



# Webinar

## Wheat Data Interoperability guidelines

research data sharing without barriers  
[rd-alliance.org](http://rd-alliance.org)

# Introduction and contexte

- Created in 2011 following endorsement by G20 Agriculture Ministries to improve food security
- A framework to identify synergies and facilitate collaborations for wheat improvement at the international level
- The Wheat Initiative members
  - Countries: Argentina, Australia, Brazil, Canada, China, France, Germany, Hungary, India, Ireland, Italy, Japan, Spain, Turkey, UK, USA
  - International organizations: CIMMYT, ICARDA
  - Private companies: Arvalis, Bayer CropScience, Florimond Desprez V&F, KWS UK, Limagrain, Monsanto Company, RAGT 2n Saateen Union Research, Syngenta Crop Protection

# The Wheat Data Interoperability WG

- Aims: contribute to the improvement of Wheat related data interoperability by
  - Building a common interoperability framework (metadata, data formats and vocabularies)
  - Providing guidelines for describing, representing and linking Wheat related data

## Contributors



## Sponsors

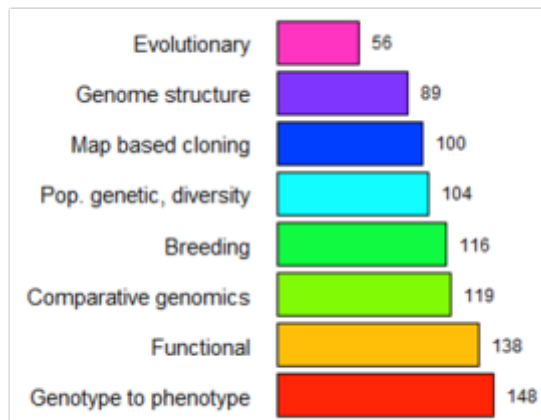


**Contributors:** Alaux Michael (INRA, France), Aubin Sophie (INRA, France), Arnaud Elizabeth (Bioversity, France), Baumann Ute (Adelaide Uni, Australia), Buche Patrice (INRA, France), Cooper Laurel (Planteome, USA), Fulss Richard (CIMMYT, Mexico), Hologne Odile (INRA, France), Laporte Marie-Angélique (Bioversity, France), Larmand Pierre (IRD, France), Letellier Thomas (INRA, France), Lucas Hélène (INRA, France), Pommier Cyril (INRA, France), Protonotarios Vassilis (Agro-Know, Greece), Quesneville Hadi (INRA, France), Shrestha Rosemary (INRA, France), Subirats Imma (FAO of the United Nations, Italy), Aravind Venkatesan (IBC, France), Whan Alex (CSIRO, Australia)

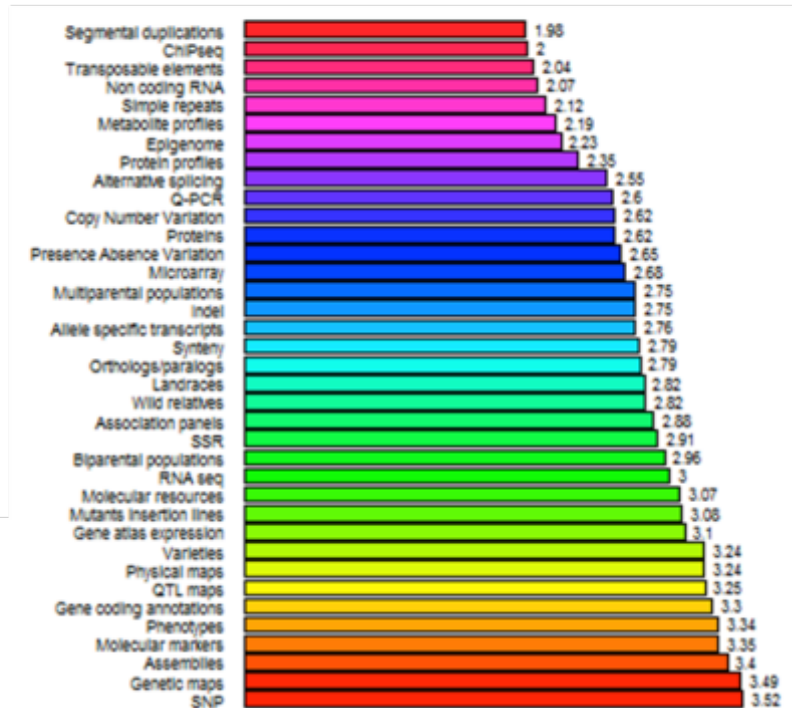
**Co-chairs:** Esther Dzali Yehou Kaboré (INRA, France), Richard Allan Fulss (CIMMYT, Mexico)

## State of the art

### Studies



### Data types



### Repositories



# The methodology

## Surveys

- Landscape of Wheat related standards and their use by the community
- Comprehensive overview of Wheat related ontologies and vocabularies

## Workshops

- Recommendations
- Mappings between different data formats
- Actions to conduct in order to improve the current level of Wheat related data interoperability
- Interoperability use cases

## Implementation

- Interactive cookbook: recommendations + guidelines
- A repository of Wheat related linked vocabularies (Bioportal)

