C.H. Aguilar, M. Oppermann, S. Weise

Impulse presentation

Phenotypic data, FAIR principles, trusted repository

Workshop of the Documentation & Information Working Group, 18–19 September 2024, Tallinn, Estonia

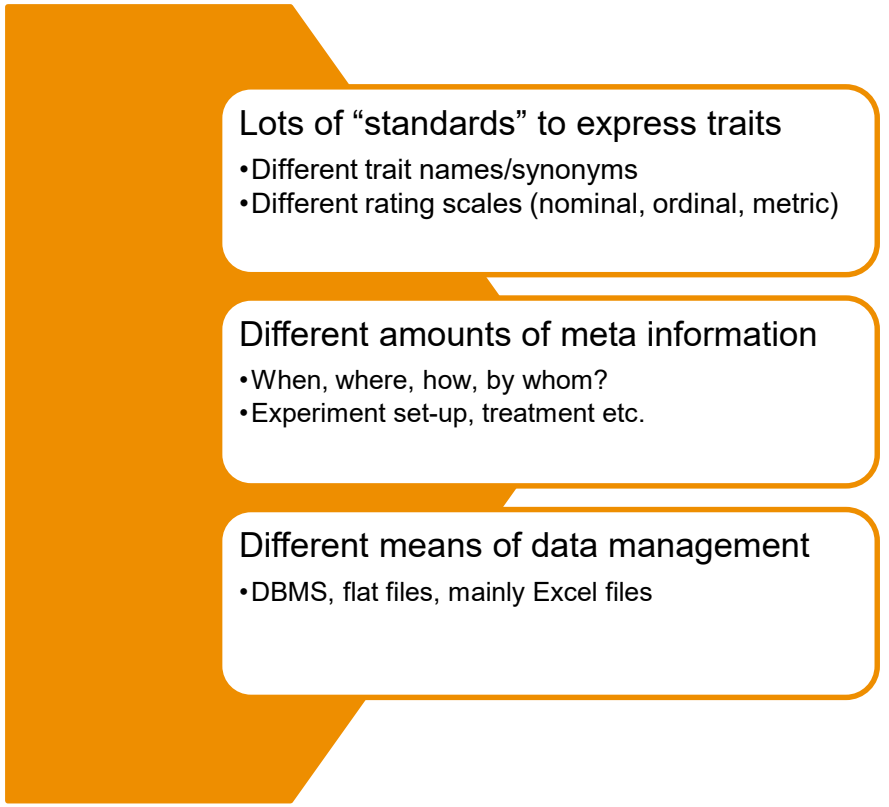
Three items to be discussed

- How to improve the availability of phenotypic data in EURISCO?
- How to increase the compliance of data with FAIR principles?
- How to develop EURISCO towards a trustable repository with acceptably high governance and data management standards?

How to improve the availability of phenotypic data in EURISCO?

Obstacles with phenotypic data

- Important: Determines value of germplasm for breeding and research
- Crop-specific traits and methods
- Many historical datasets
- Usually no data from high throughput phenotyping
- Data has to be aggregated or exchanged between organisations



