

UPDATES FROM THE
PUBLIC GENETIC RESOURCES
WORKING GROUP

AgBioData Workshop

Chicago, May 2023

Presenter: Moira Sheehan, PhD

Affiliation: Breeding Insight Director

WORKING GROUP MEMBERS



Moira Sheehan
Cornell University



Sunita Kumari
Cold Spring Harbor L.



Jodi Humann
Washington University



Shuyu Liu
Texas A&M University



Yogendra Khedikar
Nuseed



Victoria DeLeo
Bowery Farming



Mária Škrabišová
Palacký Uni. Olomouc

TYPES OF PUBLIC GENETIC RESOURCES

Data sets

Marker sets

Haplotype / Allele databases

~~Reference (pan)genomes~~

Biobank Germplasm

~~Data processing pipelines~~

Private-ness

FAIR-
ness

FAIR AND CARE DATA PRINCIPLES

<https://www.go-fair.org/fair-principles/>



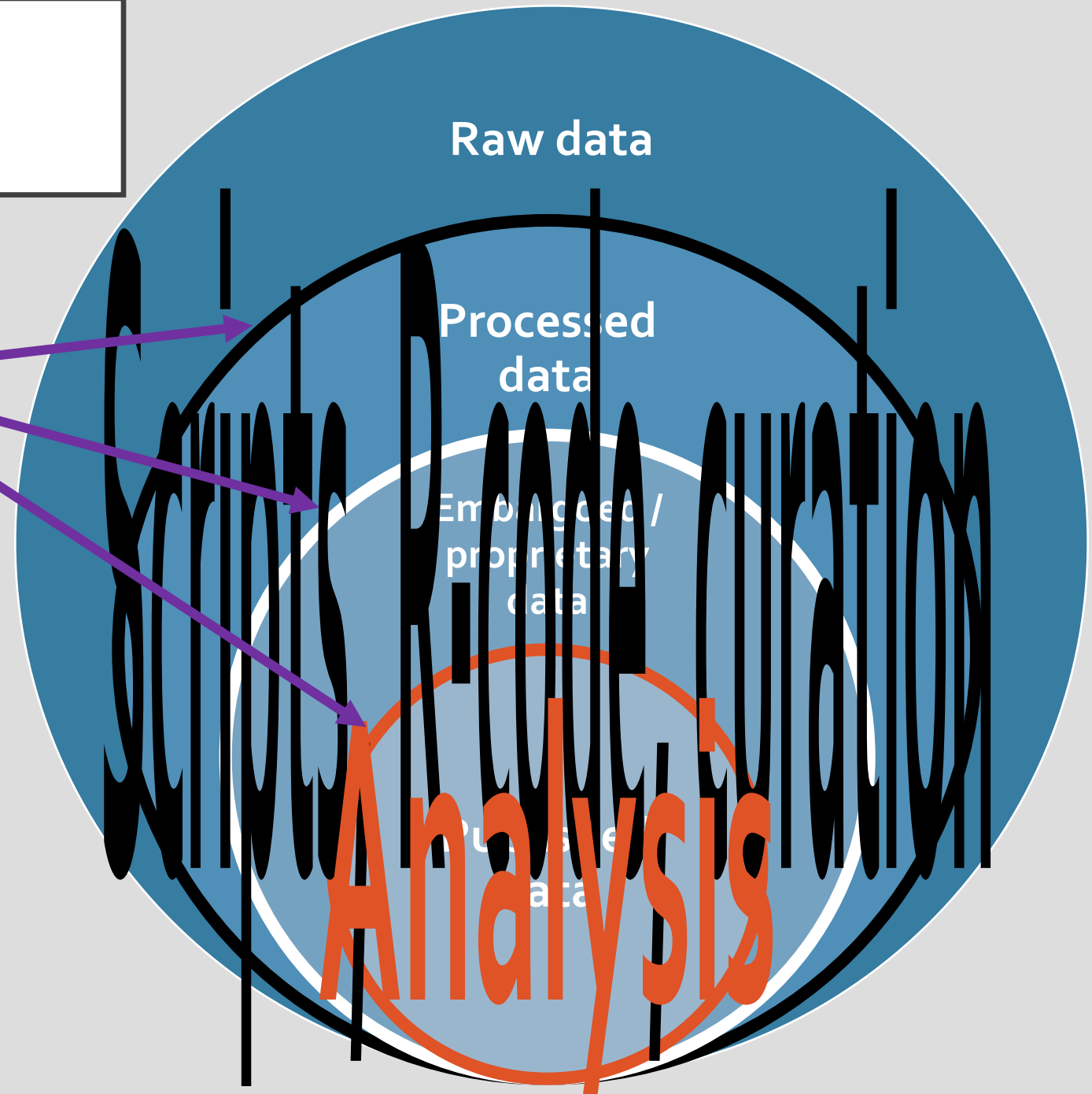
<https://www.gida-global.org/care>

DATA SETS

- Multiple versions are related to each other in a nested relationship.
- Multiple points where data is manipulated.
- Some data sets are not closed, and additional data is added at regular or irregular intervals
- Although not shown, metadata is associated with each of these subsets.

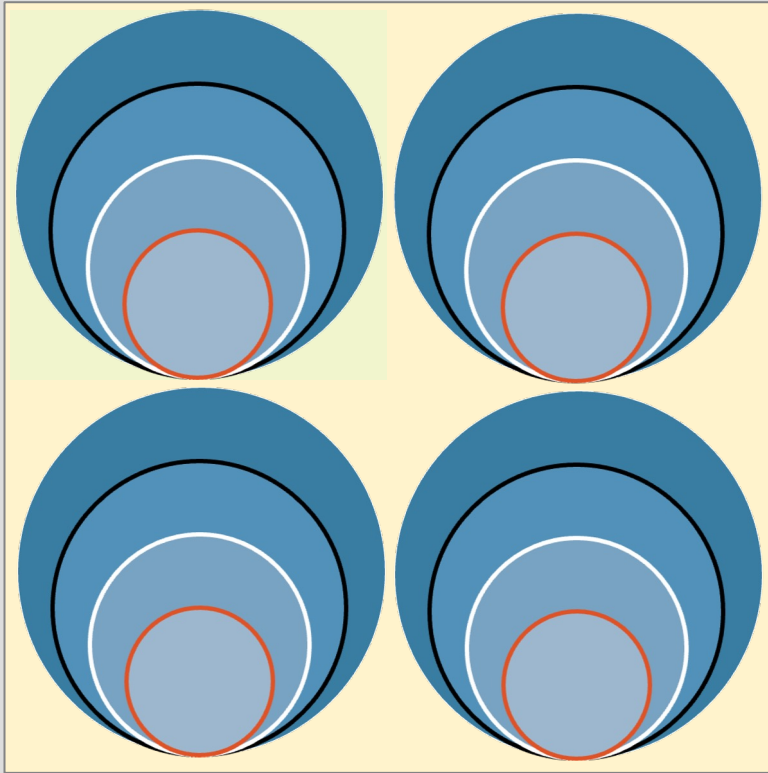
Questions

How do we apply FAIR and CARE data principles to these different facets?



