ANALYSIS OF SURVEY: DATA STANDARDS IN THE WHEAT RESEARCH COMMUNITY

WHEAT DATA INTEROPERABILITY WG

Start date: 07/04/2014

End date: 03/06/2014

SOMMAIRE

GENERAL STATISTICS

Total number of answers: 201

Number of complete answers: 125

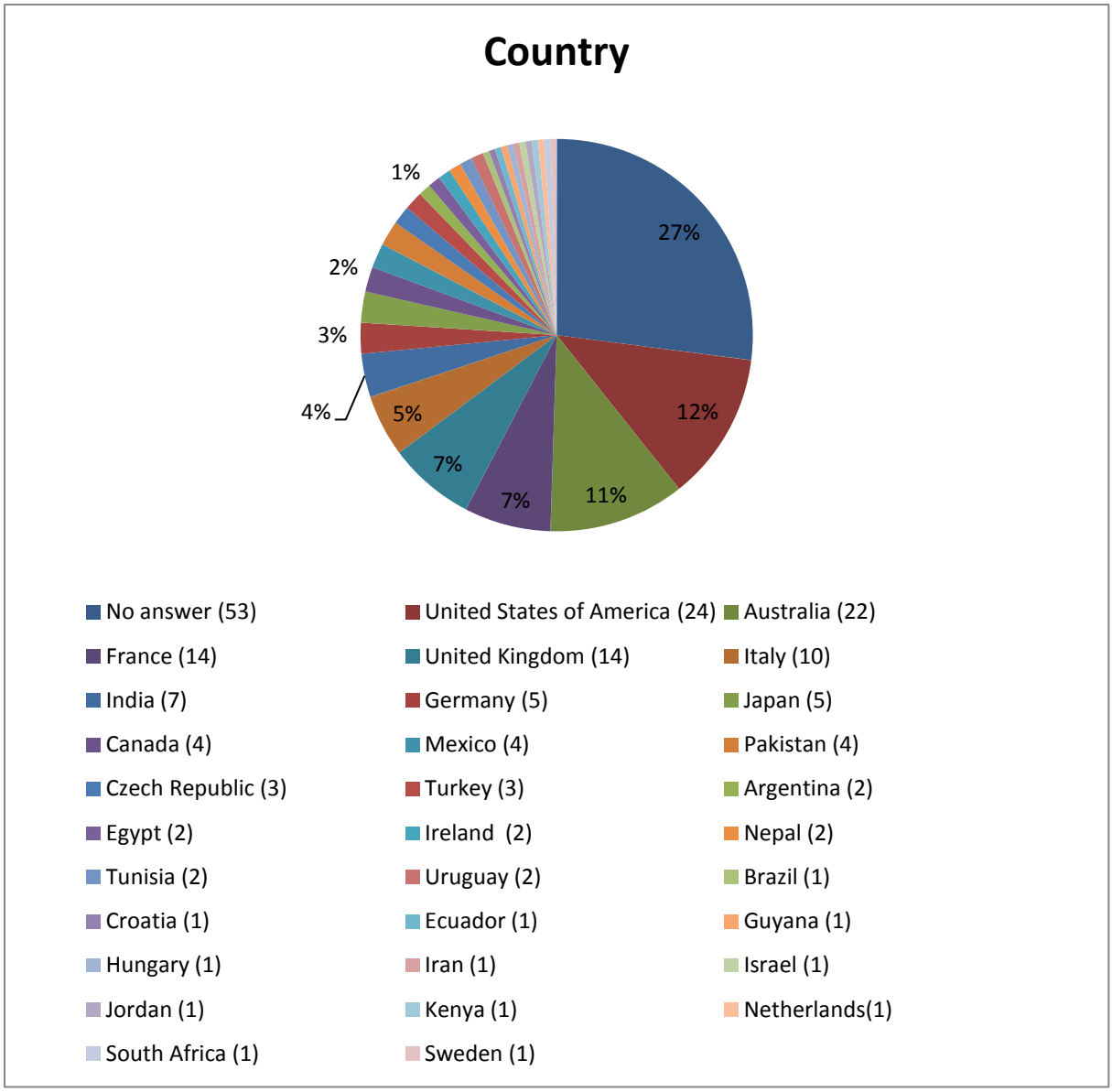
Total number of incomplete answers: 77 (6 doubles removed: people who answered twice)

Number of answers considered: **196**

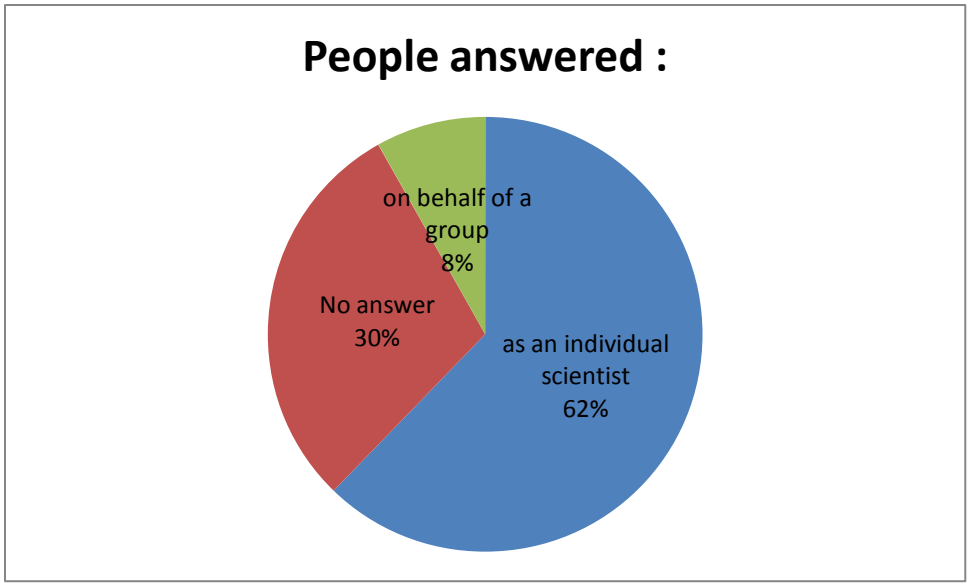
All answers, i.e. complete and incomplete, are taken into account in the following as questions were not mandatory.

