# Widening the impact of DOI for traceability and ABS

## Global Policy Meets Data Interoperability

Catalysing advanced conservation, management and traceability of PGR
DivSeek Workshop | January 18, 2022


Pankaj Jaiswal

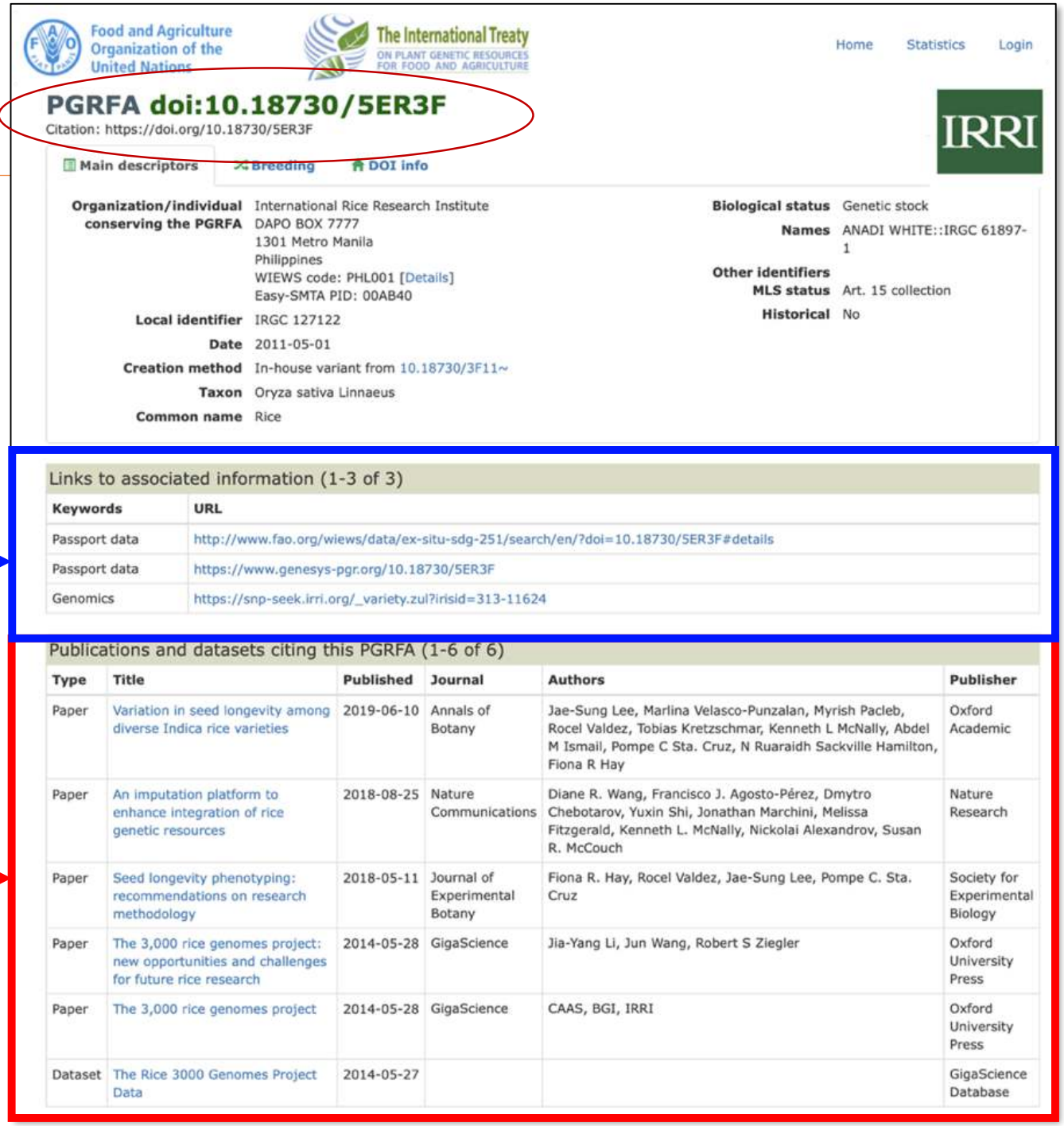
Daniele Manzella
Marco Marsella

Elizabeth Arnaud
Brian King


*Slide Source: B. King*

# Global Information System of the ITPGRFA (GLIS)

Launched Oct. 2017
Over 1.1M DOIs assigned
All of CGIAR
Research centres and Universities
National collections
Started with *ex situ* and breeding
Soon *in situ* and on farm too
DOIs to address identification
Link to web resources
Link to publications and datasets



