WHEAT DATA INTEROPERABILITY GUIDELINES

Authors: Dzalé Yeumo Esther[1], Fulss Richard[2], Alaux Michael[1], Aubin Sophie[1], Arnaud Elizabeth[3], Baumann Ute[4], Cooper Laurel[5], Davey Robert P.[12], Laporte Marie-Angélique[3], Larmande Pierre[6], Letellier Thomas[1], Pommier Cyril[1], Protonotarios Vassilis[7], Shrestha Rosemary[2], Subirats Imma[8], Venkatesan Aravind[6], Whan Alex[9], Wheat Data Interoperability working group[11]

Contributors: Lucas Hélène [1], Quesneville Hadi[1], Hologne Odile[1], Keizer Johannes[8], Mohellibi Nacer[1], Buche Patrice[1], Jonquet Clément[10], Caracciolo Caterina[8], Reverte Carme[8]

Affiliations: [1] L'Institut National de la Recherche Agronomique (INRA), [2]International Maize and Wheat Improvement Center (CIMMYT), [3]Bioversity International, [4]Australian Centre for Plant Functional Genomics (ACPFG), [5]Planteome, [6] L'Institut de Recherche pour le Développement, [7]Agroknow, [8] Food and Agriculture Organization (FAO), [9]Commonwealth Scientific and Industrial Research Organisation (CSIRO), [10]Laboratoire d'Informatique, de Robotique et de Microélectronique (LIRMM), [11]Research Data Alliance (RDA), [12]Earlham Institute (EI)

ABSTRACT

The International Wheat Initiative has identified the easy access and interoperability of all wheat related data as a top priority for the wheat research community. An important goal is to make the best use of existing genetic, genomic, and phenotypic data in fundamental and applied wheat science. Hence, data interoperability has become a priority in this community, given the ever-growing data deluge coming from improvements in technologies and numeric methods for DNA (Deoxyribonucleic acid) and RNA (Ribonucleic acid) sequencing, high throughput genotyping and phenotyping, high throughput imaging and satellite monitoring.

The Wheat Data Interoperability working group (WDI-WG) was created in 2014 as a Research Data Alliance (RDA) working group under the umbrella of Wheat Information System (WheatIS) expert working group, which is endorsed by the International Wheat Initiative to build an international information system for genetic, genomic, and phenomic data. The primary mission of the WDI-WG is to provide a common framework for describing and representing data with respect to existing open standards. The WDI-WG is formed by data and information practitioners and scientists from different organizations and countries. The standpoint of the WDI-WG was to avoid the creation of new standards. Therefore, the WDI-WG started by surveying the practices of the wheat research community, then converged and agreed on specific recommendations. The proposed recommendations have been endorsed by the RDA and the WheatIS expert working group.

SUMMARY OF THE GUIDELINES

➔ Please visit the website for the full and updated version of the guidelines.

Data types	Data exchange formats	Metadata	Ontologies, vocabularies	Additional recommendations
DNA	The Variant Call Format (VCF).	Provide associated metadata to		The currently most commonly used