PEOPLE

COUNTRIES



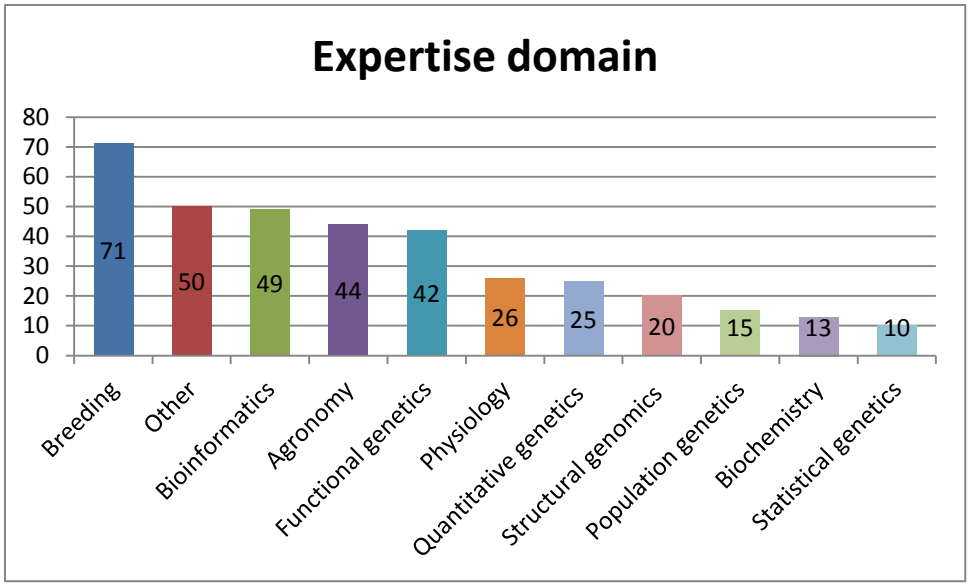
PEOPLE ANSWERS AS



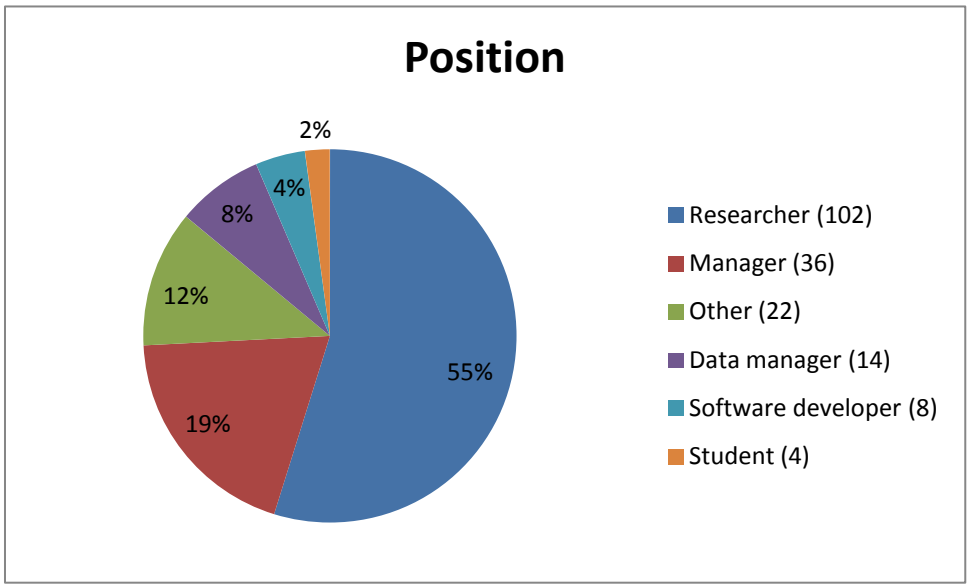
When on behalf of a group, those groups are:

- Agricultural Model Intercomparison and Improvement Project (AgMIP)
- Applied Bioinformatics
- CIMMYT Data Management team
- CIMMYT Global Wheat Program
- Genetic and Biotechnology Lab. UNS
- Genetic Resources
- Jülich Plant Phenotyping Center JPPC
- LEPSE
- Phenomics group at ACPFG
- SEVEN - UMR GDEC
- Soft Wheat Quality Lab
- UCD Crop Science
- UMT CAPTE Avignon
- Wheat Germplasm Bank, CIMMYT