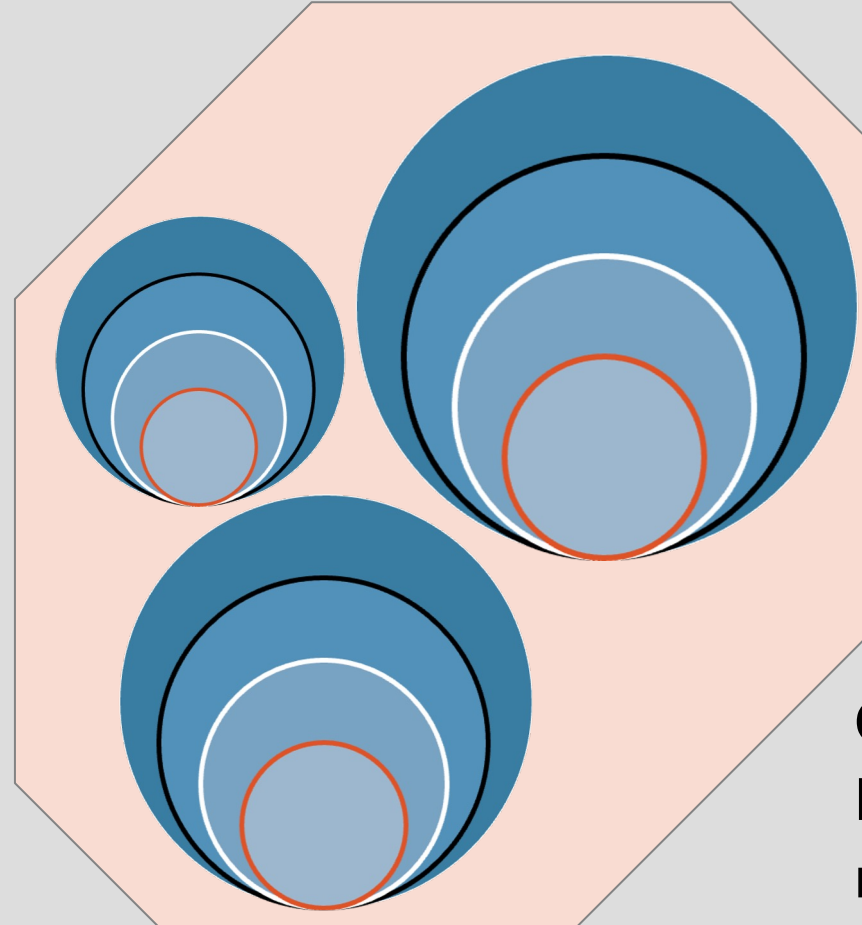
COMBINING DATA SETS

Regular, equally-sized data sets



Weather data, sensor data, etc.
(years, 365 days, 24 hours, etc.)

Irregular, unequally-sized data sets



- Trials
- Locations
- Sub-observations
- Genotypes
- Lab analyses
- Transcriptomics
- Metabolomics
- Others?

