RESEARCH DATA ALLIANCE

Wheat Data Interoperability WG outputs

research data sharing without barriers
rd-alliance.org

The Wheat Initiative

- 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 problem

Top management or breeders

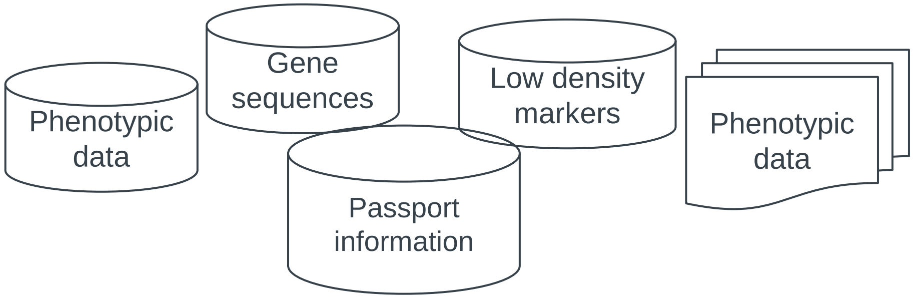
- What are the sources of resistance to stem rust (UG99) and tolerance to drought conditions in bread wheat?

Data are

Dispersed

Heretogeneous

Abundant



Phenotypic data

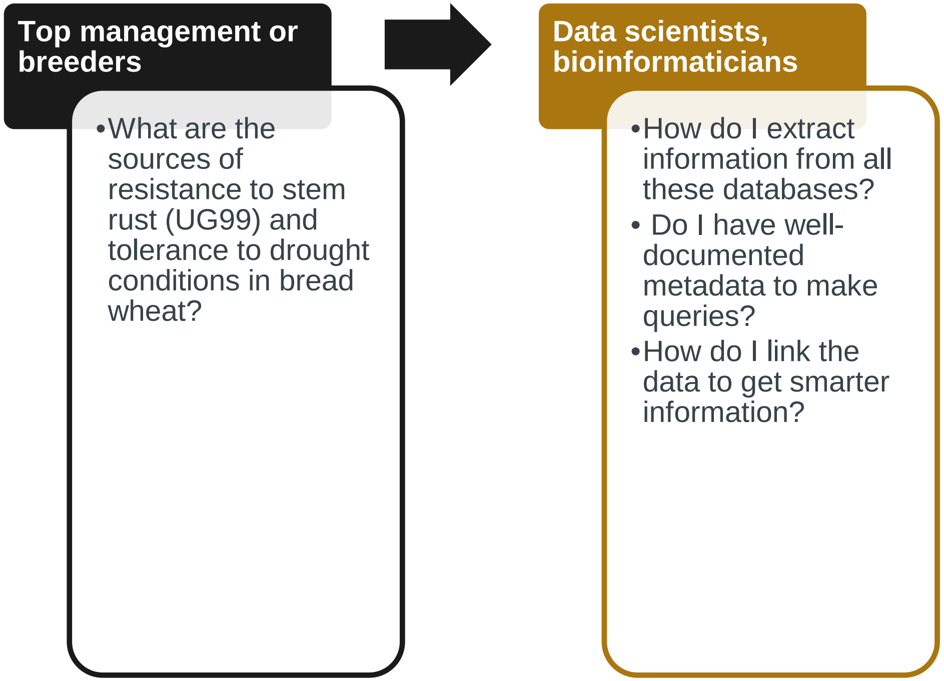
Gene sequences

Low density markers

Phenotypic data

Passport information

The problem



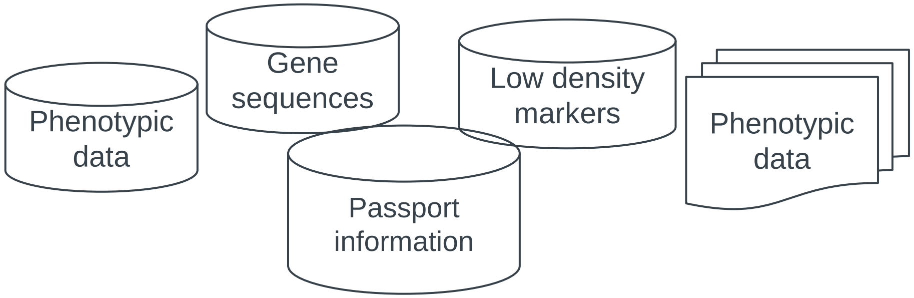
Top management or breeders

- What are the sources of resistance to stem rust (UG99) and tolerance to drought conditions in bread wheat?

Data scientists, bioinformaticians

- How do I extract information from all these databases?
- Do I have well-documented metadata to make queries?
- How do I link the data to get smarter information?