FIELDS OF EXPERTISE OF PEOPLE



POSITION OF PEOPLE

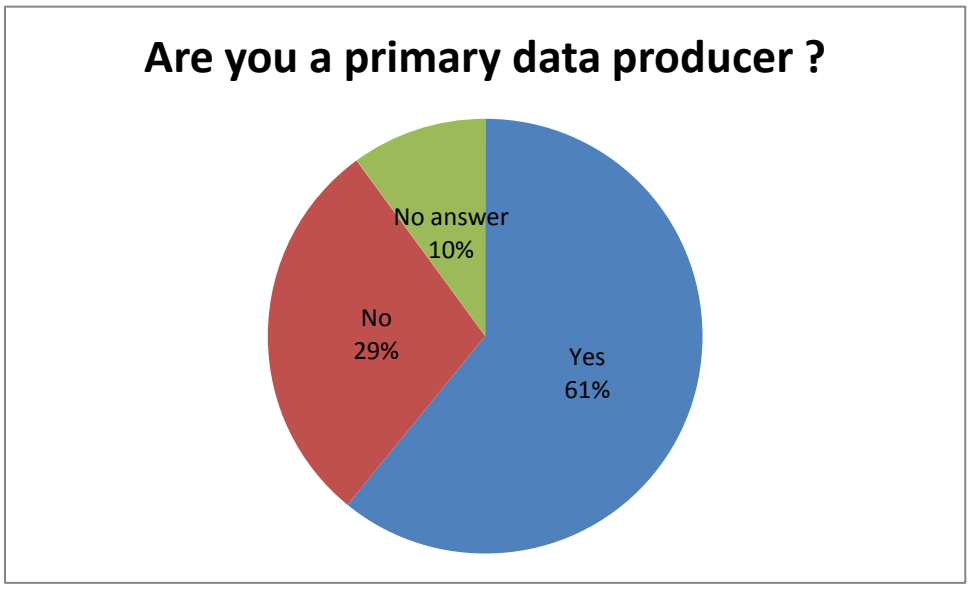


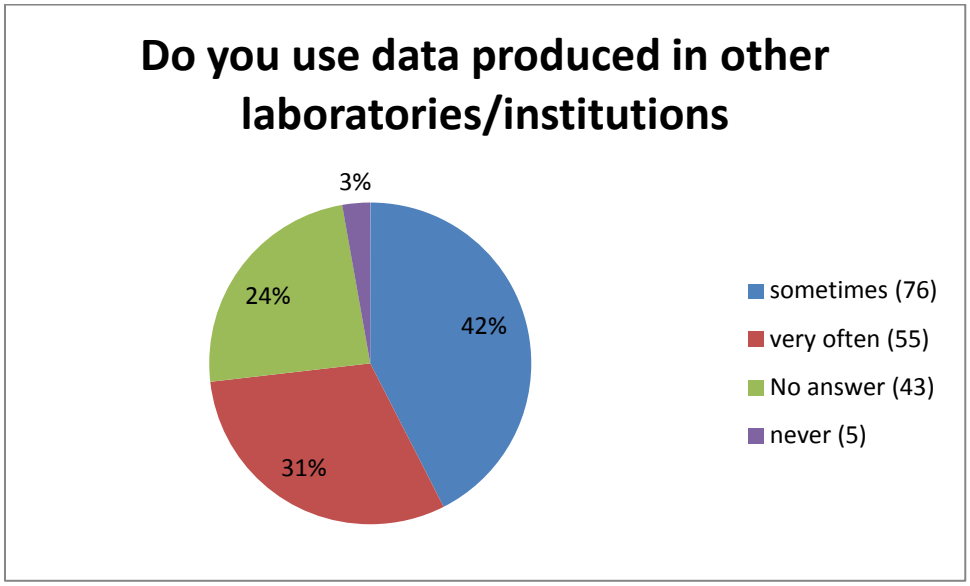
Other positions mentioned are:

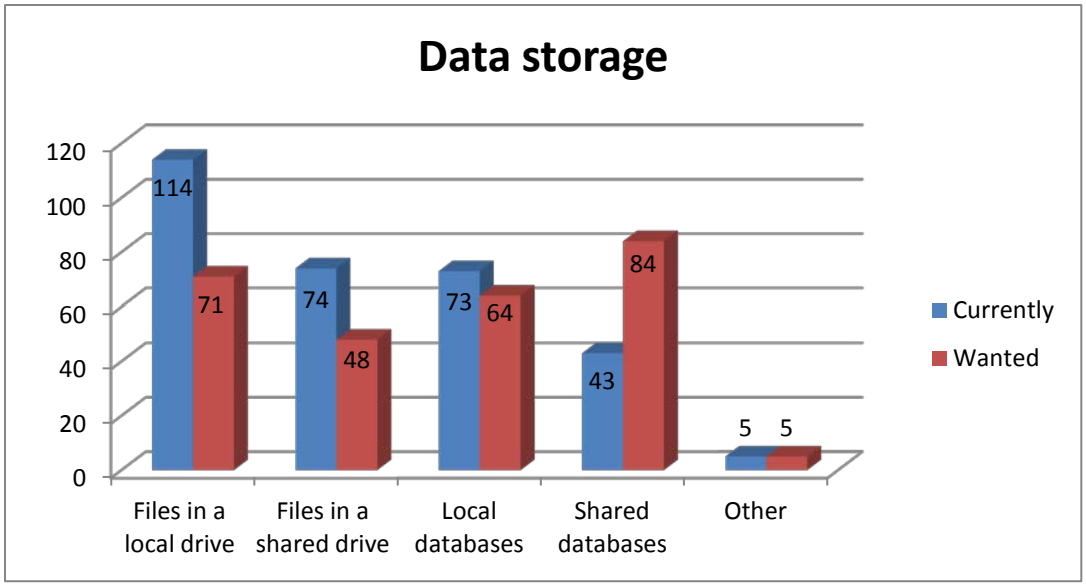
- Academic
- Assistant Professor (2)
- Breeder (2)
- Coordinator
- curator
- Director
- Emeritus
- genebank curator
- Group Leader
- head of breeding dep.
- Honorary Associate Professor

- Lecturer
- Linux Sysadmin
- Professor (3)
- Sales and Marketing
- senior lecturer
- teacher
- technician

ACTIVITIES REGARDING DATA







Shared databases used:

Count	Name
4	GrainGenes
3	CerealsDB
3	various, lots of them
2	NCBI
2	T3
2	Triticeae Toolbox
1	AgMIP (data.agmip.org, under development)
1	Broad Institute
1	Chado database
1	collaborating Institute
1	cortex.ivec.org
1	EST
1	Galaxy
1	Genbank
1	GeneSys,
1	GnpIS
1	goa (government of Alberta)
1	Goverment and also we are control field before harvest to estimate yield every harvest season
1	Gramene
1	GRIN
1	GRIN Global
1	HarvEST
1	Institutional
1	Integrated Breeding Platform
1	iPlant
1	iPlant Collaborative
1	miRBase
1	mySQL
1	Phenotyping DB
1	Plant Ontology
1	plexdb
1	PODD - Phenomics Ontology Driven Data and Metadata repository
1	postgres
1	Postgres SQL

1	publicly available biological databases
1	R2N Internal databases
1	Raw Genome Data
1	Reference Genomes
1	SQL Server
1	T3 Toolbox
1	the Climate Change Repository of Evaluation Trials (Agtrials)
1	The metadata are in the Crop Ontology database
1	the online geospatial database on collected samples
1	Those of the herbarium database
1	tropgenedb
1	UNIPROT
1	URGI DB
1	urgi gnpIS
1	Wheat genome database
1	wheat sequences hosted by URGI
1	WISP

