Wheat data interoperability

Windpouire Esther Dzale Yeumo

To cite this version:

Windpouire Esther Dzale Yeumo. Wheat data interoperability. Transitionning Cereal Systems to Adapt Climate Change, Nov 2015, Minneapolis, United States. hal-02794093

Transitioning Cereal Systems to Adapt to Climate Change

November 13-14, 2015

Funded through Award #2011-68002-30191 from USDA National Institute of Food and Agriculture

Wheat Data Interoperability

Esther Dzalé Yeumo
Co-chair RDA Wheat Data Interoperability WG
Chair INRA competence center for data management and sharing services

WHEAT INITIATIVE

http://www.wheatinitiative.org/

An international research partnership for wheat improvement

- 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,

The WheatIS Expert Working Group

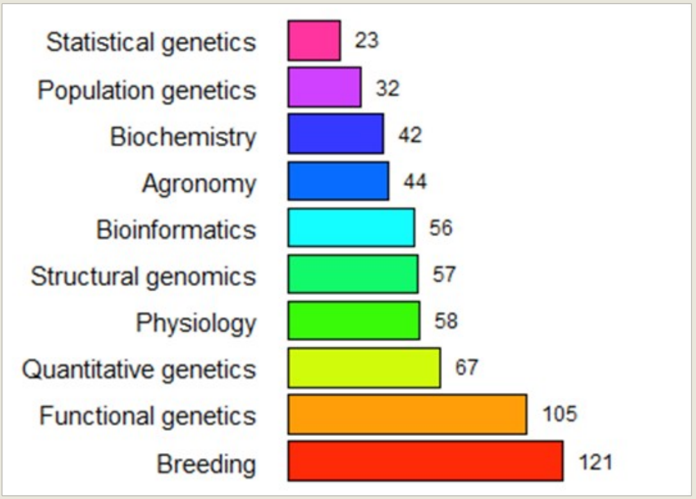
The WheatIS EWG goals

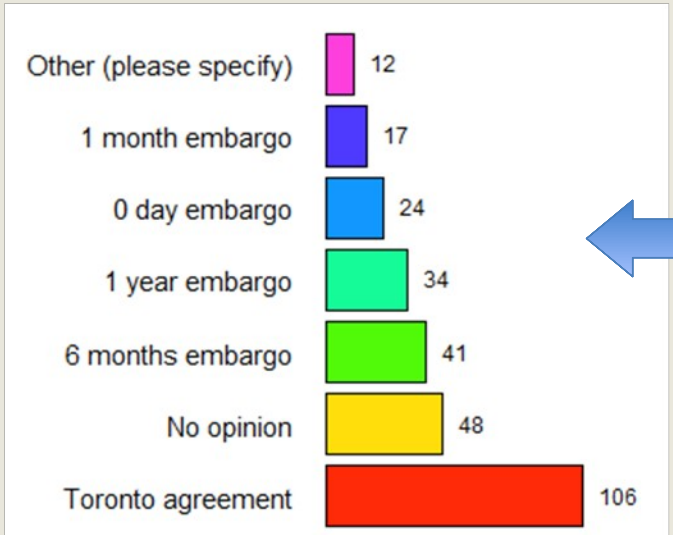
WheatIS Expert Worging Group

User survey

Full results at: http://ist.blogs.inra.fr/wdi/wp-content/uploads/sites/8/2015/06/wheat-info-system-report.pdf

Fields of expertise of the respondents



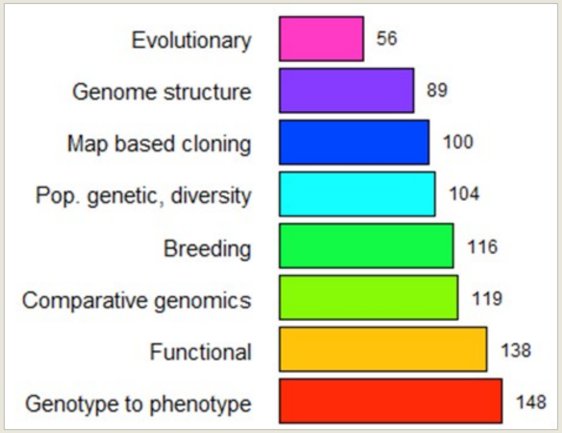


Most of the participants supported the data reuse policy promoted by the Bermuda/ Fort Lauderdale / Toronto agreements (Nature 461, 168F170, doi:10.1038/461168a), that promotes the early dissemination of whole genome datasets but preserves the rights for the data generators to lead the analysis and publication of their data in peer reviewed journals