## The outputs of the WDI working group

- Guidelines (<http://wheatis.org/DataStandards.php>)
  - Data exchange formats
    - Example: VCF (Variant Call Format) for sequence variation data, GFF3 for genome annotation data, etc.
  - Data description best practices
    - Consistent use of ontologies, consistent use of external database cross references
  - Data sharing best practices
    - Share data matrices along with relevant metadata (example: trait along with method, units and scales or environmental ones)
  - Useful tools and use cases that highlight data formats and vocabularies issues
- A repository of wheat related ontologies and vocabularies (<http://wheat.agroportal.lirmm.fr/ontologies>)
  - Allows the access to the ontologies and vocabularies through APIs.
- A prototype
  - Implementation of use cases of wheat data integration within the AgroLD (Agronomic Linked Data) tool: <http://www.agrold.org>



About

Collaborators

Search

Data Standards

Submit Data

Tools

Links

WheatIS Nodes

## WheatIS

@ PRATT J.C. / INRA



## About

This project aims at building an International Wheat Information System, called hereafter WheatIS, to support the wheat research community. The main objective is to provide a single-access web base system to access to the available data resources and bioinformatics tools.

This project is based on the principles listed below:

- Collective building of the WheatIS to better respond to the needs of the international wheat community;
- Incremental implementation to offer rapidly an operational information system;
- Emphasis on Quality Assurance to serve as a framework for an approach with incremental implementation;
- Promotion of an open-access model for data exchange;
- Reliance on a distributed system;
- Use of Virtual Machine and Cloud Computing technologies to facilitate sharing data and tools;
- Promotion of the visibility of each participating platform to contribute to their sustainability.

Home

Guidelines

Ontology

## Sequence variations

The sequence variations are the nucleotides differences between two (or several) sequences at the same locus (usually between a reference sequence and another sequence). Three types of sequence variations—single-nucleotide polymorphisms (SNPs), insertions and deletions (indels), and short tandem repeats (STRs)—have been mainly reported in plant genomes. The most currently available sequence variations for wheat are SNPs.

## Recommendations

### Summary

For Variant (e.g. SNP) calling performed by bioinformaticians:

1. Use a reference wheat genome sequence
2. Data format: Use the VCF
3. Provide associated metadata

### 1. Reference sequence

The currently most commonly used reference bread wheat sequence is the IWGSC survey sequence (cv Chinese Spring), available at the [IWGSC Sequence Repository](#) and EBI.

When available, we encourage the use of the chromosomes reference sequence.

### 2. Data format

We recommend to use the latest VCF file format.

**Description**

The Variant Call Format (VCF) is a text file used in bioinformatics for storing gene sequence variations. The format has been developed with the advent of large-scale genotyping and DNA sequencing projects, such as the 1000 Genomes Project. VCF format specifications can be found [here](#).

**Warning:** The VCF files generated for exome capture need to be labeled as such and can not be merged with those from IWGSC context.

### 3. Metadata

We recommend to provide a minimal set of metadata to contextualize the provenance of the SNPs and to provide information about the SNP quality analysis.

**Data sharing**

For data sharing, the following information should be provided in the header section of the VCF file (header lines have to be preceded by “##” characters) or as a separate tabulated file.

Name	Description
RUN NAME	Name of the sequencing run that produced the data we are interested in.
RUN DESCRIPTION	Description of this run.
SUB RUN NAME	Part of a sequencing run that produced the data we are interested in. According to the sequencing technology involved, the sub run can be a lane (for 454 sequencers), a flowcell (for Illumina sequencers)...
ANALYSIS NAME	Name of the SNP calling analysis
ANALYSIS SOFTWARE NAME	Software used for the SNP calling analysis
ANALYSISCONTACT NAME	Person who performed the analysis
PROTOCOL NAME	Name of the sequencing protocol
MAPPING GENOME NAME	Name and version of the reference genome used to call the variations
MAPPING GENOME TAXON NAME	Taxon of the reference genome used to call the variations
MAPPING_GENOME DESCRIPTION	Description of the reference genome used to call the variations
GENOTYPE NAME	Name of the sample/individual that has been sequenced.
GENOTYPE TAXON	Taxon of the sample/individual that has been sequenced.
PROJECT NAME	Name of the project that funded the sequencing
FILTERS	Filters applied to call SNPs (ex: DP > 10)

**Warning:** BAM/SAM files should be kept for traceability of further analysis since they are not suitable for sharing.

**Data submission**

For data submission in international repositories (EBI, NCBI), we advise to fill the dedicated XML format (<http://www.ebi.ac.uk/ena/submit/preparing.xmls#vcf>).

## Most popular Tools

Identification of sequence variations includes 3 steps :

1. Mapping of the reads on the reference genome
2. Calling the sequence variations
3. Filtering out irrelevant results regarding mainly depth and sequence quality and mapping quality.