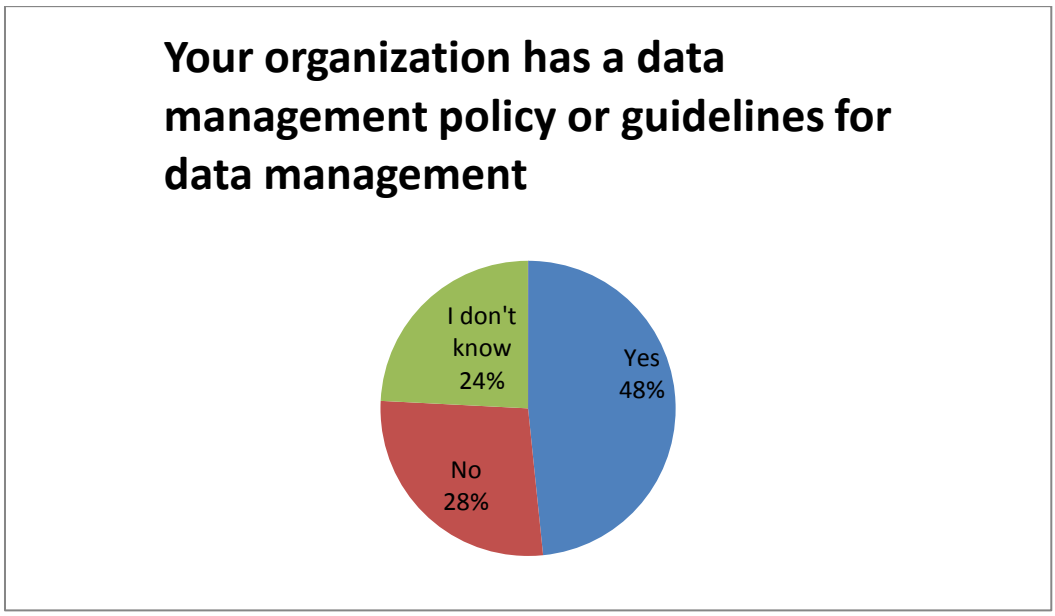
INSTITUTIONS, COMPANIES

INSTITUTIONS, COMPANIES OF PEOPLE ANSWERING THE SURVEY

Count	Name
7	INRA
5	ACPFG, University of Adelaide
5	CIMMYT
3	Australian Centre for PLant Functional Genomics
3	Forschungszentrum Jülich GmbH
3	Kansas State University
3	USDA ARS
2	Bioversity Interantional
2	Cornell University
2	INIA
2	INRA GDEC
2	Institute of experimental botany
2	Kyoto University
2	Nepal Agriculture Research Council
2	NIAB
2	The University of Adelaide
2	The University of Bristol
2	University of Tuscia
1	AgMIP
1	Agricultural Institute Osijek
1	Agricultural Institute, Centre for Agricultural Research, Hungarian Academy of Sciences, H-2462, Martonvásár, POB 19, Hungary
1	Agricultural Research Center, Field Crops Res. Inst., Wheat Res. Dep
1	Agricultural Research Council - Small Grain Institute
1	Agriculture & Agri-Food Canada
1	Agriculture and Agri-Food Canada, Plant Gene Resources of Canada
1	ARVALIS - Institut du végétal

1	Australian Centre for Plant Functional Genomics (ACPFG) Adelaide SA
1	CBBC
1	Central University of Himachal Pradesh
1	CIRAD
1	CRA Genomics Research Centre
1	CSIRO
1	CSIRO - High Resolution Plant Phenomics Centre
1	CSIRO, Plant Industry
1	Curtin University
1	DAFNE, UNIVERSITY OF TUSCIA
1	Department of Agriculture and Food Western Australia
1	department of Genetics Hazara university Mansehra
1	Dept plant breeding
1	DSV UK Ltd
1	Elsoms Wheat Ltd
1	Embrapa - Brazilian Agricultural Research Corporation
1	ENEA
1	Field Crop Development Centre, Alberta Agriculture and Rural Development
1	GAP International Research and Training Center
1	General Dir. of Agricultural Reserch and Politics
1	General Directorate of Agriculture Research and Policies
1	Genomics Research Centre, CRA
1	Indian institute of science education and research
1	Institute of Crop Science
1	Institute of Soil & Environmental Sciences, University of Agriculture Faisalabad
1	Instituto Nacional Autonomo de Investigaciones Agropecuarias - INIAP
1	INTA (National Institute for Agriculture and Husbandry Technology)
1	Intermountain Herbarium, Utah State University
1	IPK Gatersleben
1	IRD
1	ITC Ltd
1	Jamia Hamdard
1	jki
1	John Innes Centre
1	Kenya Agricultural Research Institute
1	Kihara Institute for Biological Research, Yokohama City University
1	Limagrain Cereal Seeds
1	Mendel University in Brno
1	NAREI
1	NARO
1	National Agronomic Institute of Tunis
1	National Bureau of Plant Genetic Resources, Pusa Campus, New Delhi-12
1	National Center of Agricultural Research and Extension (NCARE)
1	National Research Centre
1	National Research Centre on Plant Biotechnology
1	National Research Council (CNR)-Institute of Biosciences and Bioresources
1	North Dakota State University
1	Nuclear Institute of Agriculture, Tando Jam
1	Open University
1	oregon state
1	Oregon State University
1	PMAS Arid Agriculture University
1	Punjab Agricultural University ,Ludhiana, INDIA
1	R2N

1	Rothamsted Research
1	Royal Holloway University of London
1	Teagasc
1	Tel Aviv University
1	The James Hutton Institute
1	The University of Queensland
1	U.S. Department of Agriculture (USDA)
1	UC Davis
1	Universidad Nacional del Sur - CERZOS CONICET
1	Università Politecnica delle Marche
1	University College Dublin
1	University of bari
1	University of Cambridge
1	University of Modena and Reggio Emilia
1	University of Nebraska
1	University of Saskatchewan
1	University of Southampton
1	University of Sydney Plant Breeding Institute
1	UQ
1	USDA Agricultural Research Service
1	USDA-ARS WRRC
1	USDA-ARS, Agricultural Systems Research Unit
1	USDA-ARS, ALARC
1	wageningen-ur
1	Warren Farms
1	Washington State University