Question

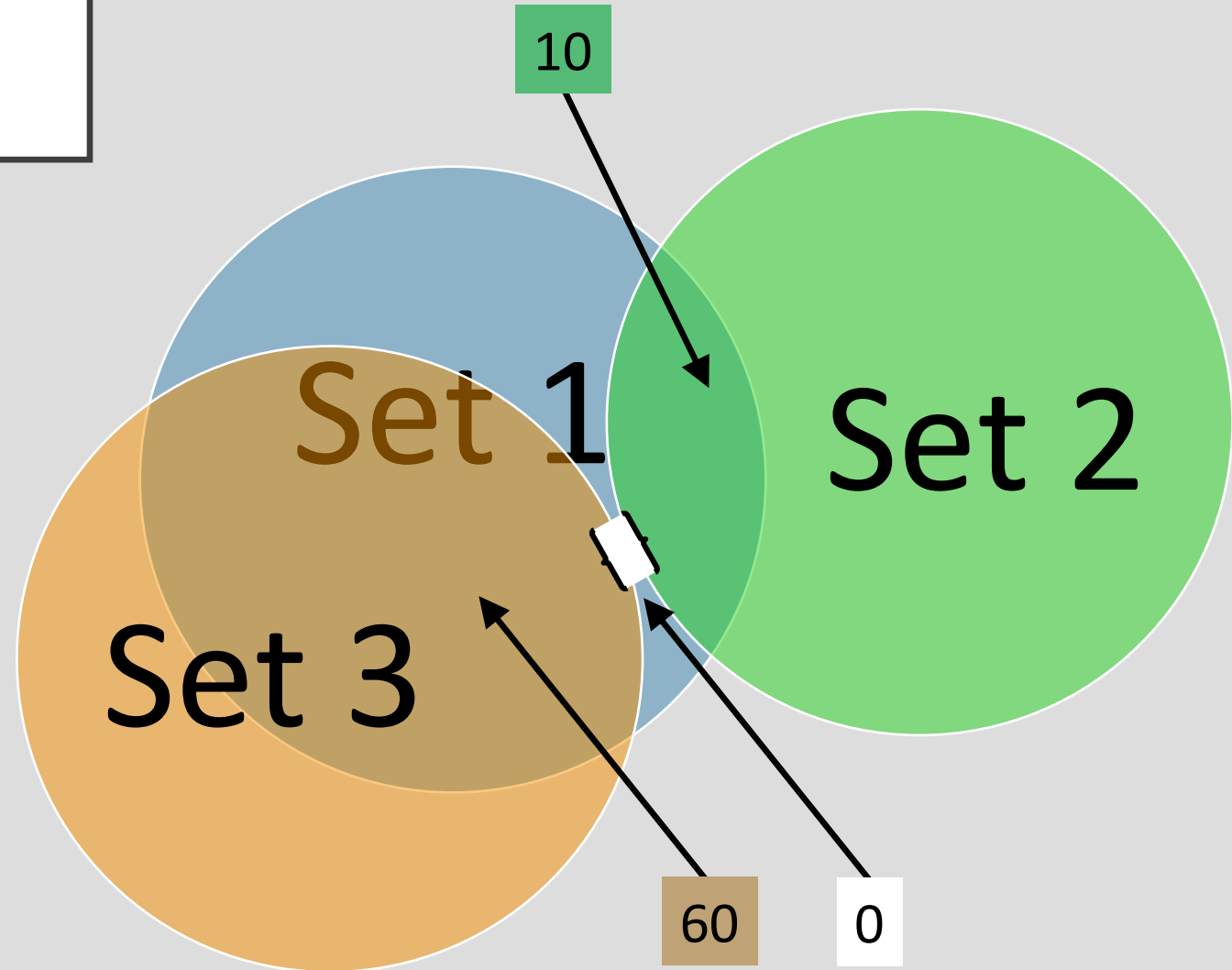
Do irregular datasets need different/more documentation?

MARKER SETS

- Every set of markers is a complete unit, where the markers within each set define the properties and usage.
- Marker sets can be predefined (arrays and pools) or they can be determined post-analysis (GBS).

Questions

1. How do we apply FAIR and CARE data principles to make clear these different sets?
2. What metadata is necessary to maintain?



In the above example, all three sets have a total of 100 markers.

- Set 1 has 10 markers in common with Set 2 and 60 markers in common with Set 3
- Set 3 and Set 2 have 0 markers in common.