## Mapping tools

- BWA
- Bowtie
- Bowtie 2

## SNP calling tools

- GATK
- SAM tools

## Filter tools

- VCF tools
- VCF utils
- SAM tools

## Example

Example of a VCF file dedicated to wheat data:

```
##fileformat=VCFv4.1
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT 102 403 407-IV_60 93 ACBarrie A
labasskaja CS Estacao M6 Marquis Neepawa PI153785 PI166180 PI166333 PI177943
PI185715 PI192001 PI192147 PI192569 PI210945 PI222669 PI245368 PI262611 PI278
297 PI349512 PI366716 PI366905 PI382150 PI406517 PI445736 PI470817 PI477870 P
I481718 PI481923 PI565213 PI82469 PI8813 PR267 Roemer Taxi Utmost acc1 acc2 a
cc3 acc4 acc5 berkut chakwa186 cham6 clear_white dharwar_dry hidhab klein_cha
maco opata pavon pbw343 rac875 vorobey
3929455_1al 1623 . T C 245.53 . AC=18;AF=0.196;AN=92;BaseQRankSum=-0.079;DP=48
;DeIs=0.00;FS=0.000;HaplotypeScore=0.1087;InbreedingCoeff=0.2057;MLEAC=18;MLE
AF=0.196;MQ=100.00;MQ0=0;MQRankSum=-1.426;QD=27.28;ReadPosRankSum=-0.158 GT:A
D:DP:GQ:PL 0/0:1,0:1:3:0,3,41 0/0:1,0:1:3:0,3,41 1/1:0,1:1:3:41,3,0 1/1:0,1:1
:3:41,3,0 ./ 0/0:1,0:1:3:0,3,41 0/0:1,0:1:3:0,3,39 ./ 0/0:1,0:1:3:0,3,39 ./
./ 1/1:0,1:1:3:39,3,0 0/0:1,0:1:3:0,3,39 ./ 1/1:0,1:1:3:38,3,0 1/1:0,1:1:
3:39,3,0 0/0:1,0:1:3:0,3,39 0/0:1,0:1:3:0,3,39 1/1:0,1:1:3:39,3,0 / 0/0:1,0
```

**PROMOTE**

the adoption of common standards, vocabularies and best practices for Wheat data management

Guidelines

Use Cases



# Wheat Data Interoperability Guidelines

[Home](#) [Guidelines](#) [Ontologies & Vocabularies](#) [Use cases](#) [Getting involved](#) [About](#)

[Home](#) / [Ontologies & Vocabularies](#)

## Ontologies & Vocabularies

In the context of Research Data, the use of vocabularies play a key role in managing, sharing and publishing data. Vocabularies enhance the quality of the interoperability and effectiveness of data exchange, thus facilitating the re-usage of data by others and in the process adding value to the local researcher.

This section focuses on vocabularies, their benefits and current situation in the context of Wheat Research Data. The aim is to provide a tool to support researchers in the selection of vocabularies to adopt according to the Wheat Data Interoperability Guidelines.

### What type of vocabularies exist in the context of the Wheat Initiative?

There are different types of vocabularies like ontologies, thesauri, classification systems, controlled lists, syntax encoding standards, authority data, controlled vocabularies, taxonomies, glossaries, etc.

### Why are vocabularies important for the Wheat Initiative?

#### What benefits can vocabularies bring to your daily work as a researcher?

They are many, including:

- › research visibility
- › research usage
- › research uptake

### What are currently the most used and relevant vocabularies in the context of Wheat Initiative?

From December 2014 to January 2015 the editorial team conducted a survey "Towards a Comprehensive Overview of Ontologies and Vocabularies for Research on Wheat". The objective was to collect information about the visibility, interoperability, domain, content and other technical aspects of relevant ontologies and vocabularies. As a result, in February 2015 a report (link) was published, and also a list of vocabularies listed as follows:

#### GETTING INVOLVED



WheatIS

## Why adopt the WDI guidelines?

# Benefits for 3 main target users

13



## As a data producer or manager

- Easily conform to the well-recognized data repositories and facilitate the deposit of your data within these repositories;
- Share common meanings of the words you utilize to describe your data and make your data more machine-readable and computable
- Contribute to foster the development of smarter search tools and make your data more visible and discoverable

## As a wheat related information system or tool developer

- Basing your tool or information system on the recommended data formats and vocabularies will make it easier to integrate data from various data sources, deliver smarter outputs for a wider audience