Food and Agriculture Organization of the United Nations
The International Treaty ON PLANT GENETIC RESOURCES FOR FOOD AND AGRICULTURE
Home Statistics Login

**PGRFA doi:10.18730/5ER3F**
Citation: https://doi.org/10.18730/5ER3F

IRRI

Main descriptors | Breeding | DOI info

| | | | |
|---|---|---|---|
| **Organization/individual conserving the PGRFA** | International Rice Research Institute<br>DAPO BOX 7777<br>1301 Metro Manila<br>Philippines<br>WIEWS code: PHL001 [Details]<br>Easy-SMTA PID: 00AB40 | **Biological status** | Genetic stock |
| | | **Names** | ANADI WHITE::IRGC 61897-1 |
| | | **Other identifiers** | |
| | | **MLS status** | Art. 15 collection |
| **Local identifier** | IRGC 127122 | **Historical** | No |
| **Date** | 2011-05-01 | | |
| **Creation method** | In-house variant from 10.18730/3F11~ | | |
| **Taxon** | Oryza sativa Linnaeus | | |
| **Common name** | Rice | | |

Links to associated information (1-3 of 3)

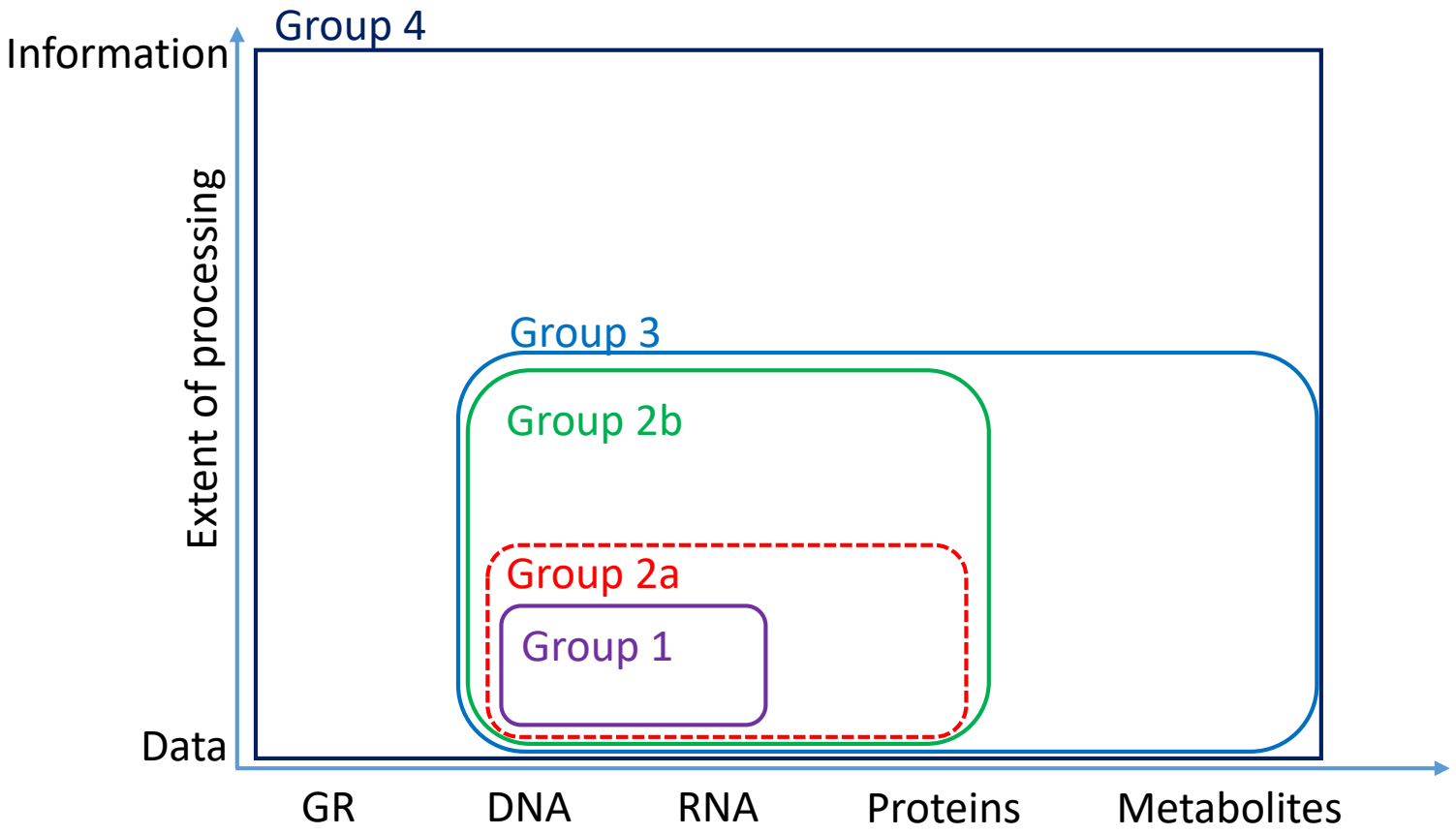
| Keywords | URL |
|---|---|
| Passport data | http://www.fao.org/wiews/data/ex-situ-sdg-251/search/en/?doi=10.18730/5ER3F#details |
| Passport data | https://www.genesys-pgr.org/10.18730/5ER3F |
| Genomics | https://snp-seek.irri.org/_variety.zul?irisid=313-11624 |