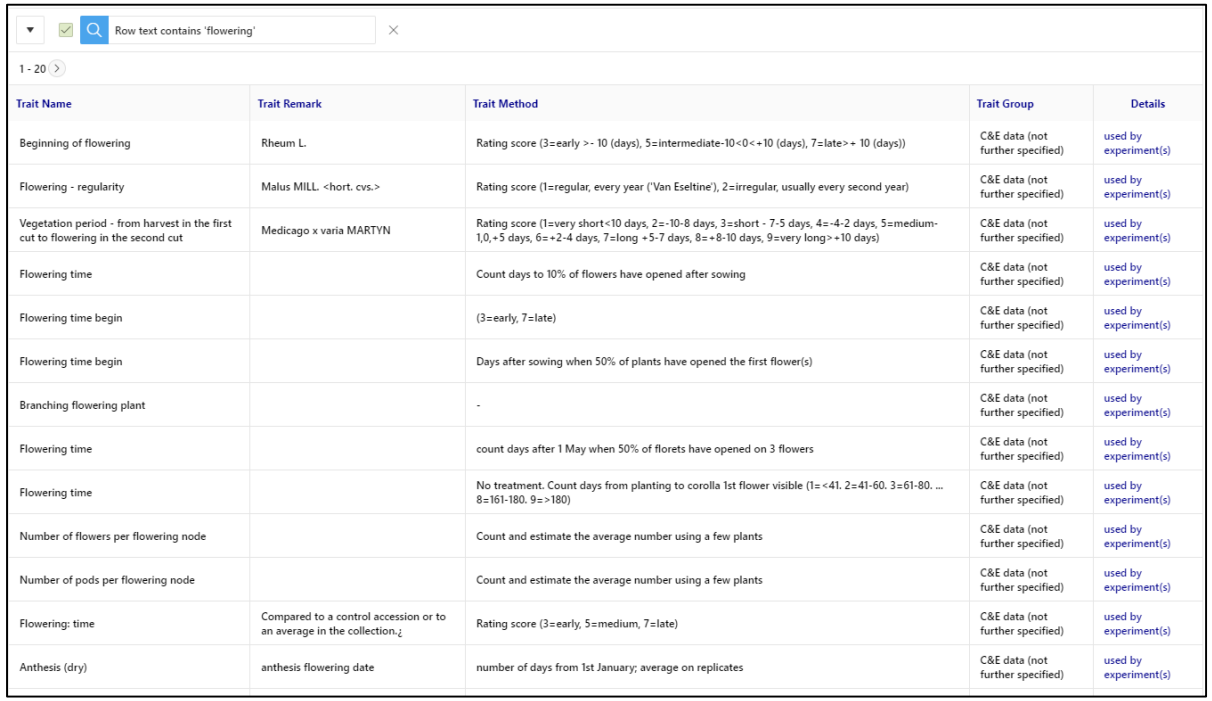
Current approach

- Data standardisation
  - About 600 germplasm collections in Europe, around 400 in EURISCO
  - No standardisation of trait, scale or experimental design
  - Pragmatic approach: Import of existing data as-is to reach critical mass
- Data exchange
  - Only standardisation of exchange format
    - As simple as possible
    - As few fields as possible

  → “minimum consensus”
- Data management
  - Highly abstracted, following the single-observation concept (van Hintum et al. 1992)
  - Omitting fine-grained metadata

Phenotypic data in EURISCO

- Extension available since 2016
- 2,729,636 records of data
- 21 countries
- 74 phenotypic datasets
- 3,919 experiments
- 9,764 traits
- 91,443 accs. with phenotypic data



Row text contains 'flowering'