## As a wheat related ontology developer

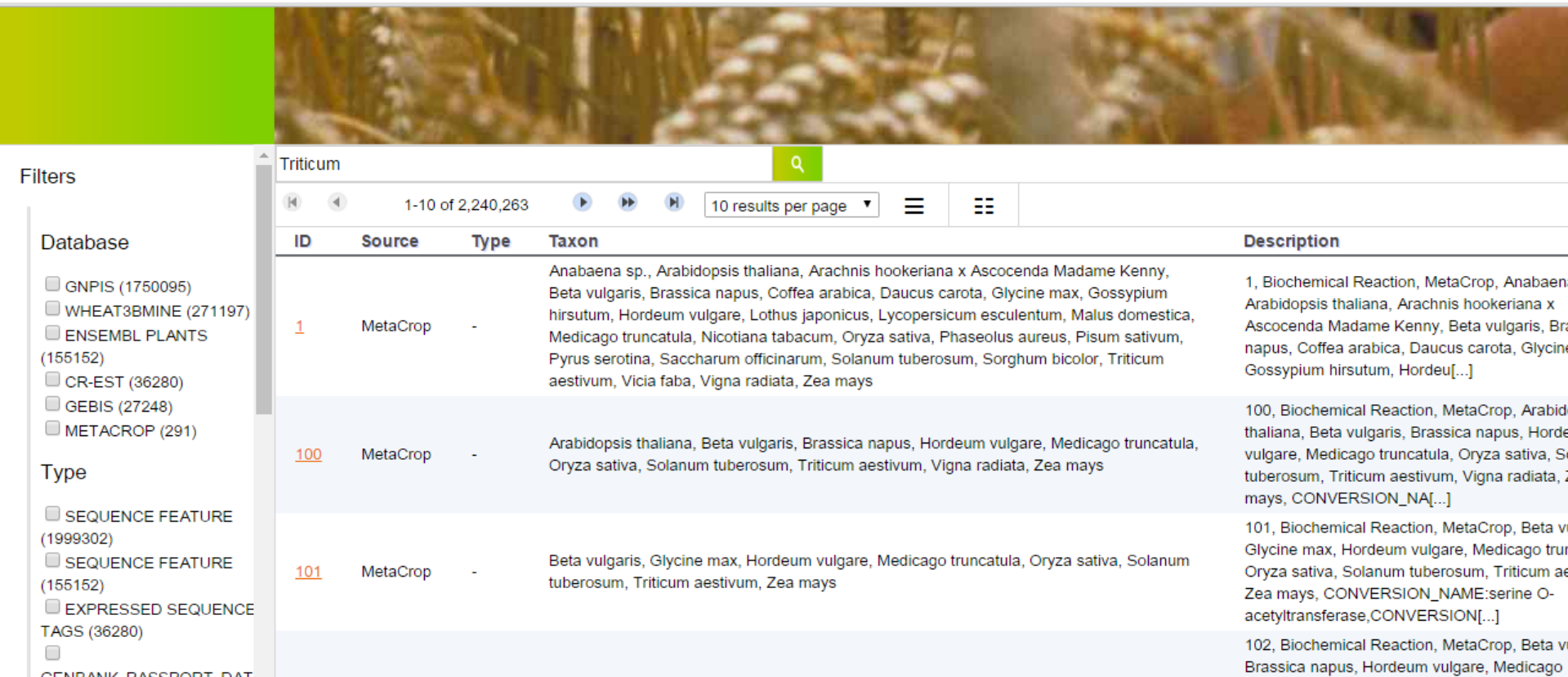
- Share your ontologies through the WDI wheat ontologies portal and make them more visible to the community
- Reuse or link your ontologies to existing concepts and terms in wheat related ontologies to enrich them, make them more visible and in some cases save you time.

# The international wheat information system

## (WheatIS: <https://urgi.versailles.inra.fr/wheatis/>)

14

- Provide a single-access web based system to access to the available data resources and bioinformatics → work in progress
- 6 nodes already connected to the WheatIS search. Work in progress to connect more nodes. More information in <http://wheatis.org/WheatIS%20nodes.php>



The screenshot displays the WheatIS web interface. On the left, there is a sidebar with filters for 'Database' and 'Type'. The 'Database' filter includes GNPIS (1750095), WHEAT3BMINE (271197), ENSEMBL PLANTS (155152), CR-EST (36280), GEBIS (27248), and METACROP (291). The 'Type' filter includes SEQUENCE FEATURE (1999302), SEQUENCE FEATURE (155152), EXPRESSED SEQUENCE TAGS (36280), and GENBANK PASSPORT DAT. The main content area shows a search for 'Triticum' with a magnifying glass icon. Below the search bar, there is a table with 5 columns: ID, Source, Type, Taxon, and Description. The table displays 10 results per page, showing results 1, 100, and 101. The first result (ID 1) is a Biochemical Reaction from MetaCrop, involving various plant species. The second result (ID 100) is a Biochemical Reaction from MetaCrop, involving Arabidopsis thaliana and other species. The third result (ID 101) is a Biochemical Reaction from MetaCrop, involving Beta vulgaris and other species.

Filters

Database

- ☐ GNPIS (1750095)
- ☐ WHEAT3BMINE (271197)
- ☐ ENSEMBL PLANTS (155152)
- ☐ CR-EST (36280)
- ☐ GEBIS (27248)
- ☐ METACROP (291)

