., Sackville Hamilton, R. (2018a). Using Genomic Sequence Information to Increase Conservation and Sustainable Use of Crop Diversity and Benefit-Sharing. Biopreserv Biobank. (5):368-376. doi: 10.1089/bio.2018.0043. PMID: 30325667; PMCID: PMC6204560.

Halewood, M., Chiurugwi, T., Sackville Hamilton, R., *et al.,* (2018). Plant genetic resources for food and agriculture: opportunities and challenges emerging from the science and information technology revolution. New Phytol. 217(4):1407-1419. doi: 10.1111/nph.14993. Epub 2018 Jan 23. PMID: 29359808.

Hammer, K., Arrowsmith, N., Gladis, T. (2003). Agrobiodiversity with emphasis on plant genetic resources. Naturwissenschaften. 2003 Jun;90(6):241-50. doi: 10.1007/s00114-003-0433-4. PMID: 12835833.

Hanson J, Lusty C, Furman B, Ellis D, Payne T, Halewood M (2024). Opportunities for strategic decision making in managing *ex situ* germplasm collections. Plant Genetic Resources: Characterization and Utilization 1–6. https://doi.org/10.1017/S1479262123000357

Haug K, Salek RM, Conesa P, Hastings J, de Matos P, Rijnbeek M, Mahendraker T, Williams M, Neumann S, Rocca-Serra P, Maguire E, González-Beltrán A, Sansone SA, Griffin JL, Steinbeck C. (2012) MetaboLights--an open-access general-purpose repository for metabolomics studies and associated meta-data. Nucleic Acids Res. 2013 Jan;41(Database issue):D781-6. doi: 10.1093/nar/gks1004. PMID: 23109552; PMCID: PMC3531110.

Harrow, J., Hancock, J., ELIXIR-EXCELERATE Community, Blomberg, N. (2021) ELIXIR-EXCELERATE: establishing Europe's data infrastructure for the life science research of the future. EMBO J. 15;40(6):e107409. https://doi.org/10.15252/embj.2020107409

Hatfield, J.L. and Walthall, C.L. (2015), Meeting Global Food Needs: Realizing the Potential via Genetics × Environment × Management Interactions. Agronomy Journal, 107: 1215-1226. https://doi.org/10.2134/agronj15.0076

Heacock, M.L., Lopez, A.R., Amolegbe, S.M., Carlin, D.J., Henry, H.F., Trottier, B.A., Velasco, M.L., and Suk, W.A. (2022) Enhancing Data Integration, Interoperability, and Reuse to Address Complex and Emerging Environmental Health Problems. Environ. Sci. Technol. 2022, 56, 12, 7544–7552. https://doi.org/10.1021/acs.est.1c08383

Hermjakob, H., & Apweiler, R. (2006). The Proteomics Identifications Database (PRIDE) and the ProteomExchange Consortium: making proteomics data accessible. Expert Review of Proteomics, 3(1), 1–3. https://doi.org/10.1586/14789450.3.1.1

Hinterberger V, Douchkov D, Luck S, Kale S, Mascher M, Stein N, *et al.,* (2022) Mining for new sources of resistance to powdery mildew in genetic resources of winter wheat. Front Plant Sci. https://doi.org/10.3389/fpls.2022.836723.

Hu T., Chitnis N., Monos D. *et al.* (2021) Next-generation sequencing technologies: an overview. Hum. Immunol., 82, 01–811.

IPGRI (1993). Diversity for development. Rome: International Plant Genetic Resources Institute.

Kearns, C. A., & Inouye, D. W. (1997). Pollinators, flowering plants, and conservation biology. BioScience, 47(5), 297-307. https://doi.org/10.2307/1313191.

Joly A., Goëau H., Bonnet P., Bakić, V., Barbe, J. *et al.,* (2014) Interactive plant identification based on social image data. Ecological Informatics 23, pp.22-34. https://doi.org/10.1016/j.ecoinf.2013.07.006.