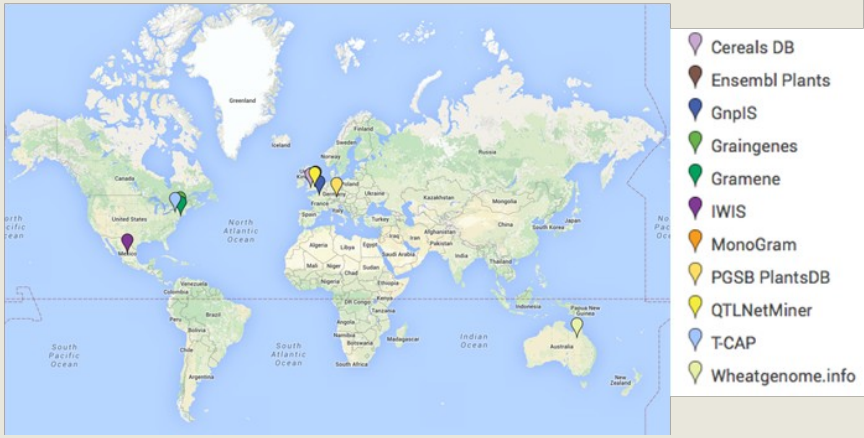
WheatIS Expert Worging Group

State of the art

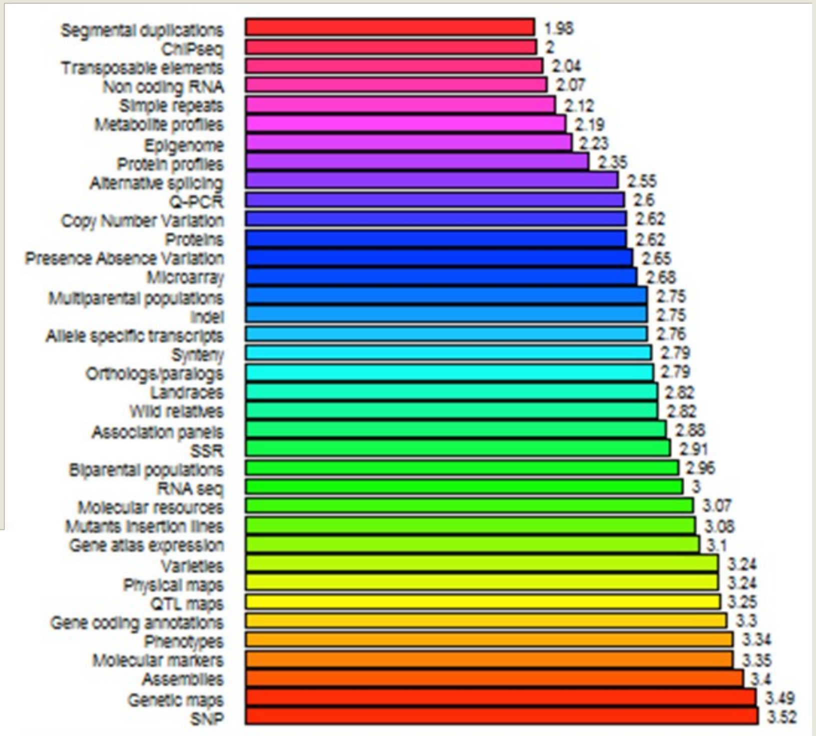
Studies



Repositories



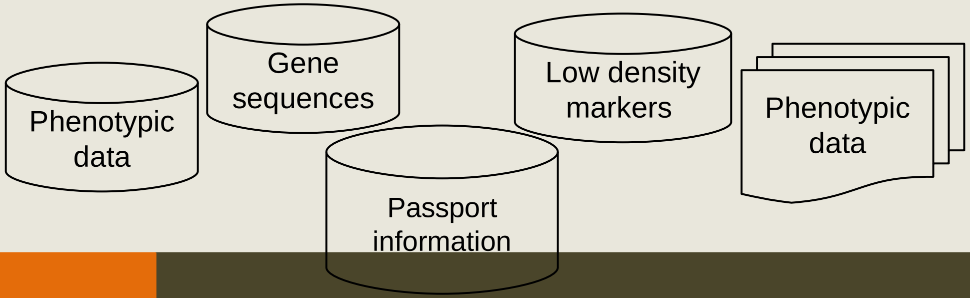
Data types



The interoperability challenge illustrated

Data are

- Dispersed
- Heretogeneous
- Abundant



Phenotypic data

Gene sequences

Passport information

Low density markers

Phenotypic data

RDA Research Data Sharing without barriers

RESEARCH DATA ALLIANCE

WHEAT INITIATIVE

The Wheat Data Interoperability WG

- Created in March 2014 within the frame of RDA
- 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 data