1 - 20

Trait Name	Trait Remark	Trait Method	Trait Group	Details
Beginning of flowering	Rheum L.	Rating score (3=early >- 10 (days), 5=intermediate-10<0<+10 (days), 7=late>+ 10 (days))	C&E data (not further specified)	used by experiment(s)
Flowering - regularity	Malus MILL. <hort. cvs.>	Rating score (1=regular, every year ('Van Eseltine'), 2=irregular, usually every second year)	C&E data (not further specified)	used by experiment(s)
Vegetation period - from harvest in the first cut to flowering in the second cut	Medicago x varia MARTYN	Rating score (1=very short<10 days, 2=-10-8 days, 3=short - 7-5 days, 4=-4-2 days, 5=medium-1,0,+5 days, 6=+2-4 days, 7=long +5-7 days, 8=+8-10 days, 9=very long>+10 days)	C&E data (not further specified)	used by experiment(s)
Flowering time		Count days to 10% of flowers have opened after sowing	C&E data (not further specified)	used by experiment(s)
Flowering time begin		(3=early, 7=late)	C&E data (not further specified)	used by experiment(s)
Flowering time begin		Days after sowing when 50% of plants have opened the first flower(s)	C&E data (not further specified)	used by experiment(s)
Branching flowering plant		-	C&E data (not further specified)	used by experiment(s)
Flowering time		count days after 1 May when 50% of florets have opened on 3 flowers	C&E data (not further specified)	used by experiment(s)
Flowering time		No treatment. Count days from planting to corolla 1st flower visible (1=<41. 2=41-60. 3=61-80. ... 8=161-180. 9=>180)	C&E data (not further specified)	used by experiment(s)
Number of flowers per flowering node		Count and estimate the average number using a few plants	C&E data (not further specified)	used by experiment(s)
Number of pods per flowering node		Count and estimate the average number using a few plants	C&E data (not further specified)	used by experiment(s)
Flowering: time	Compared to a control accession or to an average in the collection.¿	Rating score (3=early, 5=medium, 7=late)	C&E data (not further specified)	used by experiment(s)
Anthesis (dry)	anthesis flowering date	number of days from 1st January; average on replicates	C&E data (not further specified)	used by experiment(s)

- Limitations
  - EURISCO data exchange format represents a “minimum consensus”
  - Difficult to compile files manually
  - Very limited reproducibility and comparability

To be discussed

- Simplification of data collection → one column per trait to support manual recording

- Additional metadata
  - Experiment
  - Trait
  - Range of values

Experiment

- **EXPERIMENT_ID**: Unique numeric value necessary for uploading the data *(mandatory)*.
- **NAME**: Brief name of the experiment *(mandatory)*.
- COUNTRY_CODE: ISO3 code of the country in which the experiment took place *(3 alphanumeric characters)*
- SITE: Name of the location where the experiment took place *(max. 100 alphanumeric characters)*
- DESCRIPTION: Brief description of the experiment. Information relevant for the interpretation of the scores in the experiment *(max. 2000 alphanumeric characters)*.
- **YEAR_START**: The year the experiment was performed (started) *(4 numeric characters; mandatory)*.
- YEAR_END: The year in which the experiment ended *(4 numeric characters)*.
- LONGITUDE: The longitude of the experimental site, provided it was an experiment in the open field *(decimal format)*.
- LATITUDE: The latitude of the experimental site, provided it was an experiment in the open field *(decimal format)*.
- REMARKS: Any general remark that helps to interpret the experiment *(max. 2000 alphanumeric characters)*.

EXPERIMENT_ID	NAME	COUNTRY_CODE	SITE	DESCRIPTION	YEAR_START	YEAR_END	LONGITUDE	LATITUDE	REMARKS
1	Drought stress trial	DEU	Gatersleben	…	1982	1983	11.278414	51.826059	…
2	Multiplication trial	DEU	Gatersleben	…	1990	1991	11.278414	51.826059	…

Trait

- **TRAIT_ID**: Unique number of the trait, necessary for uploading *(mandatory)*.
- **NAME**: Name of the trait *(max. 100 alphanumeric characters; mandatory)*.
- DESCRIPTION: A description of the method for measuring *(max. 2000 alphanumeric characters)*.
- UNIT: The unit used for measuring the trait value *(max. 100 alphanumeric characters) (mandatory if applicable).*
- **TYPE**: The type of the trait, with type in {Date, Measurement, Rating score} *(mandatory).*
- CO_TERM: Crop Ontology term to enable subsequent harmonisation of traits *(max. 50 alphanumeric characters)*
- REMARKS: Any general remark that helps to interpret the trait *(max. 2000 alphanumeric characters)*.