Kell, S., Marino, M., Maxted, N. (2017) Bottlenecks in the PGR use system: stakeholders' perspectives. Euphytica 213, 170. https://doi.org/10.1007/s10681-017-1935-z

Kell, S.P., Knüpffer, H., Jury, S.L., Maxted, N., Ford-Lloyd, B.V. (2005) *Catalogue of Crop Wild Relatives for Europe and the Mediterranean*. University of Birmingham, United Kingdom, available online via the Crop Wild Relative Information System (CWRIS – http://www.pgrforum.org/cwris/cwris.asp) and on CD-ROM.

Khoury, C.K., Laliberte, B., Guarino, L. (2010). Trends in ex situ conservation of plant genetic resources: A review of global crop and regional conservation strategies. Genet. Resour. Crop. Evol. 57, 625–639.

Kodama, Y., Shumway, M., Leinonen, R.. (2012). International Nucleotide Sequence Database Collaboration. The Sequence Read Archive: explosive growth of sequencing data. Nucleic Acids Res. 40(Database issue):D54-6. doi: 10.1093/nar/gkr854. Epub 2011 Oct 18. PMID: 22009675; PMCID: PMC3245110.

Kopittke, P.M., Menzeis, N.W., Wang, P., McKenna B.A., Lombi, E. (2019) Soil and the intensification of agriculture for global food security. Environment International 132: 105078 https://doi.org/10.1016/j.envint.2019.105078

Kotni, P., van Hintum, T., Maggioni, L., Oppermann, M., and Weise S. (2022). EURISCO update 2023: the European Search Catalogue for Plant Genetic Resources, a pillar for documentation of genebank material. Nucleic Acids Research, 51(D1):D1465-D1469, 2023.

Kreide, S., Oppermann, M., and Weise S. (2019). Advancement of taxonomic searches in the European search catalogue for plant genetic resources. Plant Genetic Resources, 17(6):559-561.

Laird, P. (2010). Principles and challenges of genome-wide DNA methylation analysis. Nat Rev Genet 11, 191–203 https://doi.org/10.1038/nrg2732

Lassoued, R., Macall, D.M., Smyth, S.J, Phillips, P.W.B., Hesseln, H. (2021) Data challenges for future plant gene editing: expert opinion. Transgenic Res. 30(6):765-780. https://doi.org/:10.1007/s11248-021-00264-9

Leal, A.C. (2024). Data Analytics in Agriculture. In: Priyadarshan, P.M., Jain, S.M., Penna, S., Al-Khayri, J.M. (eds) Digital Agriculture. Springer, Cham. https://doi.org/10.1007/978-3-031-43548-5_17

Lewin, H.A., Robinson, G.E., Kress, W.J., *et al.,* (2018) Earth BioGenome Project: Sequencing Life for the Future of Life. Proc Natl Acad Sci U S A. 115(17):4325–4333. 10.1073/pnas.1720115115

Li L., Mao X.G., Wang J.Y., Chang X.P., Reynolds M., Jing R.L. (2019) Genetic dissection of drought and heat-responsive agronomic traits in wheat. Plant Cell Environ. 42:2540–2553. https://doi.org/10.1111/pce.13577.

Li, Y. and He, N. (2024), Innovations and prospectives of multidimensional trait integration. New Phytol. https://doi.org/10.1111/nph.19909

Lim, P. K., Zheng, X., Goh, J. C. & Mutwil, M. (2022). Exploiting plant tran-scriptomic databases: Resources, tools, and approaches. Plant Commun, 3, 100323.

Liu, Y., Lu, S., Liu, K. *et al.* (2019). Proteomics: a powerful tool to study plant responses to biotic stress. Plant Methods 15, 135. https://doi.org/10.1186/s13007-019-0515-8

Luo, H., Zhang, H., & Wang, H. (2023). Advance in sex differentiation in cucumber. Frontiers in Plant Science, 14. https://doi.org/10.3389/fpls.2023.1186904

Lusty C, Sackville Hamilton R, Guarino L, Richards C, Jamora N, Hawtin G. (2021). Envisaging an Effective Global Long-Term Agrobiodiversity Conservation System That Promotes and Facilitates Use. Plants. 10(12):2764. https://doi.org/10.3390/plants10122764

Magos Brehm, J., Kell, S.P., Thormann, I., Gaisberger, H., Dulloo, E., Maxted, N., (2017). Occurrence data collation template v.1, doi:10.7910/DVN/5B9IV5, Harvard Dataverse, V1. Available here: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/5B9IV5 (accessed 04.01.19).