CSIRO

Food and Agriculture Organization of the United Nations

INRA SCIENCE & IMPACT

ACPFG

CIMMYT International Maize and Wheat Improvement Center

Bioversity International

Planteome

AGRO-KNOW

IRD Institut de recherche pour le développement

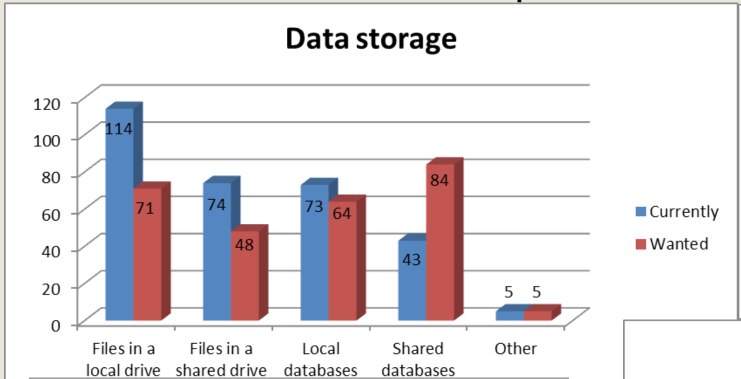
The achievements

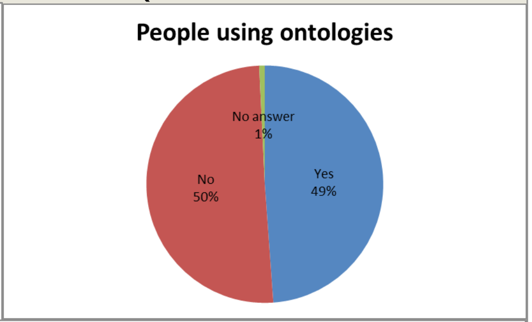
Data management practices survey

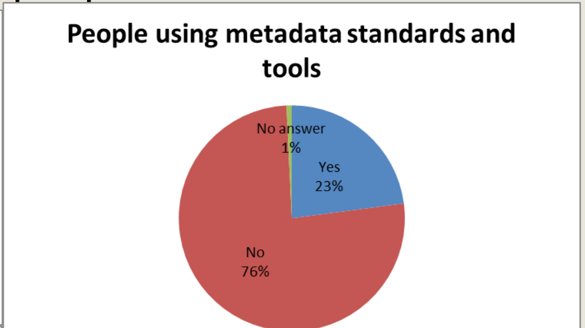
- Objective: identify
  - Data storage practices
  - Data management policy or guidelines in use
  - Data formats in use
  - Ontologies and vocabularies in use
- Complete results
  - http:// ist.blogs.inra.fr/wdi/wp-content/uploads/sites/8/2015/

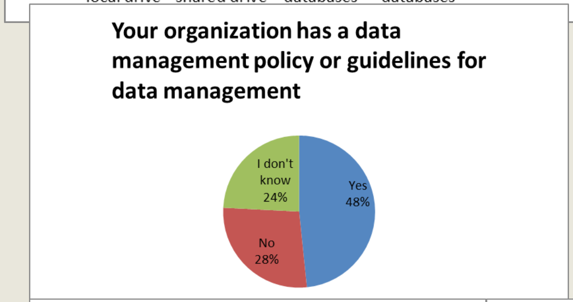
Data management practices survey

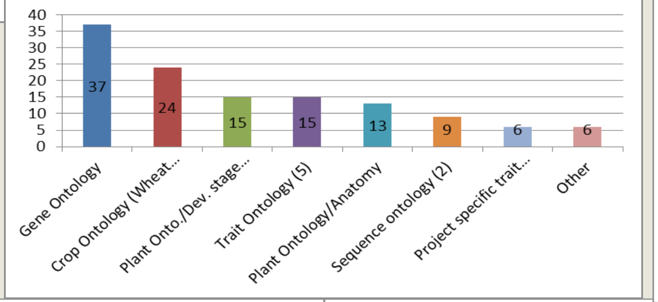
- Total number of answers: 201
- Number of complete answers: 125
- Total number of incomplete answers: 77 (6 doubles removed: people who answered