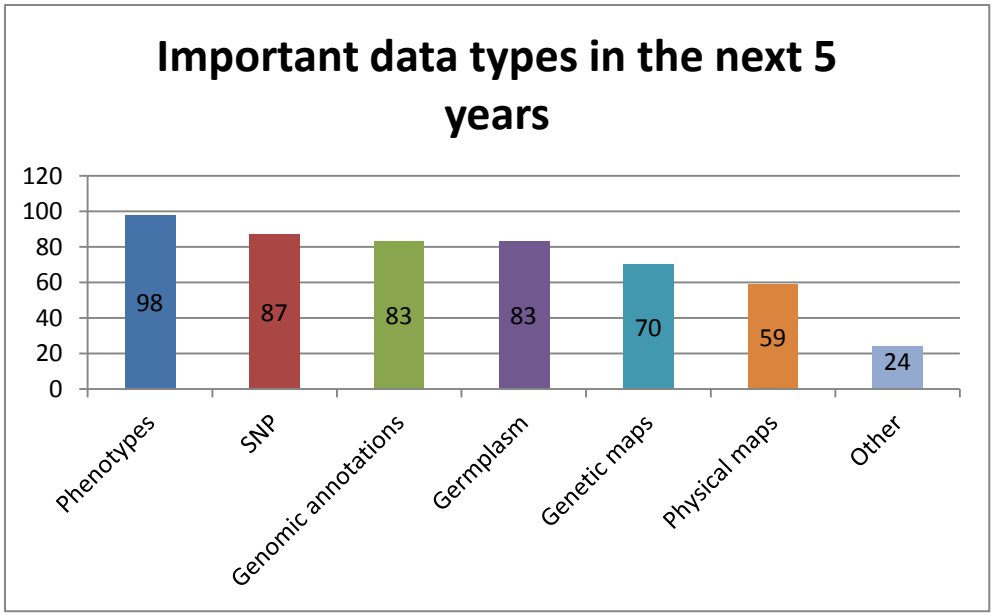
People who answered '**YES'** are from:

- ACPFG
- AgMIP
- Agricultural Institute, Centre for Agricultural Research, Hungarian Academy of Sciences
- Agricultural Research Center, Field Crops Res. Inst., Wheat Res. Dep
- Agriculture & Agri-Food Canada

- ARVALIS - Institut du végétal
- Australian Centre for Plant Functional Genomics
- Bioversity International
- CIMMYT
- CSIRO
- Curtin University
- Dept plant breeding
- DSV UK Ltd
- Elsoms Wheat Ltd
- Embrapa - Brazilian Agricultural Research Corporation
- Field Crop Development Centre, Alberta Agriculture and Rural Development
- Forschungszentrum Jülich GmbH
- INRA
- IPK Gatersleben
- ITC Ltd
- John Innes Centre
- Kansas State University
- Kenya Agricultural Research Institute
- Kihara Institute for Biological Research, Yokohama City University
- Kyoto University
- Limagrain Cereal Seeds
- NARO
- National Bureau of Plant Genetic Resources, Pusa Campus, New Delhi-12
- National Research Centre on Plant Biotechnology
- Nepal agricultural research council
- Nuclear Institute of Agriculture, Tando Jam
- Open University
- Oregon State University
- R2N
- Teagasc
- U.S. Department of Agriculture (USDA)
- University College Dublin
- University of Adelaide
- University of Modena and Reggio Emilia
- University of Nebraska
- University of Sydney Plant Breeding Institute
- UQ
- USDA Agricultural Research Service
- USDA-ARS, Agricultural Systems Research Unit
- USDA-ARS, ALARC
- wageningen-ur