Maltsev, Y., Erst, A. (2023) Recent Advances in the Integrative Taxonomy of Plants. Plants (Basel). 12(24):4097. https://doi.org/10.3390/plants12244097.

Manickam S, Rajagopalan VR, Kambale R, Rajasekaran R, Kanagarajan S, Muthurajan R. (2023). Plant Metabolomics: Current Initiatives and Future Prospects. Curr Issues Mol Biol. 45(11):8894-8906. doi: 10.3390/cimb45110558. PMID: 37998735; PMCID: PMC10670879.

Manzella, D., Marsella, M., Jaiswal, P., Arnaud, E., King, B. (2023) Digital Sequence Information and Plant Genetic Resources: Global Policy Meets Interoperability. In: Williamson, H.F., Leonelli, S. (eds) Towards Responsible Plant Data Linkage: Data Challenges for Agricultural Research and Development. Springer, Cham. https://doi.org/10.1007/978-3-031-13276-6_10.

Marden, E,, Sackville Hamilton, R., Halewood, M., McCouch, S. (2023) International agreements and the plant genetics research community: A guide to practice. Proc Natl Acad Sci U S A. 120(14):e2205773119. https://doi.org/10.1073/pnas.2205773119.

Maxted N., Amri. A., Castañeda-Álvarez, N.P., Dias, S., Dulloo, M.E., Fielder, H., Ford-Lloyd, B.V., Iriondo, J.M., Magos Brehm, J., Nilsen, L-B., Thormann, I., Vincent, H. and Kell, S.P., (2016). Joining up the dots: a systematic perspective of crop wild relative conservation and use. In: Maxted, N., Ehsan Dulloo, M. and Ford-Lloyd, B.V. (eds.), Enhancing Crop Genepool Use: Capturing Wild Relative and Landrace Diversity for Crop Improvement. Pp. 87-124. CAB International, Wallingford, UK.

Maxted, N., Castañeda Álvarez, N.P., Vincent, H.A., Magos Brehm, J., (2012). *Gap analysis: a tool for genetic conservation*. In Guarino L, Ramanatha Rao V, Goldberg E (editors). Collecting Plant Genetic Diversity: Technical Guidelines. 2011 update. Bioversity International, Rome. Available online: http://cropgenebank.sgrp.cgiar.org/index.php?option=com_content&view=article&id=678

Maxted, N., Ford-Lloyd, B.V., Hawkes, J.G. (1997). Complementary Conservation Strategies. In: Plant genetic conservation: the *in situ* approach, Maxted N, Ford-Lloyd BV, Hawkes JG (eds.) pp. 20–55. Chapman & Hall, London. https://doi.org/10.1007/978-94-009-1437-7_2

Maxted, N., Ford-Lloyd, B.V., Jury, S. *et al.* (2006) Towards a definition of a crop wild relative. Biodivers Conserv 15, 2673–2685. https://doi.org/10.1007/s10531-005-5409-6

Maxted, N., Guarino, L., Myer, L., Chiwona, E.A. (2002). Towards a methodology for on-farm conservation of plant genetic resources. Genetic Resources and Crop Evolution 49: 31–46. https://doi.org/10.1023/A:1013896401710

Maxted, N., Amri, A., Castañeda-Álvarez, N.P. *et al.,* (2016). Joining up the dots: a systematic perspective of crop wild relative conservation and use. In: Maxted, N. (et al. (eds.)) Enhancing crop genepool use: capturing wild relative and landrace diversity for crop improvement. Oxfordshire (UK): CABI, p.87-124 ISBN: 978-1-78064-613-8

Maxted, N., Hunter, D., Ortiz Rios, R.O., (2020). Plant genetic conservation. 560 pp. Cambridge University Press, Cambridge.