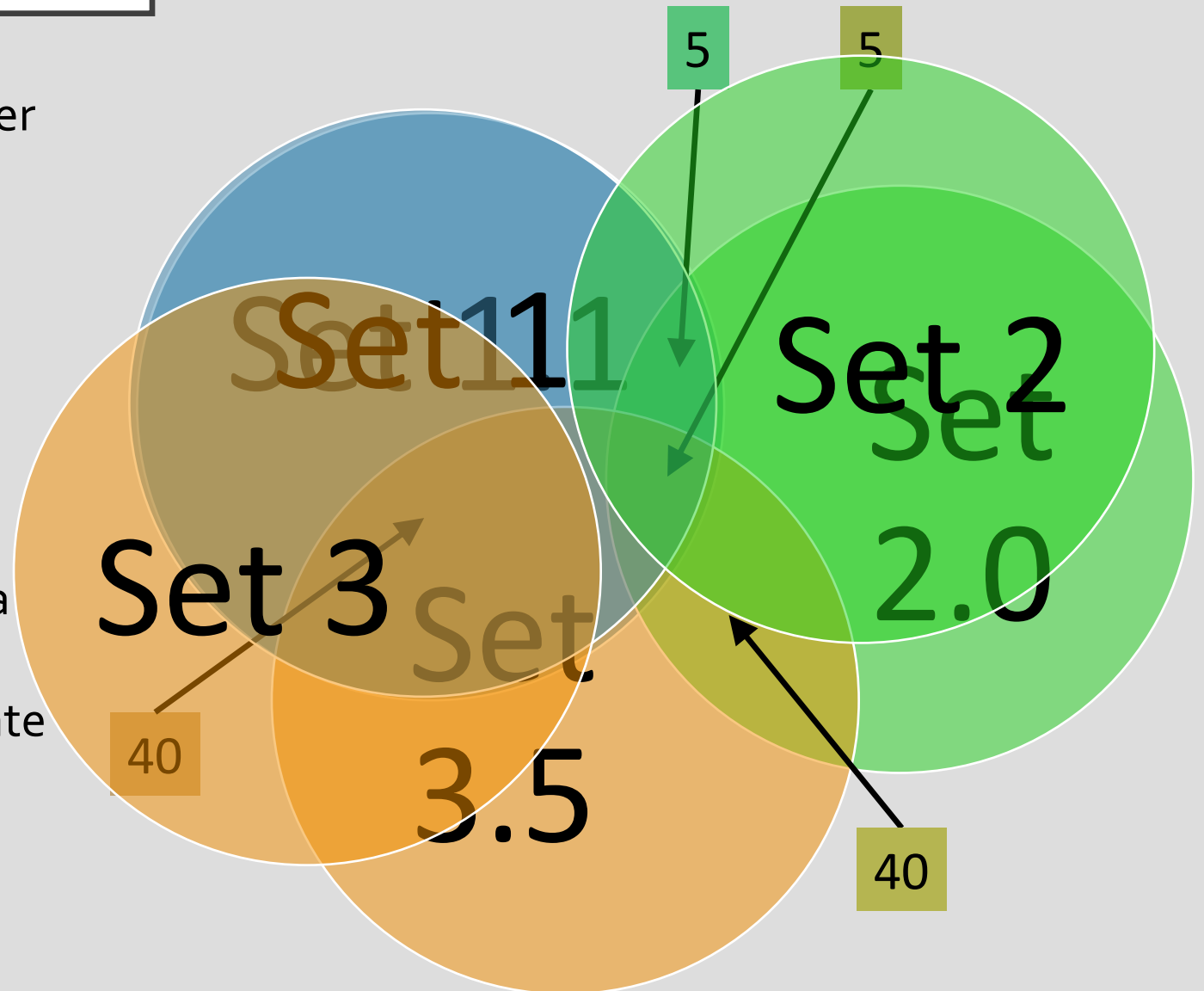
MARKER SET VERSIONING

What happens when marker sets shift over time?

- ↳ Removal of bad markers
- ↳ Addition of new markers
- ↳ Assays are refreshed but VIC/FAM are not assigned to the same alleles

Questions

1. How do we apply FAIR and CARE data principles to make clear what has changed in each set and how they relate to each other?
2. Can we find a way of versioning these changes that provides continuity of information?



HAPLOTYPE / ALLELE DATABASES

The detection of multiple alleles/haplotypes at each marker necessitates the formation of Haplotype databases.

To ensure that users can leverage all data on a marker set by uniquely calling all observed haplotype alleles consistently from experiment to experiment.

GOAL: Identify alleles/haplotypes from MADC from new project(s) for true allele database.