Data are
Dispersed
Heretogeneous
Abundant



Phenotypic data

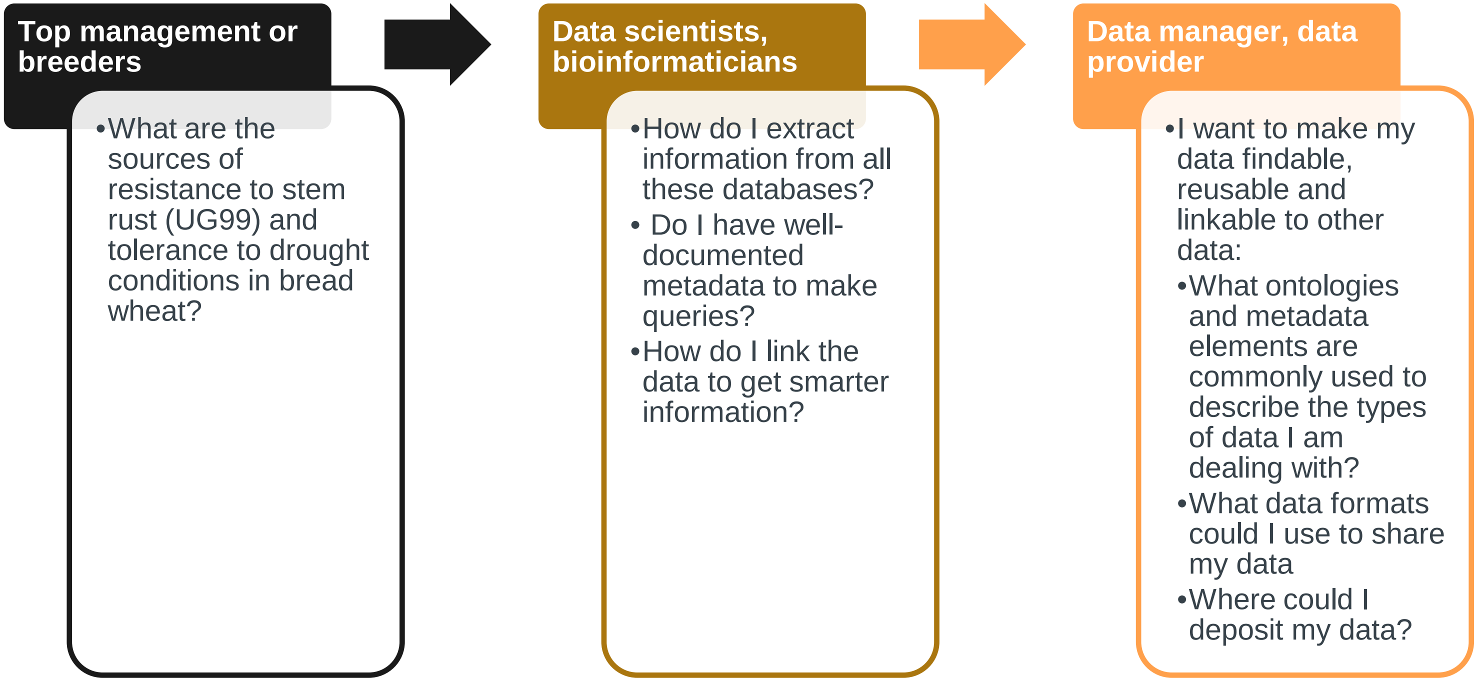
Gene sequences

Low density markers

Passport information

Phenotypic data

The problem



Top management or breeders

- What are the sources of resistance to stem rust (UG99) and tolerance to drought conditions in bread wheat?

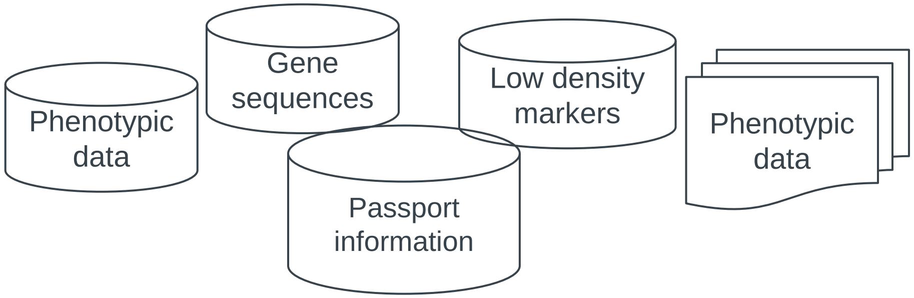
Data scientists, bioinformaticians

- How do I extract information from all these databases?
- Do I have well-documented metadata to make queries?
- How do I link the data to get smarter information?