Maxted, N., Iriondo, J., De Hond, L., Dulloo, E., Lefèvre, F., Asdal, A. Kell, S.P., Guarino, L. (2008). Genetic Reserve Management. In: Iriondo, J.M., Maxted, N., Dulloo, E. (Eds.), *Plant Genetic Population Management*. Pp. 65-87. CAB International, Wallingford.

Maxted, N., Magos Brehm, J., (2023). Maximizing the crop wild relative resources available to plant breeders for crop improvement. Frontiers in Sustainable Food Systems, 7: doi.org/10.3389/fsufs.2023.1010204.

Maxted, N., Phillips, J., Magos Brehm, J., *et al.*, (2024) Systems for describing, managing and accessing *in situ* conserved populations an interfacing them with EURISCO. Deliverable 1.3 EC funded Pro GRACE Project. Agenzia Nazionale per le Nuove Tecnologie, l'Energia e lo Sviluppo Economico Sostenibile, Roma, Italy. Pp. 1-99.

Maxted, N., van Slageren, M.W., Rihan, J., (1995). Ecogeographic surveys. In: Guarino, L., Ramanatha Rao, V. and Reid, R. (eds.), *Collecting plant genetic diversity: technical guidelines*. Pp. 255-286. CAB International, Wallingford.

Ming, R., Bendahmane, A., & Renner, S. S. (2011). Sex chromosomes in land plants. Annual Review of Plant Biology, 62, 485-514. https://doi.org/10.1146/annurev-arplant-042110-103914.

Morris, R.A., Barve, V., Carausu, M. (2013) Discovery and Publishing of Primary Biodiversity Data Associated With Multimedia Resources: The Audubon Core Strategies and Approaches. Biodiversity Informatics 8 (2). https://doi.org/10.17161/bi.v8i2.4117

Neveu, P., Tireau, A., Hilgert, N., Nègre, V., Mineau-Cesari, J., Brichet, N., Chapuis, R., Sanchez, I., Pommier, C., Charnomordic, B., Tardieu, F. and Cabrera-Bosquet, L. (2019), Dealing with multi-source and multi-scale information in plant phenomics: the ontology-driven Phenotyping Hybrid Information System. New Phytol, 221: 588-601. https://doi.org/10.1111/nph.15385

Niedbała, G., Piekutowska, M., Hara, P. 2023. New Trends and Challenges in Precision and Digital Agriculture. Agronomy 13 (8), 2136. https://doi.org/10.3390/agronomy13082136

Ninomiya S., Baret F., Cheng Z.-M. (2019). Plant phenomics: emerging transdisciplinary science. Plant Phenomics. https://doi.org/10.34133/2019/2765120

OECD (2007). OECD Best Practice Guidelines for Biological Resource Centres, OECD Publishing, Paris, https://doi.org/10.1787/9789264128767-en.

Oppermann, M., Weise, S., Dittmann, C., Knüpffer, H. (2015) GBIS: the information system of the German Genebank, Database, Volume 2015, bav021, https://doi.org/10.1093/database/bav021

Papoutsoglou, E.A., Faria, D., Arend, D., Arnaud, E., *et al.,* (2020) Enabling reusability of plant phenomic datasets with MIAPPE 1.1. New Phytol, 227: 260-273. https://doi.org/10.1111/nph.16544

Parkinson H, Kapushesky M, Shojatalab M, Abeygunawardena N, Coulson R, Farne A, Holloway E, *et al.,* 2007. ArrayExpress--a public database of microarray experiments and gene expression profiles. Nucleic Acids Res. 2007 Jan;35(Database issue):D747-50. doi: 10.1093/nar/gkl995. Epub 2006 Nov 28. PMID: 17132828; PMCID: PMC1716725.