DATA TYPES

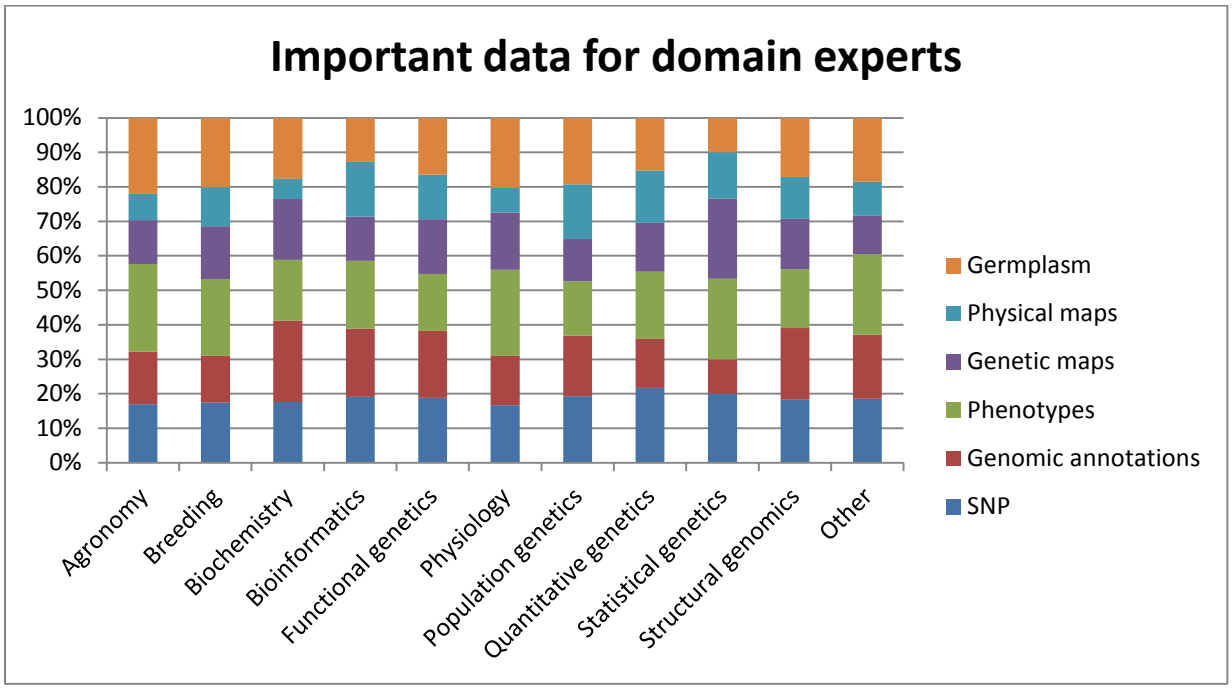
DATA TO BE CONSIDERED IMPORTANT



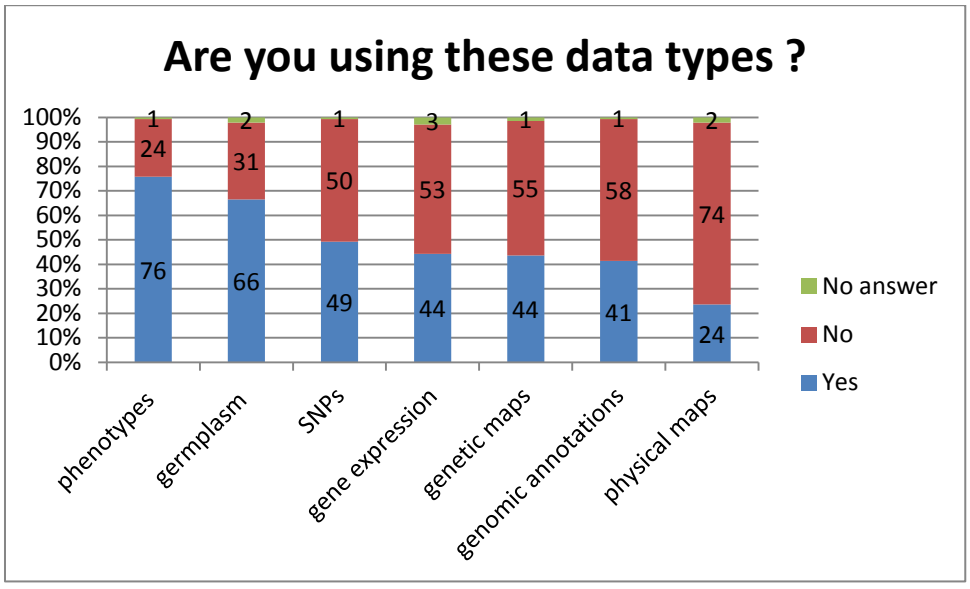
Other data are:

- all these tools and others not mentioned such as QTL, association mapping, etc.
- ALLERGENOMICS AND, IN GENERAL, HEALTH RELATED COMPONENTS
- Biochemical pathways
- Climatic or weather data
- Environmental Data related to trials
- epigenome, TE behavior
- exome capture
- Experimental data
- G*E*S
- gene expression, transcription factors, small RNAs, chromatin modifications
- Genotyping, Protocols
- I am most familiar with / interested in data for modeling including management of experiments, crop development measurements, crop growth measurements, genotype to phenotype relations, weather, soils.
- metabolomics
- methodologies
- microarray, transcriptomics
- pan genomes
- Passport data
- proteomics data
- pseudomolecules
- RNAseq
- Robust molecular markers
- Sequence data
- transcriptome data of various conditions and tissues
- foolish question - depends on kind of research involved

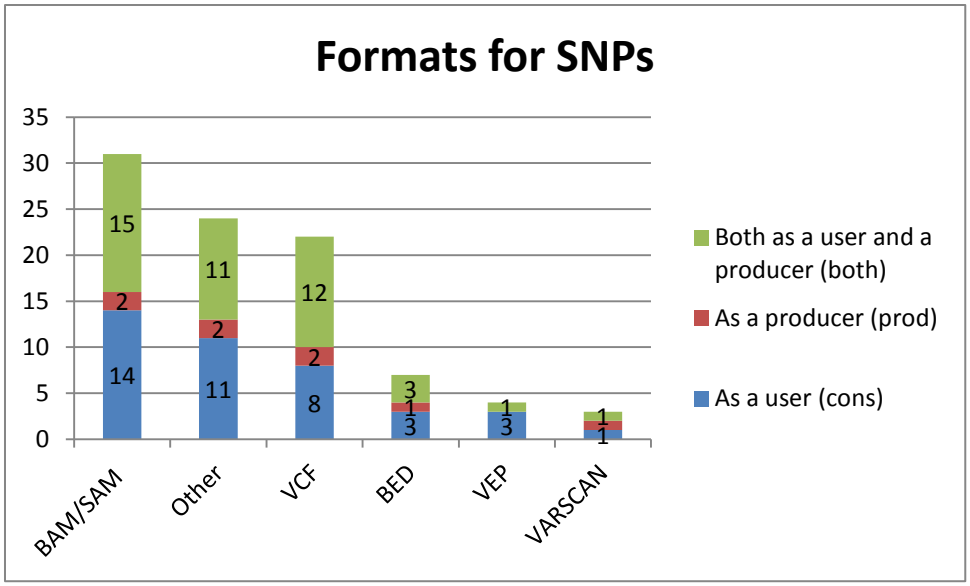
The previous question is analyzed regarding the expertise domain of the person who answered (one person can have several expertise domains and propose several important data)



DATA CURRENTLY USED/PRODUCED

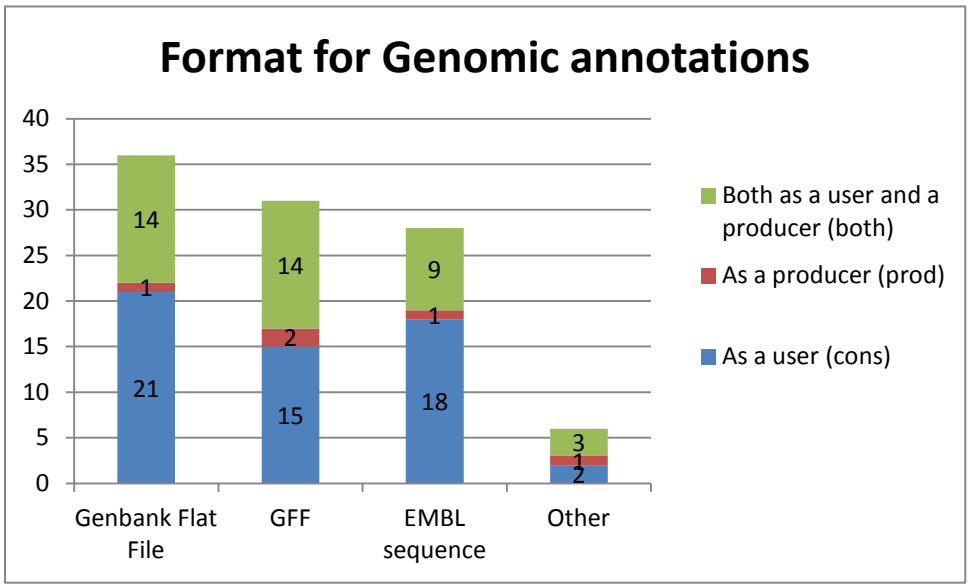


SNPS



No other formats were mentioned.