Publications and datasets citing this PGRFA (1-6 of 6)

| Type | Title | Published | Journal | Authors | Publisher |
|---|---|---|---|---|---|
| Paper | Variation in seed longevity among diverse Indica rice varieties | 2019-06-10 | Annals of Botany | Jae-Sung Lee, Marlina Velasco-Punzalan, Myrish Pacleb, Rocel Valdez, Tobias Kretzschmar, Kenneth L McNally, Abdel M Ismail, Pompe C Sta. Cruz, N Ruaraidh Sackville Hamilton, Fiona R Hay | Oxford Academic |
| Paper | An imputation platform to enhance integration of rice genetic resources | 2018-08-25 | Nature Communications | Diane R. Wang, Francisco J. Agosto-Pérez, Dmytro Chebotarov, Yuxin Shi, Jonathan Marchini, Melissa Fitzgerald, Kenneth L. McNally, Nickolai Alexandrov, Susan R. McCouch | Nature Research |
| Paper | Seed longevity phenotyping: recommendations on research methodology | 2018-05-11 | Journal of Experimental Botany | Fiona R. Hay, Rocel Valdez, Jae-Sung Lee, Pompe C. Sta. Cruz | Society for Experimental Biology |
| Paper | The 3,000 rice genomes project: new opportunities and challenges for future rice research | 2014-05-28 | GigaScience | Jia-Yang Li, Jun Wang, Robert S Ziegler | Oxford University Press |
| Paper | The 3,000 rice genomes project | 2014-05-28 | GigaScience | CAAS, BGI, IRRI | Oxford University Press |
| Dataset | The Rice 3000 Genomes Project Data | 2014-05-27 | | | GigaScience Database |

*Slide Source: Daniele Manzella & Marco Marsella*

# Digital Sequence Information




Adapted from Houssen et al. (2020)

**Granular Options for Subject Matter Groupings**

**Group 1** - Narrow: DNA and RNA

**Group 2a** includes DNA/RNA sequence data including non-coding sequences, and information on the sequence assembly, including structural annotation and genetic mapping, as well as protein sequence data.