Type

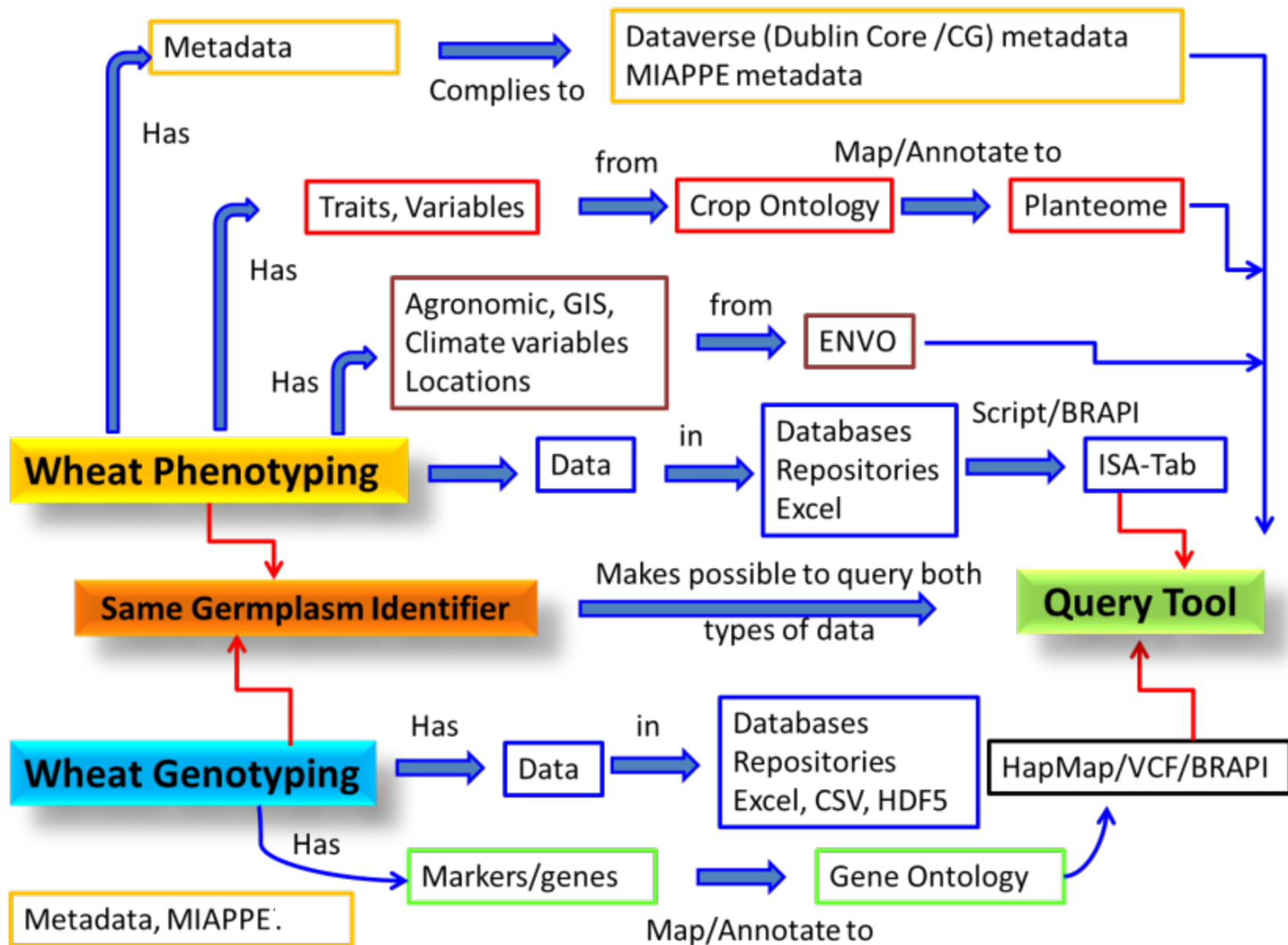
- ☐ SEQUENCE FEATURE (1999302)
- ☐ SEQUENCE FEATURE (155152)
- ☐ EXPRESSED SEQUENCE TAGS (36280)
- ☐ GENBANK PASSPORT DAT

Triticum

1-10 of 2,240,263 10 results per page

ID	Source	Type	Taxon	Description
1	MetaCrop	-	Anabaena sp., Arabidopsis thaliana, Arachnis hookeriana x Ascocenda Madame Kenny, Beta vulgaris, Brassica napus, Coffea arabica, Daucus carota, Glycine max, Gossypium hirsutum, Hordeum vulgare, Lothus japonicus, Lycopersicum esculentum, Malus domestica, Medicago truncatula, Nicotiana tabacum, Oryza sativa, Phaseolus aureus, Pisum sativum, Pyrus serotina, Saccharum officinarum, Solanum tuberosum, Sorghum bicolor, Triticum aestivum, Vicia faba, Vigna radiata, Zea mays	1, Biochemical Reaction, MetaCrop, Anabaena hookeriana x Arachnis hookeriana x Ascocenda Madame Kenny, Arabidopsis thaliana, Arachnis hookeriana x Ascocenda Madame Kenny, Beta vulgaris, Brassica napus, Coffea arabica, Daucus carota, Glycine max, Gossypium hirsutum, Hordeu[...]
100	MetaCrop	-	Arabidopsis thaliana, Beta vulgaris, Brassica napus, Hordeum vulgare, Medicago truncatula, Oryza sativa, Solanum tuberosum, Triticum aestivum, Vigna radiata, Zea mays	100, Biochemical Reaction, MetaCrop, Arabidopsis thaliana, Beta vulgaris, Brassica napus, Hordeum vulgare, Medicago truncatula, Oryza sativa, Solanum tuberosum, Triticum aestivum, Vigna radiata, Zea mays, CONVERSION_NAME[...]
101	MetaCrop	-	Beta vulgaris, Glycine max, Hordeum vulgare, Medicago truncatula, Oryza sativa, Solanum tuberosum, Triticum aestivum, Zea mays	101, Biochemical Reaction, MetaCrop, Beta vulgaris, Glycine max, Hordeum vulgare, Medicago truncatula, Oryza sativa, Solanum tuberosum, Triticum aestivum, Zea mays, CONVERSION_NAME:serine O-acetyltransferase,CONVERSION[...]
102	MetaCrop	-	Beta vulgaris, Glycine max, Hordeum vulgare, Medicago truncatula, Oryza sativa, Solanum tuberosum, Triticum aestivum, Zea mays	102, Biochemical Reaction, MetaCrop, Beta vulgaris, Glycine max, Hordeum vulgare, Medicago truncatula, Oryza sativa, Solanum tuberosum, Triticum aestivum, Zea mays

# Search for stem rust resistant Germplasm and genes associated with it



- Enriches data and enables data analysis and visualization
- Demo with QTLNetMiner (<https://ondex.rothamsted.ac.uk/QTLNetMiner/>)
  - QTLNetMiner is one of the nodes of the WheatIS
  - Use case: search for candidate genes of “drought tolerance”



- Demo with AgroLD (<http://www.agrold.org>)



- Use case: explore relationships between the following concepts: “root development”, “triticum aestivum” and “triticum urartu”



# QTLNETMINER

A tool for the discovery of candidate genes controlling complex traits.

