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], Laporte Marie-Angélique[3], Larmande Pierre[6], Letellier Thomas[1], Pommier Cyril[1], Protonotarios Vassilis[7], Shrestha Rosemary[2], Subirats Imma[8], Aravind Venkatesan[6], Whan Alex[9]
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]

Affilliations: INRA (1), Cimmyt (2), Bioversity (3), ACPFG (4), Planteome (5), IRD (6), Agroknow (7), FAO (8), CSIRO (9), LIRMM (10)

ABSTRACT

The Wheat Data Interoperability Guidelines describe a set of guiding rules to foster wheat data interoperability, with the purpose of helping researchers create, manage and exchange wheat data. These recommendations have been prepared by members of the Wheat Data Interoperability Working Group (WG), one of the WGs of the Research Data Alliance and the only WG of the Agriculture Data Interoperability Interest Group, in coordination with the Wheat Initiative, a global initiative that aims to reinforce synergies between bread and durum wheat national and international research programs, to increase food security, nutritional value and safety while taking into account societal demands for sustainable and resilient agricultural production systems.

THE WHEAT DATA DESCRIPTION AND REPRESENTATION GUIDELINES

The WDI working group produced a set of guidelines for wheat data description and representations. These guidelines are available at http://datastandards.wheatis.org and cover the following data types: sequence variations, genome annotations, phenotypes, physical maps, germplasms, and gene expression. Each of these data types appears as a submenu under the “Guidelines” menu in the website. Under each submenu you will find information on the best practices, tools, recommendations and examples to create, manage and share data related to Wheat. Examples are given at the end of each subsection. The guidelines website intends to be a one stop shop for relevant information related to wheat data management in order to raise awareness, avoid duplicated efforts, and foster adoption of common practices.

THE WHEAT-RELATED VOCABULARIES AND ONTOLOGIES PORTAL

In the context of research data, the use of vocabularies plays a key role in managing, sharing and publishing data. Simple words can take different meanings to different people and standards definition for these words is key to avoid miscommunication and enable good collaboration. Vocabularies enhance the quality of the interoperability and effectiveness of data exchange, thus facilitating the reusage of data by others and in the process adding value to the local researcher. The WDI is maintaining a list of vocabularies and ontologies relevant to wheat data description within Agroportal (http://wheat.agroportal.lirmm.fr/ontologies). AgroPortal is an ontology repository dedicated to the agronomic and plant domains that reuses the NCBO

BioPortal infrastructure. The platform enables--among other functionalities--to store ontology metadata, search within ontologies, store and retrieve 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. The content of the platform is accessible via REST API and SPARQL endpoint: http://data.agroportal.lirmm.fr/documentation and http://sparql.agroportal.lirmm.fr/test/

MAINTENANCE

To stay relevant and useful for the wheat data community, it is important to maintain the guidelines. This will be especially true as new technologies are developed requiring standards to be agreed. The maintenance of the guidelines will depend on an ongoing process of monitoring practices across the wheat community, and maintaining communication with those who are producing and using wheat data. This maintenance will be done by the WheatIS (Wheat Information System: http://wheatis.org/) within the framework of the International Wheat Initiative (http://www.wheatinitiative.org/).