Paskin, N. (2005). Digital Object Identifiers for Scientific Data. Data Science Journal 4(18):12-20

Pasquetto, I.V., Randles, B.M., Borgman, C.L., (2017). On the Reuse of Scientific Data. Data Sci. J., 16 1-9. https://doi.org/10.5334/dsj-2017-008.

Pathirana, R., Carimi F. (2022) Management and Utilization of Plant Genetic Resources for a Sustainable Agriculture. Plants. 11(15):2038. https://doi.org/10.3390/plants11152038

Pieruschka R., Schurr U. (2019). Plant phenotyping: past, present, and future. Plant Phenomics 2019, 6. https://doi.org/10.1155/2019/7507131

Pommier, C. Coppens, F., Ćwiek-Kupczyńska; H. *et al.,* (2023). Plant Science Data Integration, from Building Community Standards to Defining a Consistent Data Lifecycle. In: Williamson, H.F., Leonelli, S. (eds) Towards Responsible Plant Data Linkage: Data Challenges for Agricultural Research and Development. Springer, Cham. Pp. 149-160. https://doi.org/10.1007/978-3-031-13276-6_8

Pohl, A., Beato, M. (2014). bwtool: a tool for bigWig files. Bioinformatics. 2014 Jun 1;30(11):1618-9. doi: 10.1093/bioinformatics/btu056. PMID: 24489365; PMCID: PMC4029031.

Postman, J.D., Bretting, P.K., Kinard, G.R., Cyr, P.D., Weaver, B., Millard, M.J., and Gardner, C.A., *et al.,* 2010. GRIN-Global: An international project to develop a global plant genebank information management system. Acta Hortic. 859: 49–55. https://doi.org/10.17660/ActaHortic.2010.859.4

Pritchard, L., Brown, C.T., Harrington, B., Heath, L.S, Pierce-Ward, N.T., Vinatzer, B.A. (2022) Could a Focus on the "Why" of Taxonomy Help Taxonomy Better Respond to the Needs of Science and Society? Front Microbiol. 13:887310. https://doi.org/10.3389/fmicb.2022.887310.

Quinlan, A.R., Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 26(6):841-2. doi: 10.1093/bioinformatics/btq033. Epub 2010 Jan 28. PMID: 20110278; PMCID: PMC2832824.

Rabieyan, E., Darvishzadeh, R., Mohammadi, R., Gul, A., Rasheed, A., Akhar, F.K., Abdi, H., Alipour, H. (2023). Genetic diversity, linkage disequilibrium, and population structure of tetraploid wheat landraces originating from Europe and Asia. BMC Genomics. 24(1):682. https://doi.org/10.1186/s12864-023-09768-6

Ramanatha Rao, V., Hodgkin, T. (2002). Genetic diversity and conservation and utilization of plant genetic resources. Plant Cell, Tissue and Organ Culture 68, 1–19. https://doi.org/10.1023/A:1013359015812

Raubach, S., Kilian, B., Dreher, K., *et al.,* (2021) From bits to bites: Advancement of the Germinate platform to support genetic resources collections and prebreeding informatics for crop wild relatives. Crop Science. 61: 1538–1566. https://doi.org/10.1002/csc2.20248

Rife, T., Poland J. 2014. Field Book: An Open-Source Application for Field Data Collection on Android. Crop Breeding & Genetics 54(4): 1624-1627

Robertson, T, Döring, M, Guralnick, R, Bloom, D, Wieczorek, J, Braak, K, Otegui, J, Russell, L and Desmet, P. (2014). The GBIF integrated publishing toolkit: facilitating the efficient publishing of

biodiversity data on the internet. PloS one, 9(8): e102623. DOI: https://doi.org/10.1371/journal.pone.0102623

Scossa, F., Alseekh, S., Fernie, A.R. (2021) Integrating multi-omics data for crop improvement. J Plant Physiol. Epub 2020 Dec 17. PMID: 33360148. https://doi.org/10.1016/j.jplph.2020.153352

Shukla, B.K., Maurya, N., Sharma, M. 2023. Advancements in Sensor-Based Technologies for Precision Agriculture: An Exploration of Interoperability, Analytics and Deployment Strategies. Eng. Proc. 58, 22. https://doi.org/10.3390/ecsa-10-16051

Smith, D.T., Potgieter, A.B., Chapman, S.C. (2021) Scaling up high-throughput phenotyping for abiotic stress selection in the field. Theor Appl Genet. 2021 Jun;134(6):1845-1866. doi: 10.1007/s00122-021-03864-5. Epub 2021 Jun 2. PMID: 34076731.