PROCESS:

1. Filter missing data at SAMPLE and MARKER levels (>95% missing data (no reads))
2. Remove adapters for 81 bp amplicons
3. Determine the status of additional alleles (RefMatch and AltMatch)

RESULTS:

1. alfalfa_allele_db_v002.fa
2. alfalfa_allele_db_v002_matchCnt_lut.txt

A genotyping project



Reiterate for each one

RESOURCE:
Compiled database of true alleles for species

PROCESS:

1. Update the true allele database with new alleles
2. Update version number, readme, and look-up table

True Allele Databases

RESOURCE:
Compiled database of true alleles for LETTUCE

RESOURCE:
Compiled database of true alleles for SWEETPOTATO

BIOBANK GERMPLASM

- DOIs to function as a persistent unique identifier to help with interoperability between different banks (like Genesys PGR) in accordance with FAO specifications.
- Sustained and concerted efforts over the years to de-silo international operations.

Additional DOI Resources

- <https://www.genebanks.org/resources/dois/>
- <http://www.fao.org/plant-treaty/areas-of-work/global-information-system/faq/en/> (FAQs)



Food and Agriculture
Organization of the
United Nations

FAO Guidelines

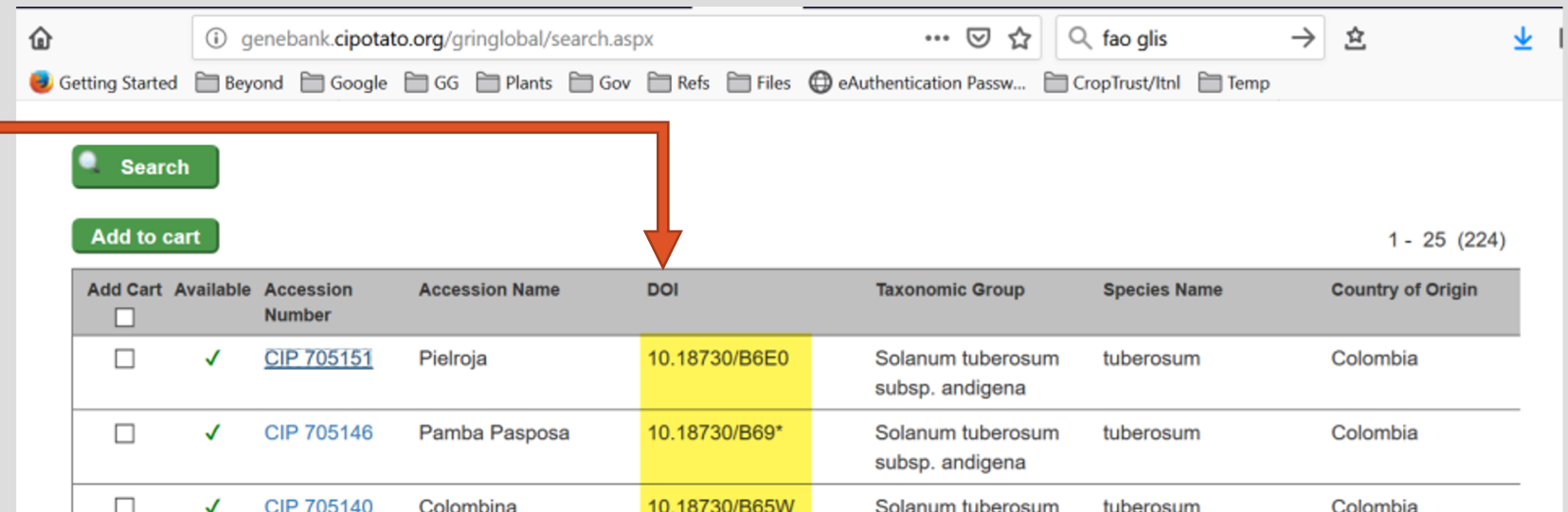
DOIs are used as Permanent Unique Identifiers (PUID) in the context of the Global Information System (GLIS) of Article 17 of the International Treaty on Plant Genetic Resources for Food and Agriculture (ITPGRFA).