**Group 2b** is the same as group 2a in addition to which it includes functional annotation of genes, gene expression information, epigenetic data, and molecular structures of proteins.

**Group 3** is the same as group 2b, but adds data on other macromolecules and metabolites, including their molecular structures

**Group 4** – Broad: DNA, RNA, protein, metabolites + traditional knowledge, ecological interactions, etc.

*Slide Source and voice: Brian King*

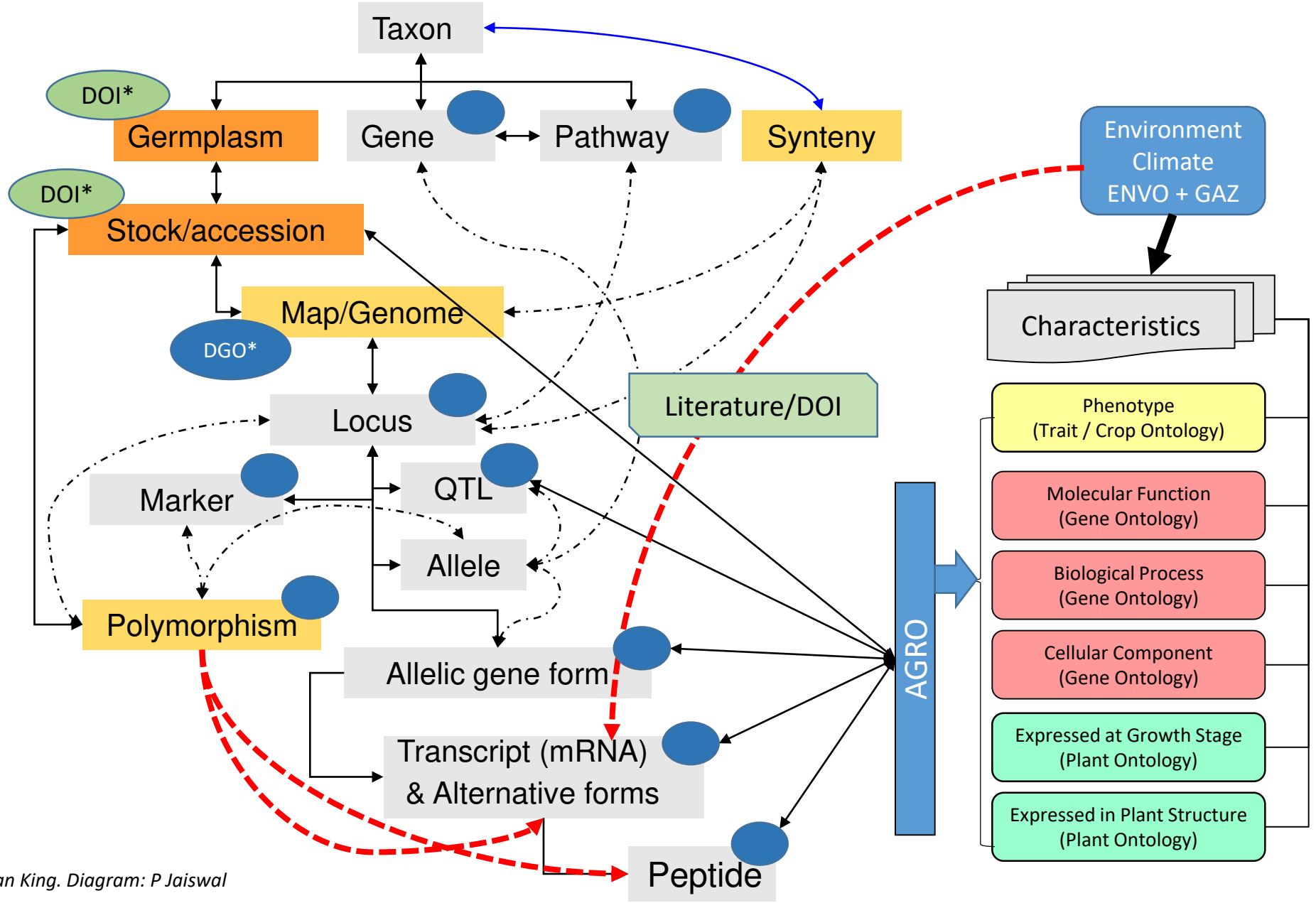
# International Nucleotide Sequence Database (INSDC) Collaboration