Song, P., Wang, J., Guo, X., Yang, W., Zhao, C. (2021). High-throughput phenotyping: breaking through the bottleneck in future crop breeding. Crop J. 9, 633–645. https://doi.org/10.1016/j.cj.2021.03.015

Steinbach D, Alaux M, Amselem J, Choisne N, Durand S, Flores R, Keliet AO, *et al.,* (2013) GnpIS: an information system to integrate genetic and genomic data from plants and fungi. Database (Oxford). https://doi.org/10.1093/database/bat058.

Sud M, Fahy E, Cotter D, *et al*., (2016) Metabolomics Workbench: An international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools, Nucleic Acids Research 44 (D1): 4 D463–D470, https://doi.org/10.1093/nar/gkv1042

Sultan, S.E. (2000) Phenotypic plasticity for plant development, function and life history Trends Plant Sci., 5: 537-542, https://doi.org/10.1016/s1360-1385(00)01797-0

Sumner LW, Amberg A, Barrett D, Beale MH, Beger R, Daykin CA, Fan TW, *et al.,* (2007) Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). Metabolomics. 3(3):211-221. https://doi.org/10.1007/s11306-007-0082-2.

Sun, Q., Zybailov, B., Majeran, W. *et al.,* 2009. PPDB, the Plant Proteomics Database at Cornell, Nucleic Acids Research, Volume 37, Issue suppl_1, 1: D969–D974, https://doi.org/10.1093/nar/gkn654

Tao, H., Xu, S., Tian, Y., Li, Z., Ge, Y., *et al.*, (2022). Proximal and remote sensing in plant phenomics: 20 years of progress, challenges, and perspectives. Plant Commun. 3, 100344 https://doi.org/10.1016/j.xplc.2022.100344.

Tardieu, F., Cabrera-Bosquet, L., Pridmore, T., Bennett, M., (2017). Plant phenomics, from sensors to knowledge. Curr. Biol. 27, R770–R783. https://doi.org/10.1016/j.cub.2017.05.055

Taylor, C., Paton, N., Lilley, K. *et al.,* (2007) The minimum information about a proteomics experiment (MIAPE). Nat Biotechnol 25, 887–893. https://doi.org/10.1038/nbt1329

Teixidor-Toneu I, Giraud NJ, Karlsen P, Annes A and Kool A (2023) A transdisciplinary approach to define and assess wild food plant sustainable foraging in Norway. Plants People Planet 5(1): 112–122.

Thiele, K.R., Conix, S., Pyle, R.L. *et al.,* (2021) Towards a global list of accepted species I. Why taxonomists sometimes disagree, and why this matters. Org Divers Evol 21, 615–622 https://doi.org/10.1007/s13127-021-00495-y

Thormann, I., Kell, S.P., Magos Brehm, J., Dulloo, M.E., Maxted, N., (2017). Checklist+Inventory_template_30122017.xlsm, CWR checklist and inventory data template v.1, https://doi.org/10.7910/DVN/B8YOQL/T4JOOV, Harvard Dataverse, V4.

Thormann, I., Parra-Quijano, M., Endresen, D.T.F., Rubio-Teso, M.L., Iriondo, M.J., Maxted, N., (2014). *Predictive characterization of crop wild relatives and landraces. Technical guidelines version 1*. Bioversity International, Rome, Italy. Available online at: http://www.bioversityinternational.org/index.php?id=244&tx_news_pi1%5Bnews%5D=4967&cHash=7cd3c6c2b8360927b83fa6ef7cc28d99 (Accessed 17.07.17).