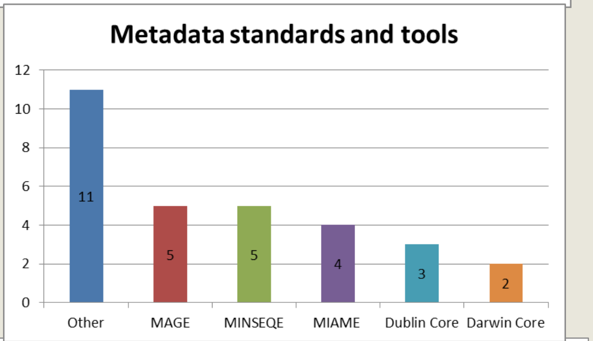


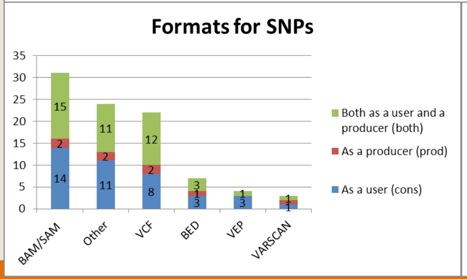


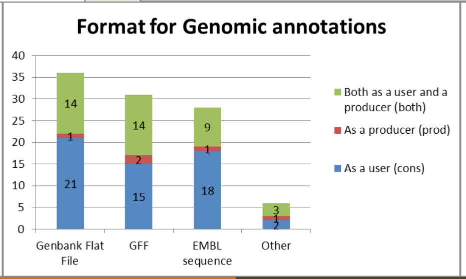


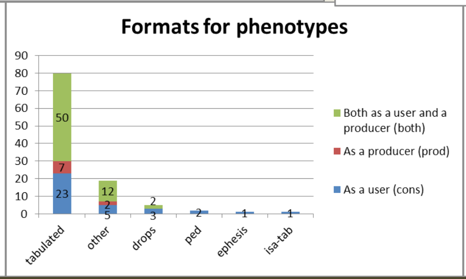


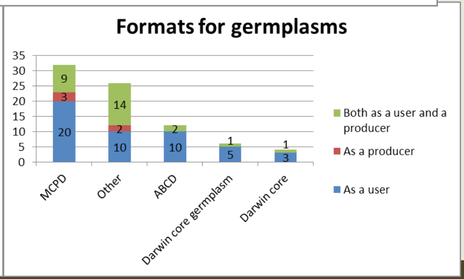








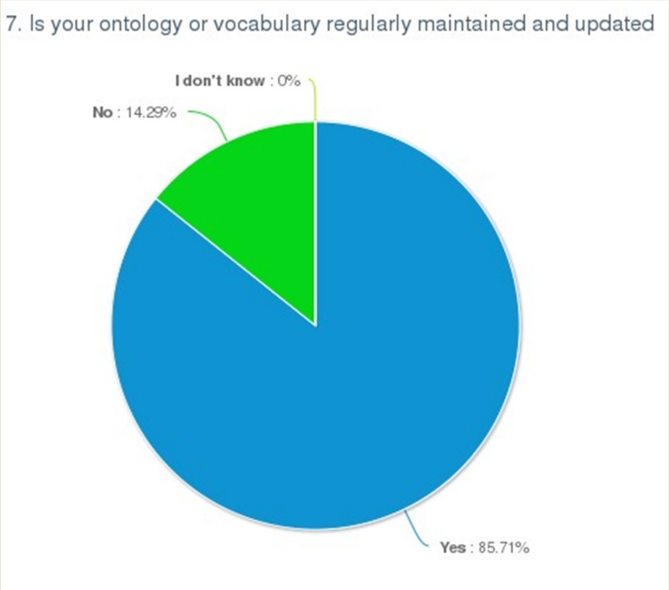


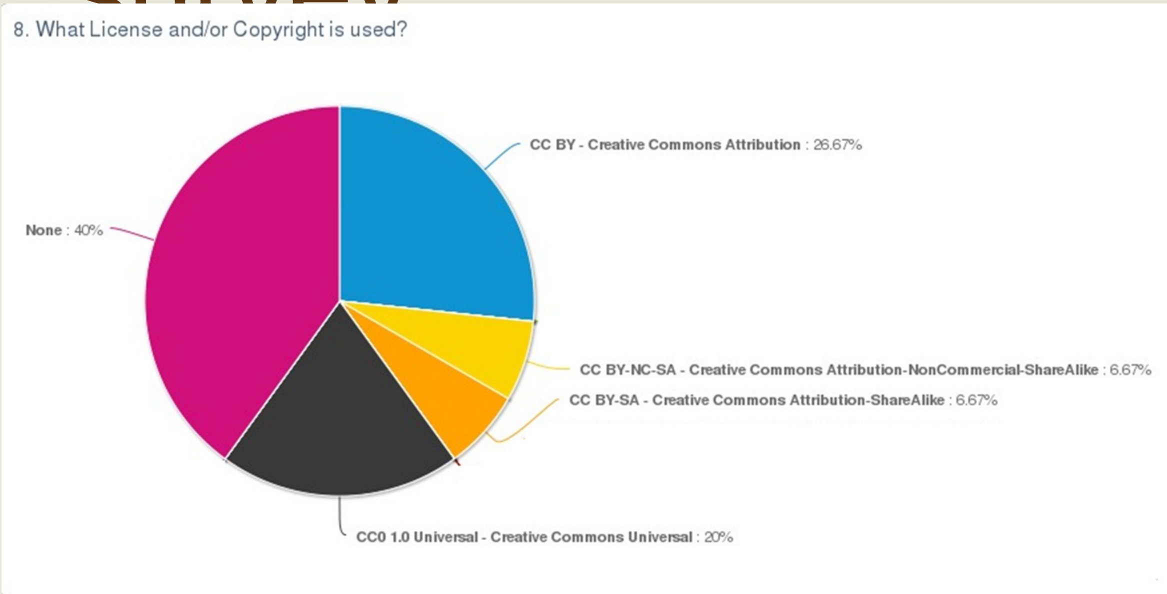