Source material is referenced through this attribute which is not mandatory

```
LOCUS       DQ884074                 257 bp    mRNA    linear   EST 24-FEB-2011
DEFINITION  DQ884074 Oryza sativa (indica cultivar-group) cv. IR64 cDNA-AFLP
            fragment Oryza sativa Indica Group cDNA clone 51_9b, mRNA sequence.
ACCESSION   DQ884074
VERSION     DQ884074.1
DBLINK      BioSample: SAMN00165158
KEYWORDS    EST.
SOURCE      Oryza sativa Indica Group (long-grained rice)
  ORGANISM  Oryza sativa Indica Group
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
            Spermatophyta; Magnoliopsida; Liliopsida; Poales; Poaceae; BOP
            clade; Oryzoideae; Oryzeae; Oryzinae; Oryza; Oryza sativa.
REFERENCE   1  (bases 1 to 257)
  AUTHORS   Ventelon-Debout,M., Tranchant-Dubreuil,C., Nguyen,T.-T.-H.,
            Bangratz,M., Sire,C., Delseny,M. and Brugidou,C.
  TITLE     Rice Yellow Mottle Virus stress responsive genes from susceptible
            and tolerant rice genotypes
  JOURNAL   BMC Plant Biol. 8, 26 (2008)
   PUBMED   18315879
COMMENT     Contact: Ventelon-Debout M
            UMR 5096
            Institut de Recherche pour le Developpement
            911 avenue d'Agropolis, BP54501, Montpellier, 34394, France.
FEATURES             Location/Qualifiers
     source          1..257
                     /organism="Oryza sativa Indica Group"
                     /mol_type="mRNA"
                     /cultivar="IR64"
                     /db_xref="taxon:39946"
                     /clone="51_9b"
                     /clone_lib="SAMN00165158 Oryza sativa (indica
                     cultivar-group) cv. IR64 cDNA-AFLP fragment"
```

*Slide Source: Daniele Manzella & Marco Marsella*

# What we propose to build




*Slide Source and voice: Brian King. Diagram: P Jaiswal*

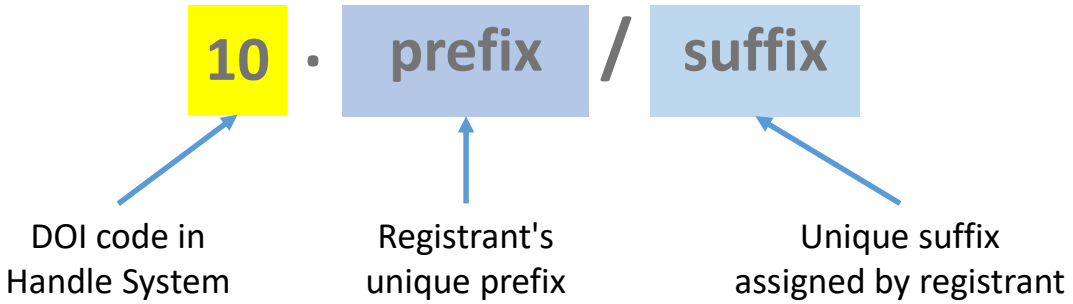
# What is a Digital Genomic Object (DGO)?