Use the query suggestor to find alternative search queries to improve your results



17

?

1464 documents and 23845 genes will be found with this query

☐ Query Suggestor ?

☐ Genome or QTL Search ?

☐ Gene List ?

In total 23845 genes were found. Top 100 genes are displayed in Map view.  
Query was found in 1464 documents related with genes (3012 documents in total)

[Download as TAB delimited file](#)  
Select gene(s) and click "Show Network" button to see the Ondex network. ?

Max number of genes to show:  Known targets: ☐ Novel targets: ☐

ACCESSION	CHROM	START	SCORE	USER	QTL	EVIDENCE	Select
-----------	-------	-------	-------	------	-----	----------	--------

Define a QTL region you are interested in

Include a list of gene names and see if they are related to your keyword

Map View

Gene View

Evidence View

Network View

18

[Download as TAB delimited file](#)

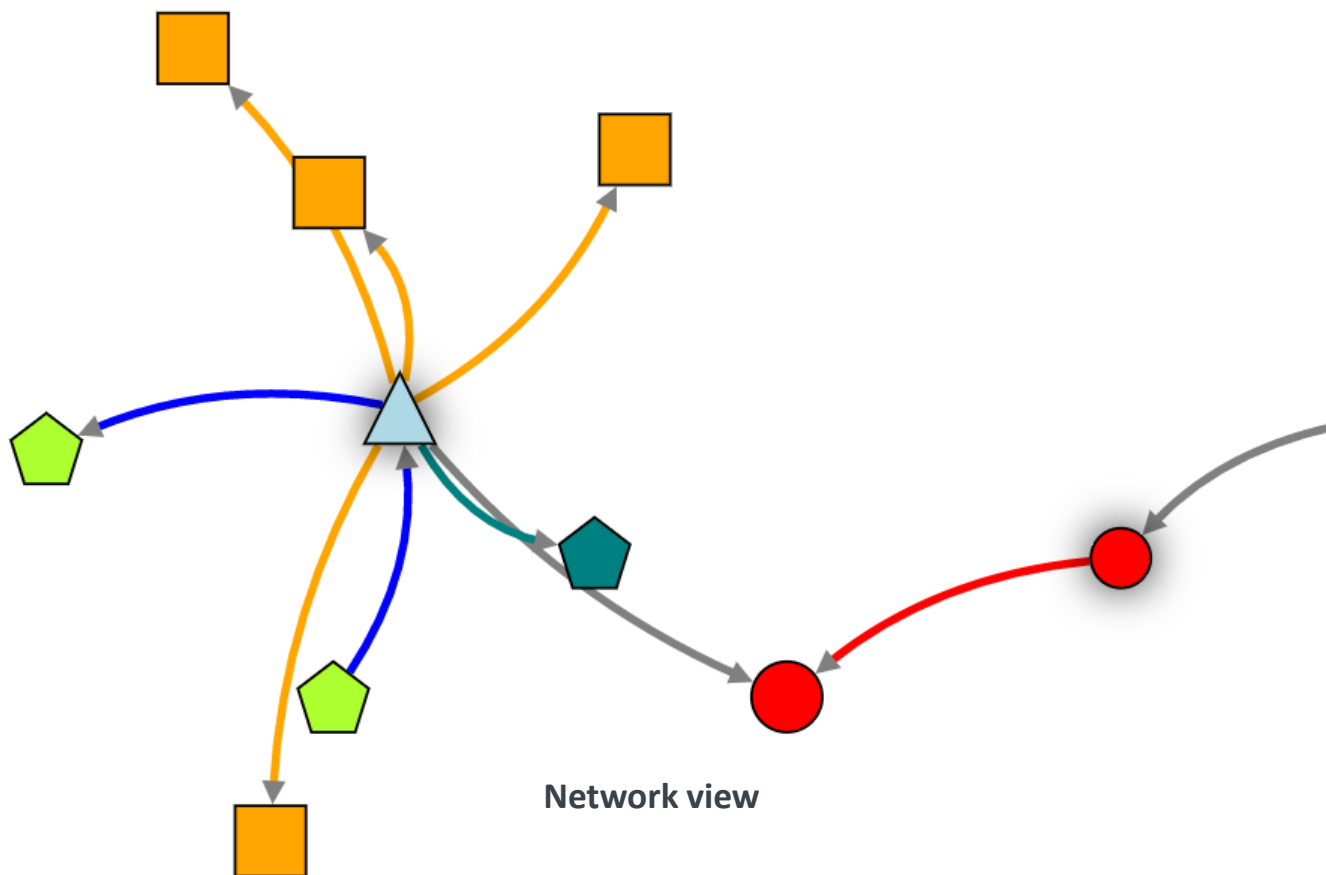
Select gene(s) and click "Show Network" button to see the Oindex network.

?