sequence variations		contextualize the provenance of the SNPs and provide information about the SNP quality analysis.		reference bread wheat sequence is the IWGSC (International Wheat Genome Sequencing Consortium) survey sequence (cv Chinese Spring), which is available at the IWGSC Sequence Repository and The European Bioinformatics Institute (EBI). When available, we encourage the use of the chromosome reference sequence.
Genome annotations	The General Feature Format (GFF3).		Use ontologies (Gene Ontology, Sequence Ontology) for functional annotation in column 9 in the GFF3 file.	Provide comprehensive content description for column 9 in the GFF3 file. Use external database cross references (Dbxref) in a consistent way.
Gene sequences	Use the existing format standards laid out by the repositories such as The National Center for Biotechnology Information (NCBI) (Gene Expression Omnibus, GEO), EBI Array Express, and The European Nucleotide Archive (ENA).	Metadata is important for all gene expression studies – whether microarray or RNA Seq data. • Microarray data: NCBI-GEO-MIAMI compliant • RNA-Seq data – follow the guidelines for Minimum Information About a Microarray Experiment (MIAME) as far as applicable (FastQ) For BAM files – additional info needed: • mapping software • mismatch settings • reference sequences used such as IWGSC survey sequences or MIPS (The Munich Information Center for Protein Sequences) gene models, or transcriptome assembly	Use ontologies and controlled vocabularies to annotate the required metadata • Plant Ontology (PO) terms to describe the plant tissues and developmental stage. • Plant Environment Ontology (EO) terms to describe the experimental conditions. • Plant Stress Ontology (PSO) terms to describe the treatments with pathogens, stress conditions (proposed). • Gene Ontology (GO) is the standard for the functional analysis. • Microarray and Gene Expression Data Ontology (MGED) Ontology (MO) terms mapped to The Ontology for Biomedical Investigations (OBI) /The Open Biomedical Ontologies (OBO) foundry ontology terms.	
Phenotype	Use data format that follows	Provide complete metadata for at least	Use following ontologies and controlled	We recommend sharing at least clean

data	minimum format principles with data matrices plus metadata for at least variables (trait along with method, units and scales or environmental ones) and germplasms. ISA-Tab, which is complaint to the Minimum Information About a Plant Phenotyping Experiment (MIAPPE) standard, is an implementation of this principle.	germplasm and observation variables.	vocabularies to annotate nurseries and trials data: • Crop Research Ontology (CO_715) to describe the metadata related to nurseries and trials, field management, field environments, study design, etc. These metadata are actively adapted for a wider use. • Wheat Trait Ontology (CO_321) describes the variables that are most commonly used in wheat phenotyping. • Biorefinery Ontology describes the concepts and terms associated with biomass composition and characterization (crystallinity, surface area, particle size, porosity, etc.), physico-chemical pretreatments, enzymatic hydrolysis, and experimental processes descriptions.	documented raw data, like plant height, leaf area, etc.
Germplasm data	Use data matrices in csv (comma-separated values) or Excel. For examples: GnpIS, Breeding Management System (BMS).	Use Multi-Crop Passport Descriptor (MCPD) as metadata format for passport and other variables data required for characterization of germplasm	Wheat Trait Ontology (CO_321) to describe germplasm characterization	Identification of germplasm must be non-ambiguous at the scale of the world wide web. To achieve this, consider using URI and mainly follow the work of the DivSeek project.
Physical maps	FPC format for physical map raw data GFF3 format for data integration			**Data submission**: INRA URGI is the official repository for IWGSC physical maps. We recommend using the data submission support at http://wheat-urgi.versailles.inra.fr/Seq-Repository/Support-to-assembly-and-data-submission

➔ In addition, the WDI-WG uses AgroPortal to maintain the list of the ontologies and vocabularies recommended by the working group. The WDI-WG encourages the developers of wheat related ontologies and vocabularies to share new resources through the same portal in order to facilitate their findability and accessibility by the community. AgroPortal allows users to search within ontologies, retrieve existing mappings between ontologies, visualize ontology content, comment about ontologies, concepts and mappings, annotate text data with ontology concepts, get a recommendation about which ontology to use for a text corpus or set of keywords, and store projects which use ontologies.

MAINTENANCE OF THE GUIDELINES

The RDA WDI working group has now become a maintenance group since June 2017. The working group will maintain the guidelines in collaboration with the WheatIS expert working group and a steering group including representatives of the adopters. The guidelines will be revised twice (frequently) a year to consider the evolving landscape of data standards and practices.

GETTING INVOLVED

There are many ways you can be part of this work:

- ➔ Give us your feedback: you can leave a comment on a specific page of the guidelines website in order to propose improvements or corrections and share best practices or useful tools.
- ➔ Join the maintenance group: the maintenance group is open and all contributions are welcome. You can freely join us at https://www.rd-alliance.org/groups/wheat-data-interoperability-wg.html
- ➔ Become an adopter: Please let us know if you wish to adopt the guidelines by contacting us at rda-wdinterop-wg@rda-groups.org

List of early adopters

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