- FAO's comprehensive guide "Digital Object Identifiers for Food Crops" is online at <http://www.fao.org/3/I8840EN/i8840en.pdf>
- Guidelines for the optimal use of Digital Object Identifiers as permanent unique identifiers for germplasm samples are found at <http://www.fao.org/3/a-bt114e.pdf>

NPGS GRIN-GLOBAL GERMPLASM

- GRIN Global DOI document (https://www.grin-global.org/docs/gg_doi.pdf)
- Presently, the NPGS is not using the DOI field. In GRIN version 1.10.4
- Grin specifies the DOI format as follows: *“Using Name Records to Store DOIs in GG”*

Some organizations running GRIN-Global have DOI data in their database.



The screenshot shows a web browser window with the URL genebank.cipotato.org/gringlobal/search.aspx. The search results are displayed in a table with the following columns: Add Cart, Available, Accession Number, Accession Name, DOI, Taxonomic Group, Species Name, and Country of Origin. The DOI column is highlighted in yellow, and a red arrow points from the text box on the left to this column.

Add Cart	Available	Accession Number	Accession Name	DOI	Taxonomic Group	Species Name	Country of Origin
<input type="checkbox"/>	✓	CIP 705151	Pielroja	10.18730/B6E0	Solanum tuberosum subsp. andigena	tuberosum	Colombia
<input type="checkbox"/>	✓	CIP 705146	Pamba Pasposa	10.18730/B69*	Solanum tuberosum subsp. andigena	tuberosum	Colombia
<input type="checkbox"/>	✓	CIP 705140	Colombina	10.18730/B65W	Solanum tuberosum	tuberosum	Colombia

INTEROPERABILITY WITH OTHER GERMPLASM BANKS

- This system is to hopefully increase compatibility with other systems e.g., [Genesys PGR](#)
- In addition to these permanent DOIs, GRIN Global maintains accession ID and inventory IDs.
 - The inventory typically represents a packet of seed, a clonal plant, pollen etc.
- Generally, **GRIN users are unaware of the inventory they are receiving**, although in a recent release, curators can specify multiple inventories to distribute (e.g. I'm using this to distribute a DH mapping population; the population is an accession and each DH is an inventory—the users can select which DH they want).

Genesys also uses the text “10.18730/1PGAP” as the link, not the full URL. <https://www.genesys-pgr.org/10.18730/1PGAP>



The screenshot shows the Genesys website interface. At the top, there is a navigation bar with the Genesys logo and a hamburger menu icon. Below the navigation bar, a light blue banner states: "This accession is in the Multilateral System of the ITPGRFA." The main content area displays the "Accession profile: IRGC 4". A table below lists various details:

DOI	10.18730/1PGAP
Holding institute	PHL001 International Rice Research Institute
Location	Philippines
Accession number	IRGC 4
Country of origin	 Malaysia

WORKING GROUP TIMELINE

We are here



Activity	Q1 2023	Q2 2023	Q3 2023	Q4 2023
Define what data/resources qualify as a Public Genetic Resource	✓			
Define the current limitation(s) to FAIR and CARE access for Public Genetic Resources				
Determine if animal resources and plant resources need to be treated differently. What about autopolyploids?				
Determine what an acceptable solution must have to fill the need.				
Determine if any currently available publishing options can be modified or amended (or created anew).				
Create a proposed solution(s) and decision tree (with examples)				
Present recommendations back to AgBioData				

DISCUSSION QUESTIONS



(Virtual and In-Person): Are there any blindspots that have not been considered for the public genetic resources we have chosen to address (can be for a virtual audience, too)?

What is the minimum amount of metadata required to accompany each type of public genetic resources? Is there any way metadata can be verified before publication?

How should genetic resources for autopolyploids be handled to best suit users?

How can we implement CARE data principles into our recommendations?