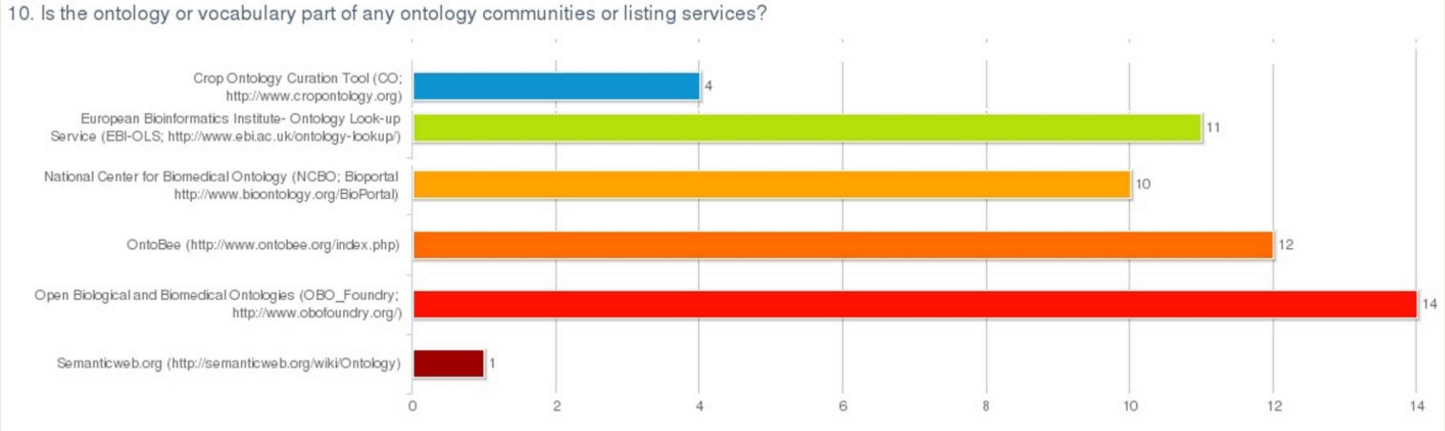
Ontologies & vocabularies survey

- Objective
  - Assess the level of visibility and interoperability of Wheat related vocabularies and ontologies
  - Identify the domain covered by the ontologies and vocabularies
  - Collect some technical details