TRAIT_ID	NAME	DESCRIPTION	UNIT	TYPE	CO_TERM	REMARKS
222	Date of flowering	Date of flowering when 50% plants in a plot have started flowering stage		Date	CO_323:0000012	…
333	Grain yield	Whole above ground biomass dry matter basis yield	kg/m²	Measurement	CO_323:0000229	..
444	Lodging	Lodging incidence per plot		Rating score	CO_323:0000021	…
555	Plant height	Height from the ground level to the top part	cm	Measurement	CO_323:0000024	…

Range of value

- **TRAIT_ID**: Unique number of the trait as defined in the TRAIT template *(mandatory)*.
- **RATING_VALUE**: Allowed rating value *(number or max. 100 alphanumeric characters; mandatory)*.
- **DESCRIPTION**: Meaning of the rating value *(max. 100 alphanumeric characters; mandatory)*.
- REMARKS: Any general remark that helps to interpret the value *(max. 2000 alphanumeric characters)*.

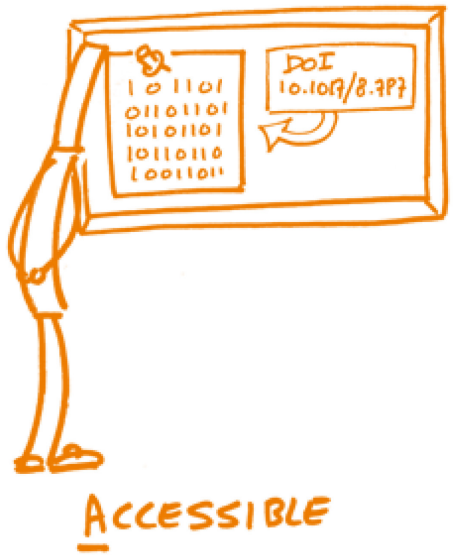
TRAIT_ID	RATING_VALUE	DESCRIPTION	REMARKS
444	1	none	…
444	2	slight	…
444	3	very low	…
444	4	low	…
444	5	intermediate	…
444	6	intermediate to high	…
444	7	high	…
444	8	very high	…
444	9	severe	…

Observed scores

- **Experiment_ID**: Unique number of the experiment as defined in the EXPERIMENT template *(mandatory)*.
- **INSTCODE**: FAO-WIEWS code of the institute maintaining the accession *(max. 100 alphanumeric characters; mandatory)*.
- **GENUS**: Genus name of the accession *(max. 100 alphanumeric characters; mandatory)*.
- **ACCENUMB**: Accession number *(max. 100 alphanumeric characters; mandatory)*.
- DOI: Digital Object Identifier of the accession, if available (*max. 100 alphanumeric characters*).
- Observations: Here follow the actual observed scores. Two columns, the observation value and the observation date.

EXPERIMENT ID	INSTCODE	GENUS	ACCENUMB	DOI	Date of flowering	Observation date	Grain yield	Observation date	Lodging	Observation date	Plant height	Observation date	…
2	DEU146	Hordeum	HOR 1	10.25642/IPK/GBIS/16271	19820503	19820503	14	19820708	2	19820609	146	19800630	…
2	DEU146	Hordeum	HOR 2	10.25642/IPK/GBIS/16272	19820507	19820507	12	19820708	7	19820610	135	19820630	…
…	…	…	…	…	…		…		…				…

How to increase the compliance of data with FAIR principles?



Bezjak et al. 2018

- Unique & persistent identifiers
- Rich metadata
- Clear data-metadata links
- Searchable

- Retrievable using a protocol
- Protocol is open & implemented
- Protocol allows authorisation when needed
- Persistently accessible metadata

- Links to other (meta) data
- Consistent vocabularies
- Formal, accessible, shared and applicable language

- Clear and accessible data usage licence
- Detailed provenance
- Rich descriptions with accurate and relevant attribute
- Meet domain-relevant standards