GETTING INVOLVED

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

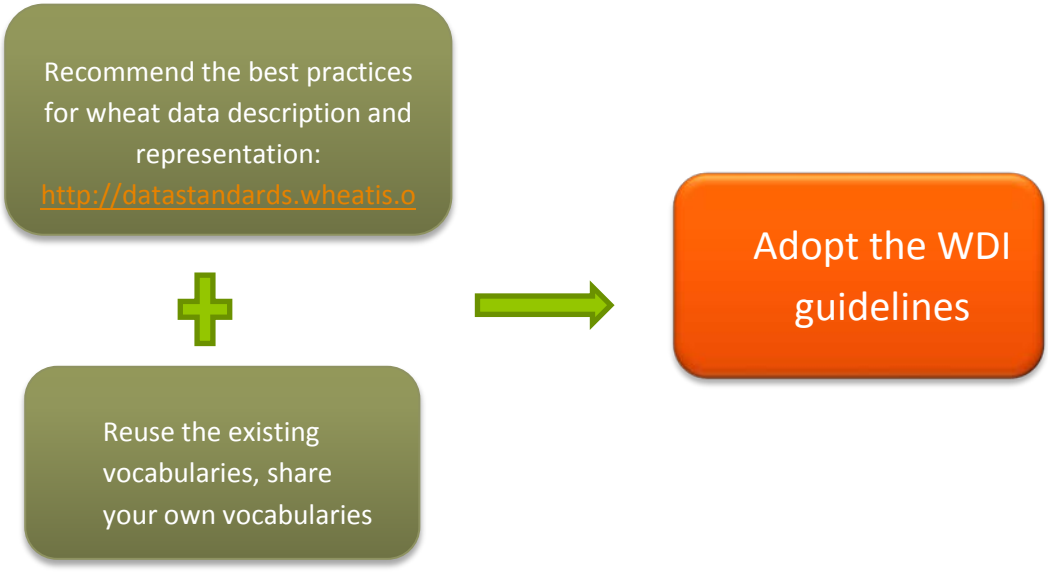
GIVE YOUR FEEDBACK

You can contact us either by email or by leaving a comment on a specific page 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 join us to help keep the guidelines accurate and up-to-date with the latest standards and practices in wheat research. Please send us an email: urgi-contact@versailles.inra.fr

BECOME AN ADOPTER



You can adopt the guidelines (http://datastandards.wheatis.org) by recommending the best practices they describe:

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

- Use a reference wheat genome sequence
- Data format: use the VCF
- Provide associated metadata

For genome annotations

- Use GFF3 as data format.
- Provide comprehensive content description for column 9 in the GFF3 file
- Consistent use of external database cross references (Dbxref)
- Use ontologies for functional annotation in column 9, such as, Gene Ontology and Sequence Ontology

For phenotypes

- Use data format that follows 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 is an implementation of this principle.
- Use complete metadata for at least germplasm and observation variables
- Keep curated data (checked outliers)

For germplasm data

- Use data matrices in csv or Excel (For example GnpIS, BMS)
- Use MCPD as metadata format for passport and other variables data required for characterization of germplasm

For gene expression

- Use the existing format standards laid out by the repositories such as NCBI (GEO) and EBI Array Express + ENA
- Use ontologies and controlled vocabularies to annotate the required metadata
    - Plant Ontology terms to describe the plant tissues and developmental stage
    - Plant Environment Ontology to describe the experimental conditions
    - Plant Stress Ontology to describe the treatments with pathogens, stress conditions (proposed)
    - Gene Ontology is the standard for the functional analysis
    - Microarray ontology (MO) terms mapped to the OBI/OBO foundry ontology terms – MGED ontology (http://bioportal.bioontology.org/ontologies/MO?p=classes)

For physical maps

- Use the FPC format for physical map raw data
- Use the GFF3 format for data integration

You can adopt the wheat vocabularies portal (http://wheat.agroportal.lirmm.fr/ontologies) by:

- Sharing your own vocabularies through the portal
- Aligning your own vocabularies with the existing ones
- Reusing the existing vocabularies within your information systems

List of current adopters

Organization	Contact
NIAB, www.niab.com	Professor Mario Caccamo Head of Crop Bioinformatics
USDA ARS and Cold Spring Harbor Laboratory, http://cshl.edu/	Doreen Ware Adjunct Associate Professor Ph.D., Ohio State University
Paul Kersey EMBL European Bioinformatics Institute, http://www.ebi.ac.uk/	Paul Kersey Team Leader Non-vertebrate Genomics
Australian Center for Plant Functional Genomics, http://www.acpfg.com.au/	Dr Baumann, Ute Bioinformatics Leader
The Genome Analysis Center, http://www.tgac.ac.uk/	Robert Davey Data Infrastructure & Algorithms Group Leader
Munich Information Center for Protein Sequences (MIPS), Helmholtz Center Munich, http://www.helmholtz-muenchen.de/	Dr. Klaus Mayer Research Director MIPS
INRA URGI, https://urgi.versailles.inra.fr/	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, http://www.rothamsted.ac.uk/	Christopher Rawlings Head of Department Computational & Systems Biology Harpenden
James Hutton Institute, http://www.hutton.ac.uk/	David Marshall Information and Computational Sciences The James Hutton Institute
CIMMYT Wheat program, http://www.cimmyt.org/en/	Richard Allan James, Head of Knowledge Management Rosemary Shrestha, Data Coordinator