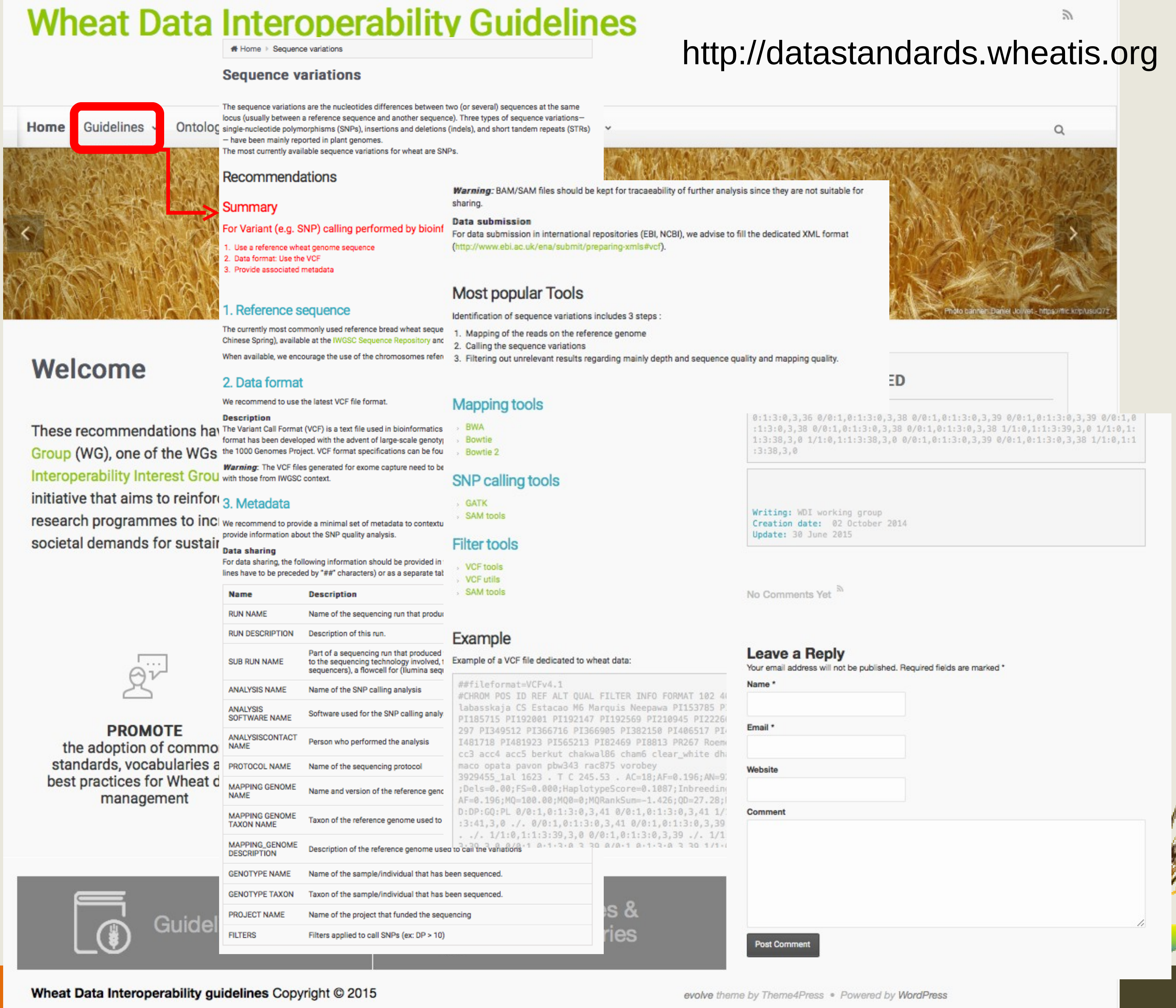
Ontologies & vocabularies survey


7. Is your ontology or vocabulary regularly maintained and updated


8. What License and/or Copyright is used?


10. Is the ontology or vocabulary part of any ontology communities or listing services?

Complete results: http://ist.blogs.inra.fr/wdi/wp-content/uploads/sites/8/2015/05/WDI-Ontologies-2015-03.pdf http://datastandards.wheatis.org


Wheat Data Interoperability Guidelines
Home
Guidelines
Ontolog
Welcome
These recommendations hav
Group (WG), one of the WGs
Interoperability Interest Grou
initiative that aims to reinfor
research programmes to inc
societal demands for sustair
PROMOTE
the adoption of commo
standards, vocabularies a
best practices for Wheat d
management
Guidel
es &
ries
ED
Photo banner: Daniel Jolivet - https://flic.kr/p/usuQ7z
No Comments Yet
Leave a Reply
Your email address will not be published. Required fields are marked *
Name *
Email *
Website
Comment
Post Comment
evolve theme by Theme4Press • Powered by WordPress

Home › Sequence variations

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 bioinf

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 seque Chinese Spring), available at the IWGSC Sequence Repository and

When available, we encourage the use of the chromosomes refer

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 format has been developed with the advent of large-scale genoty the 1000 Genomes Project. VCF format specifications can be fou

Warning: The VCF files generated for exome capture need to be with those from IWGSC context.

3. Metadata

We recommend to provide a minimal set of metadata to contextu provide information about the SNP quality analysis.