Wilkinson et al. 2016

F
FINDABLE

A
ACCESSIBLE

I
INTEROPERABLE

R
REUSABLE

IPK
LEIBNIZ INSTITUTE

Limitations

***Ex situ* passport data**	• **Non-standardised collection identifiers** • Inconsistencies in metadata labelling
Phenotypic data	• **Fragmentation across institutions & heterogeneity** • Metadata quality and standardisation • Disparate data repositories
Genotypic data	• Heterogeneity of data types and formats • Data indexing and search tools • Metadata quality and standardisation
Image data	• Lack of standardised metadata and identifiers • Absence of consistent tagging or keyword systems • No digital asset management
…	

Way forward

- Additional unique identifiers
  - FAO recommendation for PUIDs in the form of DOIs
  - Increasingly used, but still great potential for improvement (currently for approx. 20% of accessions)
  - BioSample IDs, mainly for genomic data
  - Linking with DOIs possible
  - Application of DOIs also for existing phenotypic datasets
- Better networks of IDs

F
FINDABLE

A
ACCESSIBLE

I
INTEROPERABLE

R
REUSABLE

IPK
LEIBNIZ INSTITUTE

Limitations

***Ex situ* passport data**	• **Access restrictions** • Technical limitations (e.g. obsolete platforms)
Phenotypic data	• **Different storage systems & technical infrastructure** • Depth and granularity discrepancies • Use restrictions
Genotypic data	• Specialised infrastructure for vast data sets • Advanced analytical tools requirement • **Sustainability of infrastructures**
Image data	• Inefficient or proprietary compression algorithms affecting data retrieval speed and quality • Ethical and regulatory complexities

…

Way forward

- Aggregators
  - Expand existing systems
  - Better networking
- (Further) development of trusted repositories
  - Especially for phenotypic data
  - Consistently submit project data to public repositories
- Stronger cooperation between genebanks

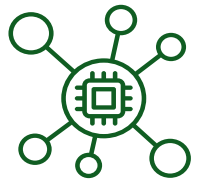
IPK LEIBNIZ INSTITUTE

Limitations

***Ex situ* passport data**	• Inconsistent adoption of MCPD standard • Additional data beyond MCPD • **Semantic interoperability**
Phenotypic data	• Diversity of trait measurements and terminologies • **Format & standard compatibility** • Data management and exchange protocols
Genotypic data	• **Diverse data formats and standards** • Data integration across platforms • Evolving landscape of research
Image data	• Inconsistent resolution, format, scale & annotations • Absence of universally adopted ontologies for PGR image data • Variability in image analysis
…	

Way forward

- Standardisation/harmonisation
  - Terminologies
  - Data standards
- Approaches to semantic standardisation
  - Not yet fully developed and not yet effective
  - Especially for phenotypic data
  - Ontologies, as used for PGR, do not work properly
- Restrictions may have to be accepted here



F
FINDABLE

A
ACCESSIBLE

I
INTEROPERABLE

R
REUSABLE

IPK
LEIBNIZ INSTITUTE

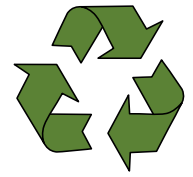
Limitations

***Ex situ* passport data**	• Legacy data issues • Insufficient metadata will reduce their potential for reuse
Phenotypic data	• Genotype x environment x cultural practice • **Comprehensive metadata and documentation** • Data quality & integrity
Genotypic data	• (Meta) data quality and completeness • Annotation and version control
Image data	• **Metadata completeness and standardisation** • Data format compatibility • Data quality assurance and preservation

…

Way forward

- Consequent use of approaches for better description
  - For example MIAPPE
  - Also for project or legacy data
  - Early involvement of data stewards



How to develop EURISCO towards a trustable repository with acceptably high governance and data management standards?

Way forward

- Develop EURISCO into an integrated European PGR information system (PRO-GRACE and beyond)
  - Add missing sources
  - Connect additional domains
  - Promote standards and protocols
- Remain committed to project cooperation
- Spread the word and raise awareness
- Expand cooperation with bioinformatics hubs