A DSI element is a DGO identified by a DOI and collected according to standard/defined procedure with very defined types of component data, can be linked via semantics

*Slide Source: Daniele Manzella & Marco Marsella*

# The DOI System

DOIs (Digital Object Identifiers) are a kind of Permanent Unique identifier (PID)




10.3204/PHPPUBDB-25818
10.18730/1W34E
10.1109/5.771073

The International DOI Foundation maintains the infrastructure

DOIs are assigned by Registration Agencies such as Crossref or DataCite

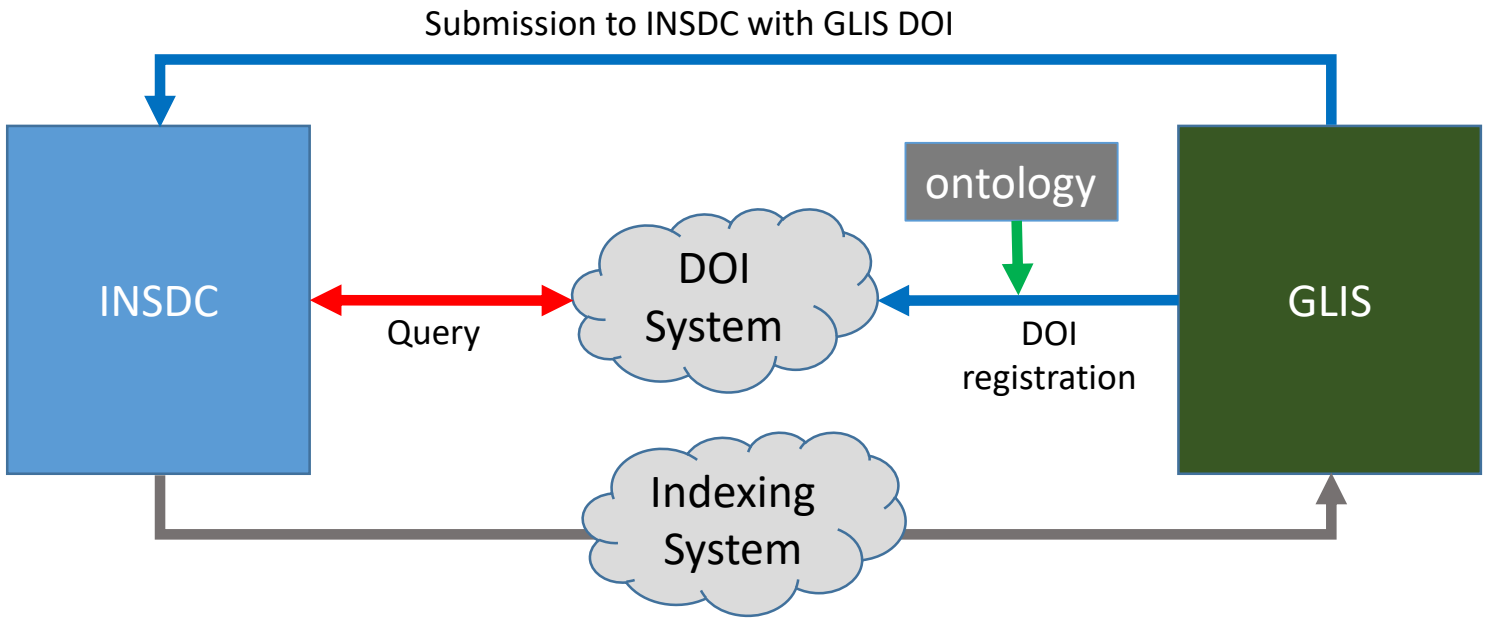
The DOI System offers concepts and tools suitable to our needs

- Flexible metadata structure to describe the different DSI components
- Provisions for establishing relationships among DOIs and other PIDs
- PID Graph, funded by EU project FREYA, records references among PIDs
- Make Data Count, records citation metrics for datasets
- DOIs are already being assigned to
  - Plant Genetic Resources
  - Papers and publications
  - Datasets

*Slide Source: D. Manzella & Marco Marsella*

# Our pilot proposal

Scope: Select a couple of  species data sources




- Researcher obtains DOI for source material from GLIS is not already available
- Researcher submits dataset to INSDC referencing GLIS DOI for source material
- GLIS discovers INSDC records referencing its own DOIs using an Indexing System
- If INSDC opts to assign DOIs to its Accessions, the Indexing System is not required