Data manager, data provider

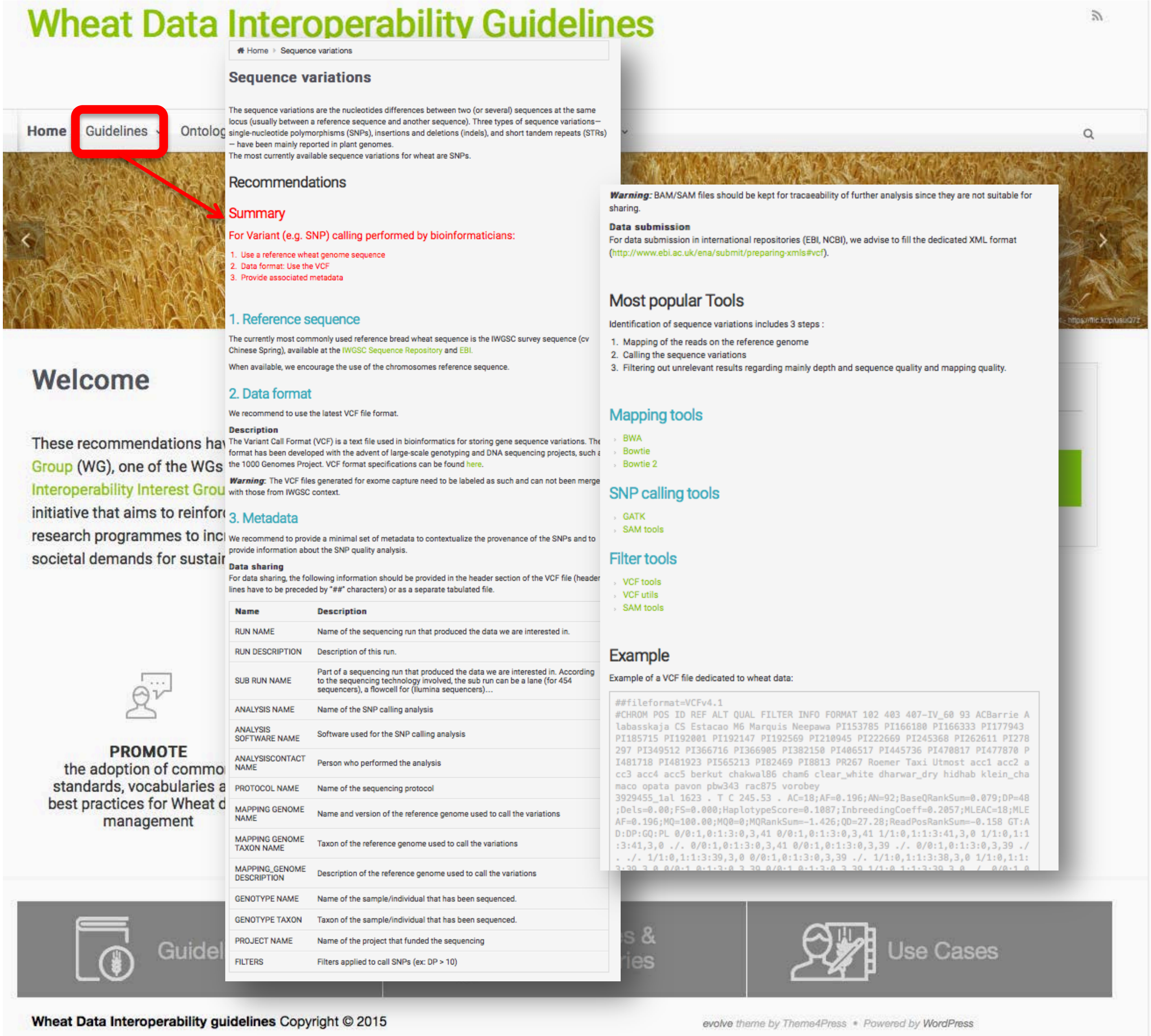
- I want to make my data findable, reusable and linkable to other data:
  - What ontologies and metadata elements are commonly used to describe the types of data I am dealing with?
  - What data formats could I use to share my data
  - Where could I deposit my data?

Data are
Dispersed
Heretogeneous
Abundant



The deliverables

- 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 portal 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://volvestre.cirad.fr:8080/agrold/



Wheat Data Interoperability Guidelines

Welcome

These recommendations ha[v]
Group (WG), one of the WGs
Interoperability Interest Grou
initiative that aims to reinfor[c]
research programmes to inc[r]
societal demands for sustai[n]

PROMOTE
the adoption of commo[n]
standards, vocabularies a[nd]
best practices for Wheat d[ata]
management

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 a[s] 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 been merge[d] 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 tracaeability 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 unrelevant 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 chakwal86 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
;Dels=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
```

Wheat Data Interoperability guidelines Copyright © 2015

Benefits for many target users

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

Benefits for many target users

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

Benefits for many target users

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.

How You Can Endorse the guidelines on data formats

- 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

How You Can Endorse the data description and sharing best practices

- Describe your data following the WDI recommendations and 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
    - INRA Wheat Ontology
    - Biorefinery ontology
    - XEO, XEML Environment Ontology

How You Can Endorse the WDI wheat related ontologies portal?

- 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.
- Whenever possible
  - Please align your ontologies to the existing ones within the WDI slice in Agroportal and share the mapping results

Endorsements/Adopters

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

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 Dzalé Yeumo Kaboré (INRA, France), Richard Allan Fulss (CIMMYT, Mexico)

Thank you!