Data sharing
For data sharing, the following information should be provided in lines have to be preceded by "##" characters) or as a separate tab

Name	Description
RUN NAME	Name of the sequencing run that produ
RUN DESCRIPTION	Description of this run.
SUB RUN NAME	Part of a sequencing run that produced to the sequencing technology involved, sequencers), a flowcell for (Ilumina seq
ANALYSIS NAME	Name of the SNP calling analysis
ANALYSIS SOFTWARE NAME	Software used for the SNP calling analy
ANALYSISCONTACT NAME	Person who performed the analysis
PROTOCOL NAME	Name of the sequencing protocol
MAPPING GENOME NAME	Name and version of the reference genc
MAPPING GENOME TAXON NAME	Taxon of the reference genome used to
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:

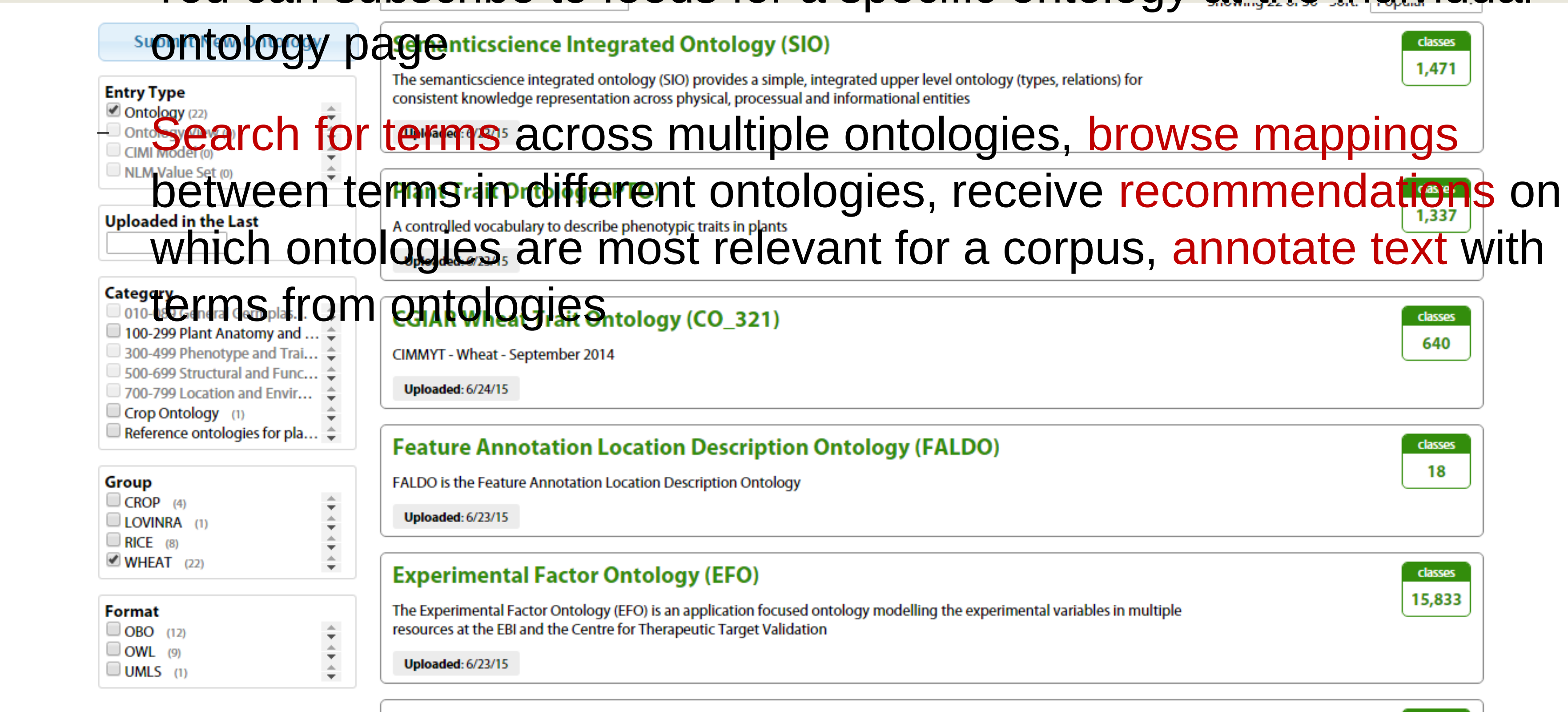
```
##fileformat=VCFv4.1
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT 102 4
labasskaja CS Estacao M6 Marquis Neepawa PI153785 P
PI185715 PI192001 PI192147 PI192569 PI210945 PI2226
297 PI349512 PI366716 PI366905 PI382150 PI406517 PI
I481718 PI481923 PI565213 PI82469 PI8813 PR267 Roem
cc3 acc4 acc5 berkut chakwal86 cham6 clear_white dh
maco opata pavon pbw343 rac875 vorobey
3929455_1al 1623 . T C 245.53 . AC=18;AF=0.196;AN=9
;Dels=0.00;FS=0.000;HaplotypeScore=0.1087;Inbreedin
AF=0.196;MQ=100.00;MQ0=0;MQRankSum=-1.426;QD=27.28;
D:DP:GQ:PL 0/0:1,0:1:3:0,3,41 0/0:1,0:1:3:0,3,41 1/
:3:41,3,0 ./. 0/0:1,0:1:3:0,3,41 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
```

```
0:1:3:0,3,36 0/0:1,0:1:3:0,3,38 0/0:1,0:1:3:0,3,39 0/0:1,0:1:3:0,3,39 0/0:1,0
:1:3:0,3,38 0/0:1,0:1:3:0,3,38 0/0:1,0:1:3:0,3,38 1/1:0,1:1:3:39,3,0 1/1:0,1:
1:3:38,3,0 1/1:0,1:1:3:38,3,0 0/0:1,0:1:3:0,3,39 0/0:1,0:1:3:0,3,38 1/1:0,1:1
:3:38,3,0
```