Max number of genes to show: 100

Known targets: ☐ Novel targets: ☐

ACCESSION	CHRO	START	SCORE	USER	QTL	EVIDENCE	Select
<a href="#">TRAES_4DL_6063F821A</a>	4D	8.36	2.20	no	0		<input type="checkbox"/>
<a href="#">TRAES_4BL_46BE39DD3</a>	4B	74.63	2.20	no	0		<input type="checkbox"/>
<a href="#">TRAES_4AS_13CE9F8B7</a>	4A	19.89	2.20	no	0		<input type="checkbox"/>
<a href="#">TRAES_7BL_57F31555F</a>	7B	167.18	1.83	no	0		<input type="checkbox"/>
<a href="#">TRAES_5AL_5D65F5235</a>	5A	2.57	1.74	no	0		<input type="checkbox"/>
<a href="#">TRAES_3B_2A5C09288</a>	3B	68.30	1.61	no	0		<input type="checkbox"/>
<a href="#">TRAES_2DL_574F25C6E</a>	2D	76.53	1.61	no	0		<input type="checkbox"/>
<a href="#">TRAES_7BS_9F5A66A59</a>	7B	66.45	1.61	no	0		<input type="checkbox"/>
<a href="#">TRAES_6AS_C1F384744</a>	6A	0.00	1.51	no	0		<input type="checkbox"/>
<a href="#">TRAES_6BS_04D7C6A64</a>	6B	65.07	1.51	no	0		<input type="checkbox"/>
<a href="#">TRAES_7DS_32CBC58AD</a>	7D	0.00	1.51	no	0		<input type="checkbox"/>



## Search > Explore

RelFinder

URL

between

examples

- (1)
- (2)
- (3)

add

clear

Find Relations

Filter by:

relations: (57/57)

☒ length ☒ class ☒ link ☒ connect...

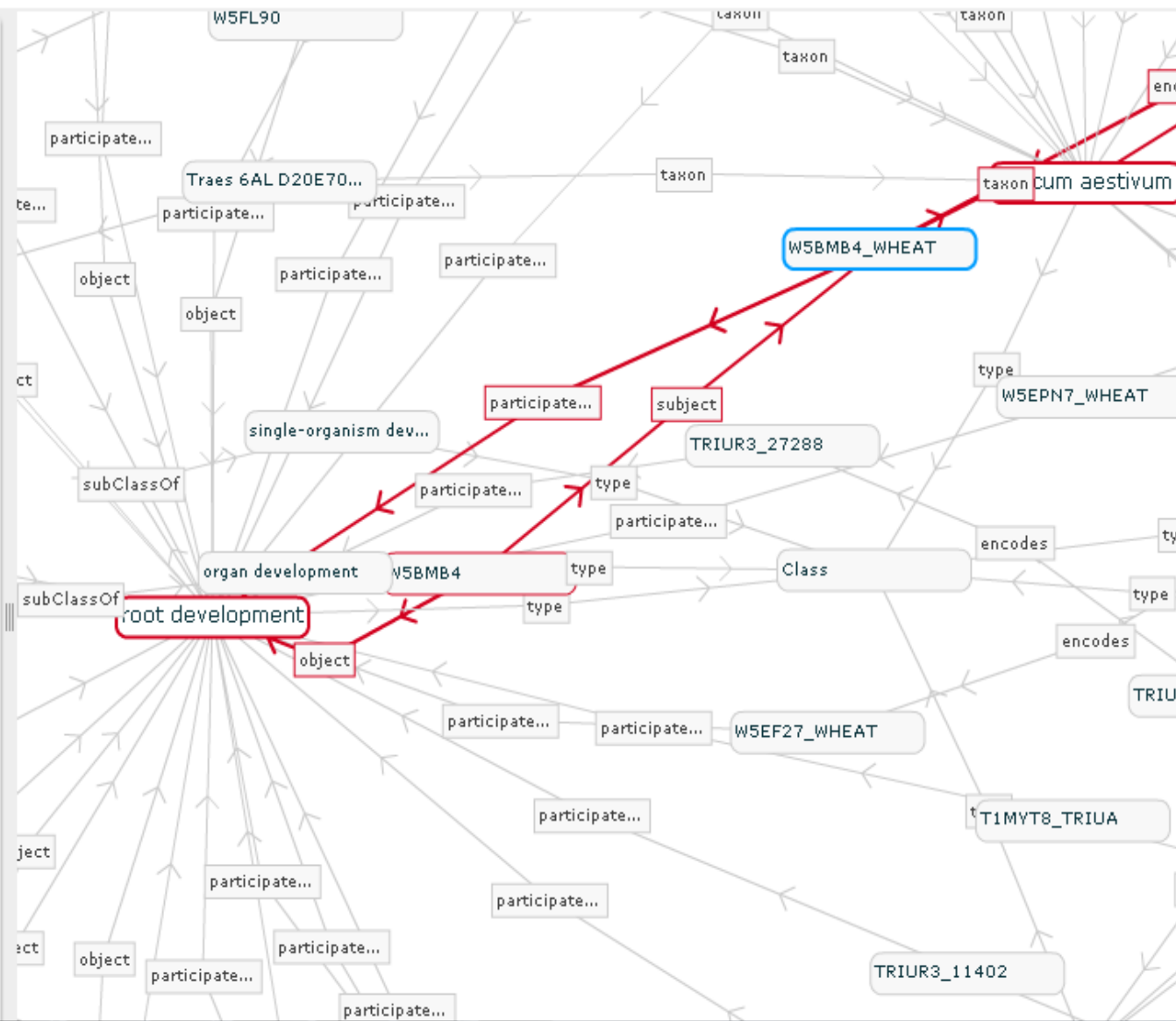
number of objects	num	vi
1	15/15	<input checked="" type="checkbox"/>
2	42/42	<input checked="" type="checkbox"/>