Tirnaz S, Zandberg J, Thomas WJW, Marsh J, Edwards D, Batley J. Application of crop wild relatives in modern breeding: An overview of resources, experimental and computational methodologies. Front Plant Sci. 2022 Nov 17;13:1008904. doi: 10.3389/fpls.2022.1008904. PMID: 36466237; PMCID: PMC9712971.

Ulian, T., Diazgranados, M., Pironon, S., Padulosi, S., *et al.,* (2020). Unlocking plant resources to support food security and promote sustainable agriculture. Plants, People, Planet, 2(5), 421–445. https://doi.org/10.1002/ppp3.10145

van Etten, J., de Sousa, K., Cairns, J.E., Dell'Acqua, M., *et al.,* (2023) Data-driven approaches can harness crop diversity to address heterogeneous needs for breeding products. PNAS 120(14): e2205771120. 10 p. ISSN: 0027-8424. https://doi.org/10.1073/pnas.2205771120

Volk GM, Byrne PF, Coyne CJ, Flint-Garcia S, Reeves PA, *et al.* (2021). Integrating genomic and phenomic approaches to support plant genetic resources conservation and use. Plants 10:2260. https://doi.og/10.3390/plants10112260

Wafula EK, Zhang H, Von Kuster G, Leebens-Mack JH, Honaas LA and dePamphilis CW (2023) PlantTribes2: Tools for comparative gene family analysis in plant genomics. Front. Plant Sci. 13:1011199. https://doi.org/10.3389/fpls.2022.1011199

Wang J, Li C, Li L, Reynolds M, Mao X, Jing R. (2021) Exploitation of Drought Tolerance-Related Genes for Crop Improvement. Int J Mol Sci. https://doi.org/10.3390/ijms221910265.

Wang L, Xu J, Wang H, Chen T, You E, Bian H, Chen W, Zhang B and Shen Y (2023) Population structure analysis and genome-wide association study of a hexaploid oat landrace and cultivar collection. Front. Plant Sci. 14:1131751. https://doi.org/10.3389/fpls.2023.1131751

Watt, M., Fiorani, F., Usadel, B., Muller, O., Schurr, U., (2020). Phenotyping: new windows into the plant for breeders. Annu. Rev. Plant Biol. 71, 1–24. https://doi.org/10.1146/annurev-arplant-042916-041124

Weise, S., Oppermann, M., Maggioni, L., van Hintum T., and Knüpffer H. (2017). EURISCO: The European Search Catalogue for Plant Genetic Resources. Nucleic Acids Research, 45(D1):D1003-1008.

Weise, S., Lohwasser, U., Oppermann, M. (2020) Document or Lose It-On the Importance of Information Management for Genetic Resources Conservation in Genebanks. Plants (Basel). 18;9(8):1050. https://doi.org/10.3390/plants9081050

Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, *et al.,* (2012) Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. PLoS ONE 7(1): e29715. https://doi.org/10.1371/journal.pone.0029715

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

Yandell, M., Ence, D. (2012) A beginner's guide to eukaryotic genome annotation. Nat Rev Genet. 13(5):329-42. doi: 10.1038/nrg3174. PMID: 22510764.

Yang W., Feng H., Zhang X. *et al.,* (2020) Crop phenomics and high-throughput phenotyping: past decades, current challenges, and future perspectives. Mol. Plant, 13, 187–214.

Yilmaz, P., Kottmann, R., Field, D., *et al.,* (2011). Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. Nature Biotechnology 29, 415–420. https://doi.org/10.1038/nbt.1823

Zhang, H., Wang, L., Jin, X., Bian, L., Ge, Y., (2023). High-throughput phenotyping of plant leaf morphological, physiological, and biochemical traits on multiple scales using optical sensing. Crop J. https://doi.org/10.1016/j.cj.2023.04.014.

Zhang, Y., Zhang, N., (2018). Imaging technologies for plant high-throughput phenotyping: a review. Front. Agric. Sci. Eng. 5, 406–419. https://doi.org/10.15302/J-FASE-2018242