*Slide Source: D. Manzella & Marco Marsella*

# INSDC record after injection of DOI

```
LOCUS       DQ884074                257 bp    mRNA    linear   EST 24-FEB-2011
DEFINITION  DQ884074 Oryza sativa (indica cultivar-group) cv. IR64 cDNA-AFLP
            fragment Oryza sativa Indica Group cDNA clone 51_9b, mRNA sequence.
ACCESSION   DQ884074
VERSION     DQ884074.1
DBLINK      BioSample: SAMN00165158
KEYWORDS    EST.
SOURCE      Oryza sativa Indica Group (long-grained rice)
  ORGANISM  Oryza sativa Indica Group
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
            Spermatophyta; Magnoliopsida; Liliopsida; Poales; Poaceae; BOP
            clade; Oryzoideae; Oryzeae; Oryzinae; Oryza; Oryza sativa.
REFERENCE   1  (bases 1 to 257)
  AUTHORS   Ventelon-Debout,M., Tranchant-Dubreuil,C., Nguyen,T.-T.-H.,
            Bangratz,M., Sire,C., Delseny,M. and Brugidou,C.
  TITLE     Rice Yellow Mottle Virus stress responsive genes from susceptible
            and tolerant rice genotypes
  JOURNAL   BMC Plant Biol. 8, 26 (2008)
   PUBMED   18315879
COMMENT     Contact: Ventelon-Debout M
            UMR 5096
            Institut de Recherche pour le Developpement
            911 avenue d'Agropolis, BP54501, Montpellier, 34394, France.
FEATURES             Location/Qualifiers
     source          1..257
                     /organism="Oryza sativa Indica Group"
                     /mol_type="mRNA"
                     /cultivar="IR64"
                     /db_xref="taxon:39946"
                     /db_xref="https://doi.org/10.18730/V918="
                     /clone="51_9b"
                     /clone_lib="SAMN00165158 Oryza sativa (indica
                     cultivar-group) cv. IR64 cDNA-AFLP fragment"
```

GLIS DOI

*Slide Source: Daniele Manzella & Marco Marsella*

What we propose to build

What key systems or types of systems we will need to connect with to be able to implement linked data management between HTP-enabled breeding and genebank operational domains--using DGOs/data interoperability as the 'connective tissue'?

Global Information System (GLIS) as a link to DOI System and onward link to Genesys/other systems

Germplasm that is publicly released gets DOI from GLIS

DGOs are generated for digital sequence information--and these are assigned DOIs.

Markers DGOs (in accomm different markers DGOs ar kind of o Sequenc This can assigned

Systems

Process steps

Questions

Use-cases

Genesys

Is new DSI ontology with links to other ontologies identified here sufficient?

do we need expanded stress, disease ontologies?

SOFA Ontology relevant?

Could Cyverse help manage and promote this infrastructure as a more public-good actor?

DOIs

GLIS public information system

CIAT Genebank DBs

Accessions linked to DOCs

CIAT Breeding Management System

DOIs are pulled from GRINGlobal into breeding process

(potential use for ICRISAT breeding portal)

image-based CNNs for trait recognition, recognizing biotic/abiotic stress, recognizing crop growth stage

Material goes to a trial

ROVER

high-throughput phenotyping data collection

Phenotyping Repository (HTP data)

High-throughput Phenotyping Data are annotated using reference Ontologies, and thus enabling linkages to associated DGOs and DOIs

Store 2 raw data types associated with trials (provenance linkages--e.g. experiment ID maybe something like temporary DOI): phenotyping data, raw 'shallow sequencing' data to pull out markers.

A subset of the raw data is chosen for analysis.

DGOs

HTP data analysis

A marker is already defined and has DGO associated with it, and you are looking for it in the raw data

annotated data assets are stored and exposed to enable other research.

Curated data assets created from the raw data

Decide what data to keep, curate from HTP e.g. each raw data over time series or "cumulative traits"

Analyzed datasets produced (includes markers and...?) a DGO is assigned to markers (mendelian or sequence) or any other Digital Sequence Data--this links back to associated DOI.

*Slide Source: Brian King*

CGIAR | Platform for Big Data in Agriculture

# THANK YOU

pankaj.jaiswal@oregonstate.edu

*Slide Source: Brian King*