W5BMB4\_WHEAT

en

More Infos: [purl.uniprot.org](http://purl.uniprot.org)

Uncharacterized protein




The more the wheat research community harmonize its practices in terms of data management, the more IS and tools like the WheatIS, QTLNetMiner and AgroLD can integrate data and provide valuable knowledge

## How to adopt the guidelines?

- For legacy data
  - Please provide your data in at least one of the recommended data formats even if, for some reasons, you need to also keep them in other non-recommended formats
- For future developments
  - Please consider using the recommended data formats from the beginning.
- Example: provide your sequence variation data in the latest VCF file format
  - Please refer to the [WDI guidelines](#) for precise recommendations on each data type

- Describe your data with the recommended metadata standards and annotate your data with the recommended vocabularies.
  - Examples:
    - For genome annotation data in GFF3 format, use of ontologies for functional annotation in column 9, such as, Gene Ontology and Sequence Ontology.
    - For observation Variables (including trait and environment variables), use existing variables, listed in the following vocabularies and ontologies :
      - [Wheat crop ontology](#)
      - Wheat INRA Phenotype Ontology (previously INRA Wheat Ontology)
      - [Biorefinery ontology](#)
      - [XEO, XEML Environment Ontology](#)
- Deposit your data in the WheatIS data repository or well established data repositories

<https://urgi.versailles.inra.fr/dspace/>



The screenshot shows the WheatIS data repository interface. At the top, there's a banner with the text "WheatIS data repository" and a "Login" button. Below the banner, the text "DSpace Home" is visible. The main heading is "WheatIS data repository". A paragraph describes the space as a digital service for collecting, preserving, and distributing public data related to wheat scientific research. It mentions that users can consult and download submitted data anonymously and need to register to submit datasets. A list of links includes "Submit Phenotyping data", "Submit Genotyping data", and "Submit SNP Discovery data". Below this, it says "You may look at [Wheat Data Interoperability Guidelines](#) for recommendations on how and what to submit." and "Or you can just discover what has already been submitted using browsing features below and in the right menu." The section "Communities in DSpace" lists "Wheat Community". On the right side, there's a "Search DSpace" section with a search box and a "Go" button, followed by a link to "Advanced Search". Below that is a "Browse" section with links for "All of DSpace", "Communities & Collections", "By Issue Date", "Authors", and "Titles". At the bottom right, there's a "My Account" section with links for "Login" and "Register".

- Share your wheat related ontologies within the [WDI slice in Agroportal](#)
- Before developing a new ontology
  - Make sure there is not an existing one within the WDI slice in Agroportal that covers your needs
- When developing a new ontology
  - Please reuse or link to exiting concepts and terms in the ontologies within the [WDI slice in Agroportal](#) whenever possible.
  - Please align your ontologies to the existing ones within the [WDI slice in Agroportal](#) and share the mapping results



# Endorsements/Adopters

25

Laboratory	Contact
NIAB,	Professor Mario Caccamo Head of Crop Bioinformatics
USDA ARS and Cold Spring Harbor Laboratory,	Doreen Ware Adjunct Associate Professor Ph.D., Ohio State University
Paul Kersey EMBL European Bioinformatics Institute,	Paul Kersey Team Leader Non-vertebrate Genomics
Australian Center for Plant Functional Genomics,	Dr Baumann, Ute Bioinformatics Leader
The Genome Analysis Center,	Robert Davey Data Infrastructure & Algorithms Group Leader
Munich Information Center for Protein Sequences (MIPS), Helmholtz Center Munich,	Dr. Klaus Mayer Research Director MIPS
INRA URGI,	Michael Alaux, Deputy leader of "Information System and data integration" team Cyril Pommier, Deputy leader, Information System and Data integration team, Phenotype thematic leader
Rothamsted Research,	Christopher Rawlings Head of Department Computational & Systems Biology Harpenden
James Hutton Institute,	David Marshall Information and Computational Sciences The James Hutton Institute
CIMMYT Wheat program,	Richard Allan James, Head of Knowledge Management Rosemary Shrestha, Data Coordinator

**WDI WG members:** Fulss Richard, co-chair (CIMMYT), Alaux Michael (INRA), Aubin Sophie (INRA), Arnaud Elizabeth (Bioversity), Baumann Ute (Adelaide University), Buche Patrice (INRA), Cooper Laurel (Planteome), Hologne Odile (INRA), Laporte Marie-Angélique (Bioversity), Larmande Pierre (IRD), Letellier Thomas (INRA), Mohellibi Nacer (INRA) Pommier Cyril (INRA), Protonotarios Vassilis (Agro-Know), Shrestha Rosemary (CIMMYT), Subirats Imma (FAO of the United Nations), Aravind Venkatesan (IBC), Whan Alex (CSIRO), Jonquet Clément (Lirmm, Agroportal)

And

Lucas Hélène (INRA, International Wheat Initiative), Quesneville Hadi (INRA, chair WheatIS EWG), Chris Rawlings (Rothamsted Research, QTLNetMiner)

**Thank you!**