Writing: WDI working group
Creation date: 02 October 2014
Update: 30 June 2015

Wheat related vocabularies in Agroportal

- http://wheat.agroportal.lirmm.fr/ontologies
  - Access to, and retrieve the ontologies through the Web interface, an API and a Sparql Endpoint
  - Subscribe a RSS feed to receive alerts for submissions of new ontologies, new versions of ontologies, new notes, and new projects. You can subscribe to feeds for a specific ontology at the individual ontology page
  - Search for terms across multiple ontologies, browse mappings between terms in different ontologies, receive recommendations on which ontologies are most relevant for a corpus, annotate text with terms from ontologies



Entry Type

☑ Ontology (22)
☐ Ontology View (0)
☐ CIMI Model (0)
☐ NLM Value Set (0)

Uploaded in the Last

Category

☐ 010-089 General Germplas...
☐ 100-299 Plant Anatomy and ...
☐ 300-499 Phenotype and Trai...
☐ 500-699 Structural and Func...
☐ 700-799 Location and Envir...
☐ Crop Ontology (1)
☐ Reference ontologies for pla...

Group

☐ CROP (4)
☐ LOVINRA (1)
☐ RICE (8)
☑ WHEAT (22)

Format

☐ OBO (12)
☐ OWL (9)
☐ UMLS (1)

Semanticscience Integrated Ontology (SIO) — classes 1,471
The semanticscience integrated ontology (SIO) provides a simple, integrated upper level ontology (types, relations) for consistent knowledge representation across physical, processual and informational entities
Uploaded: 6/23/15

Plant Trait Ontology (PTO) — classes 1,337
A controlled vocabulary to describe phenotypic traits in plants
Uploaded: 6/23/15

CGIAR Wheat Trait Ontology (CO_321) — classes 640
CIMMYT - Wheat - September 2014
Uploaded: 6/24/15

Feature Annotation Location Description Ontology (FALDO) — classes 18
FALDO is the Feature Annotation Location Description Ontology
Uploaded: 6/23/15

Experimental Factor Ontology (EFO) — classes 15,833
The Experimental Factor Ontology (EFO) is an application focused ontology modelling the experimental variables in multiple resources at the EBI and the Centre for Therapeutic Target Validation
Uploaded: 6/23/15

The benefits

For data producers, managers, providers

- One stop shop for relevant information related to wheat data management □ arise awareness, avoid duplicated efforts, foster adoption of common practices
- Facilitate the use of common data exchange formats □ easy data sharing/submission to international repositories
- Foster a standardized description of datasets with consistent use of ontologies and metadata □increase the identification, the findability and the usability of the dataset

For data scientists, bioinfomaticians

- Facilitate the access, integration and analysis of data from various sources
- Access to data of higher quality

For top management, researchers

- Increase the chance to answer complex questions

Transitioning Cereal Systems to Adapt to Climate Change

REACCH
Regional Approaches to Climate Change –
PACIFIC NORTHWEST AGRICULTURE

Acknowledgement

***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)*

And

Clément Jonquet (Lirmm, Agroportal), Hélène Lucas (Wheat Initiative) Hadi Quesneville (WheatIS EWG)

Transitioning Cereal Systems to Adapt to Climate Change

REACCH
Regional Approaches to Climate Change –
PACIFIC NORTHWEST AGRICULTURE

Thank you to our sponsors:

We will add this to the end of each presentation