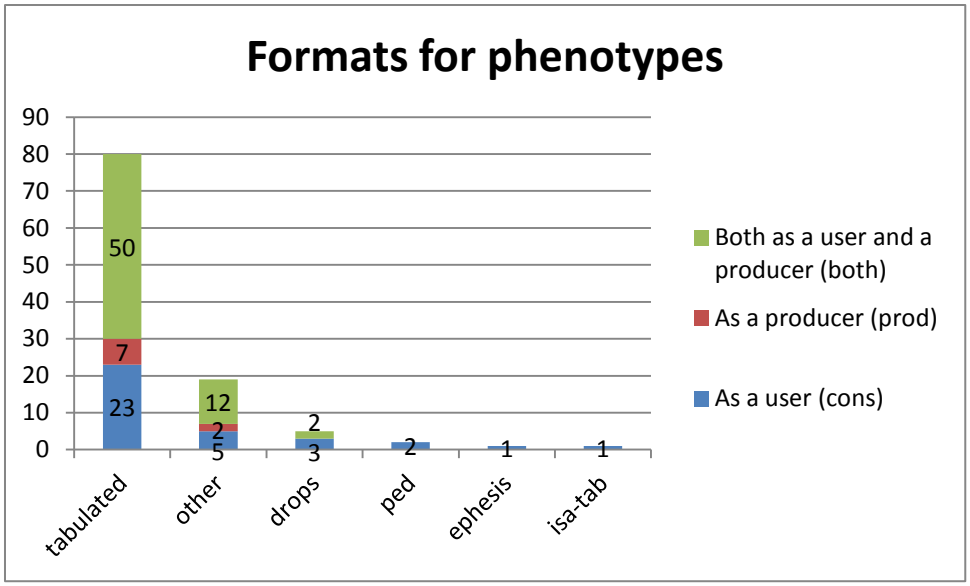
GENOMIC ANNOTATIONS



Other formats mentioned are:

- fasta
- Geneious
- GeneOntology (GO)
- gff3, dat
- Inhouse CSV + XML
- variety of custom formats, user and producer

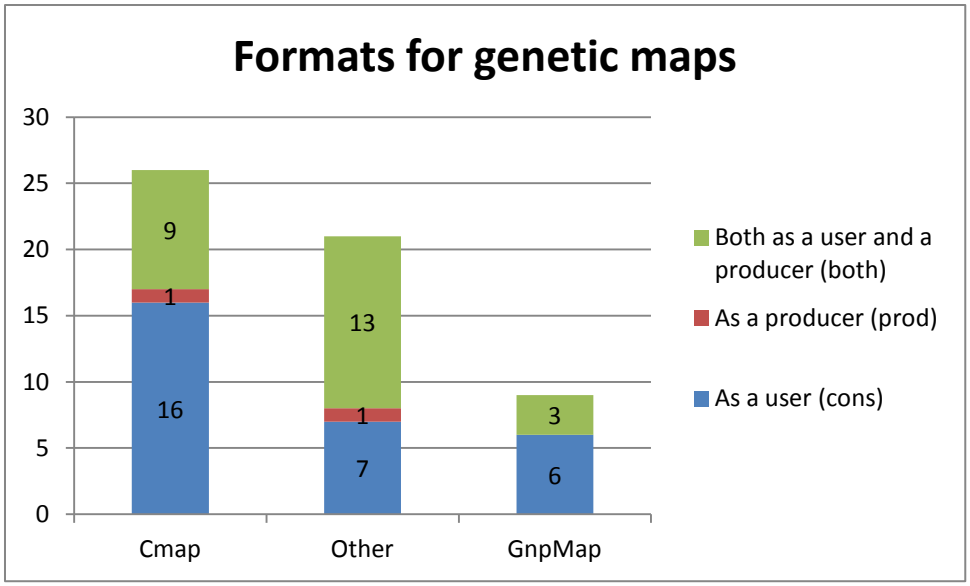
PHENOTYPES



Other formats mentioned are:

- 3D data sets (vtk etc) and images
- access database
- binary files/ matlab files / ascii files
- csv
- Databases--Foxpro
- Excel format white space delimited Written scores in field books
- greenhouse and field experiments
- Have worked with a full range of formats that people use to share data - from delimited text files, spreadsheets, Word style tables, hard copy and databases ranging from dBase and Access to MySQL and similar.
- Import traits into the program AgrobaseII.
- INRA-BBSRC and ADAPTAWHEAT (FP7) formats
- jeg,
- non-standard formats for high-throughput data
- Observed experimental data for crop models in AgMIP harmonized format.
- RDF triples
- SAS
- Some phenotyping undertaken as imaging therefore original images sometimes need to be available for re analysis
- T3 data files
- various databases
- We run our own systems in EPPN and DPPN that are more adequate than the ones mentioned

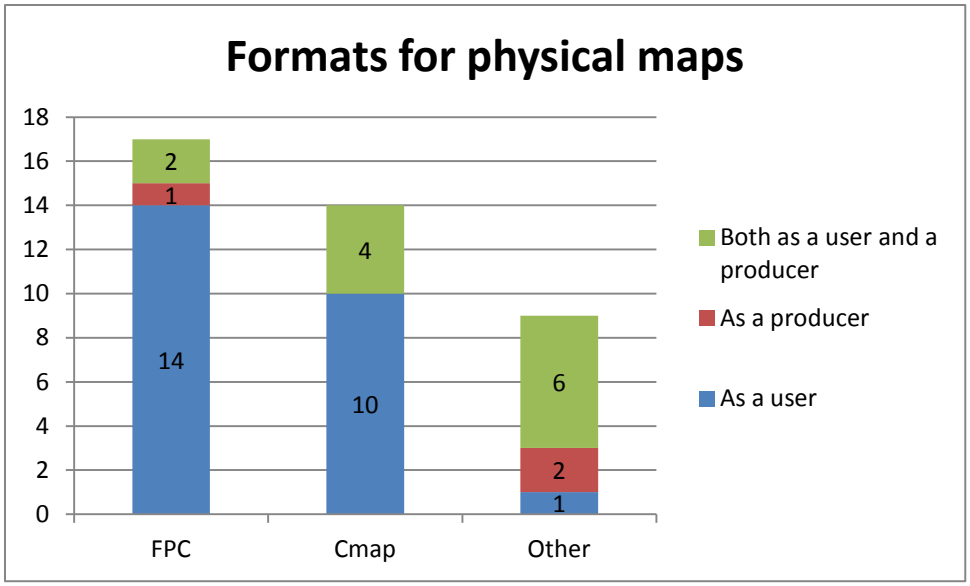
GENETIC MAPS



Other formats mentioned are:

- AB formatted (consumer) Cluster calls (Ilumina GenomeStudio (producer)
- csv table (as a user)
- custom use/produce
- custom, stroodle
- Depending on the QTL software: RQTL, QTLcartographer, ...
- excel
- GBrowse gff3
- Inhouse CSV + XML
- join map
- joinmap / Rqtl
- mpmap (consumer)
- No idea
- R map format, csv
- R, MapDisto, DST, JoinMap
- R/QTL files in the form: marker name, chromosome, postion in many (most?) packages with various delimiters and orders
- tab
- text
- text, csv
- txt format as a producer and user
- we are user

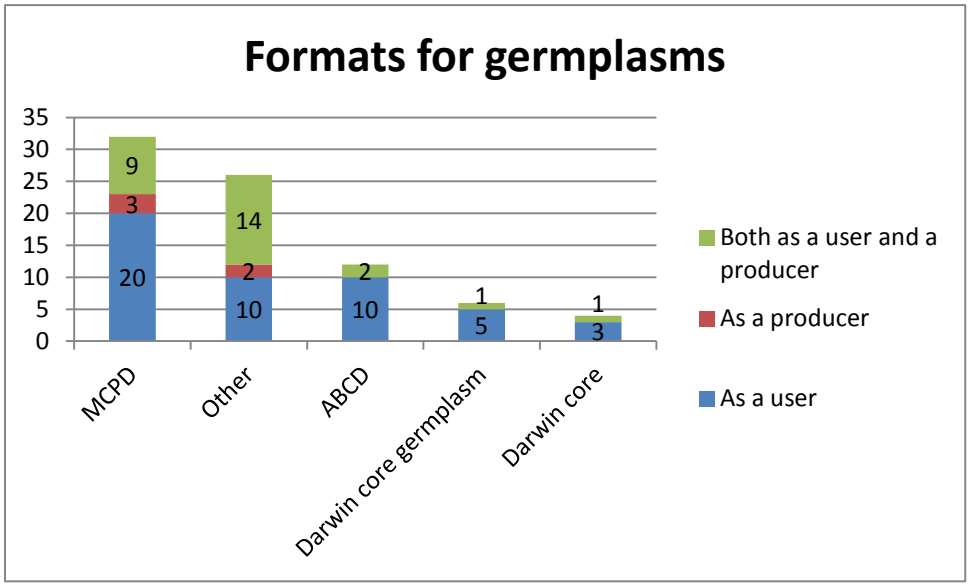
PHYSICAL MAPS



Other formats mentioned are:

- carthagene
- custom use/produce
- Farm works
- GBrowse
- genetically anchored physical maps (POPSEQ)
- LTC
- no idea
- various and custom

GERMPLASMS

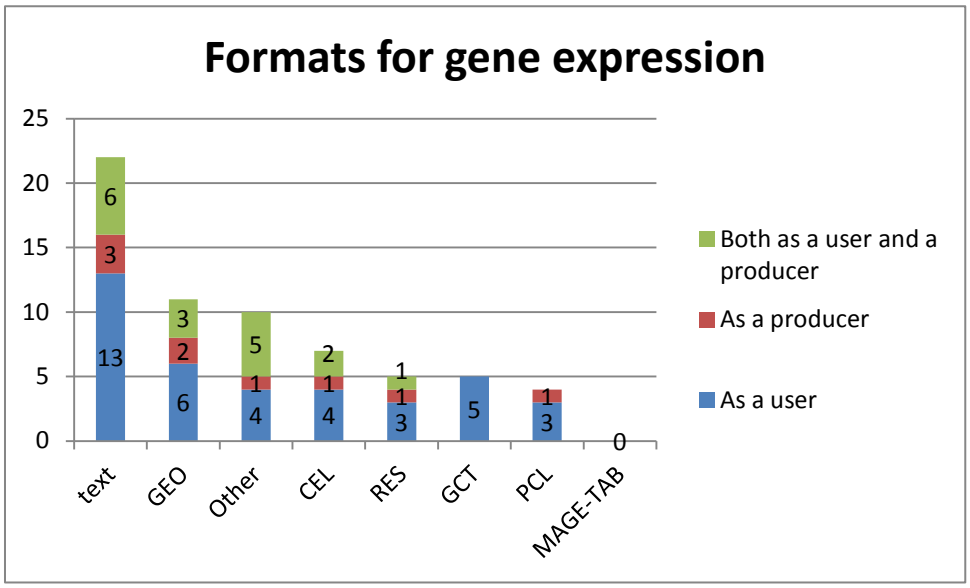


Other formats mentioned are:

- EVIGEZ
- excel

- EXCEL
- Excel
- Excel (producer)
- Excel files.
- from Specifics Seed Banks
- GRIN
- GRIN files
- I use the pedigree method of Purdy et al Crop Sci.8 (1968) as both producer and consumer. I also use the PI numbers from GRIN (USDA-ARS)and the SA numbers from the Germplasm collection at our Institute as a consumer.
- in house
- internal nomemclature for maize, TAIR/NASC code for arabidopsis thaliana
- Katmandoo, T3, custom use/produce
- local database
- no idea
- Our own
- Prepare my own
- private format
- tab
- text
- upov
- USDA GRIN
- various
- xcel

GENE EXPRESSION



Other formats mentioned are:

- custom
- custom use/produce
- Excel (both)
- fasta

- Format for proteomics analysis (mascot, mgf)
- Inhouse CSV + XML
- RNAseq data (from Illumina reads); CLC Genomic Workbench outputs
- variant of UniGene data

OTHER TYPES OF DATA

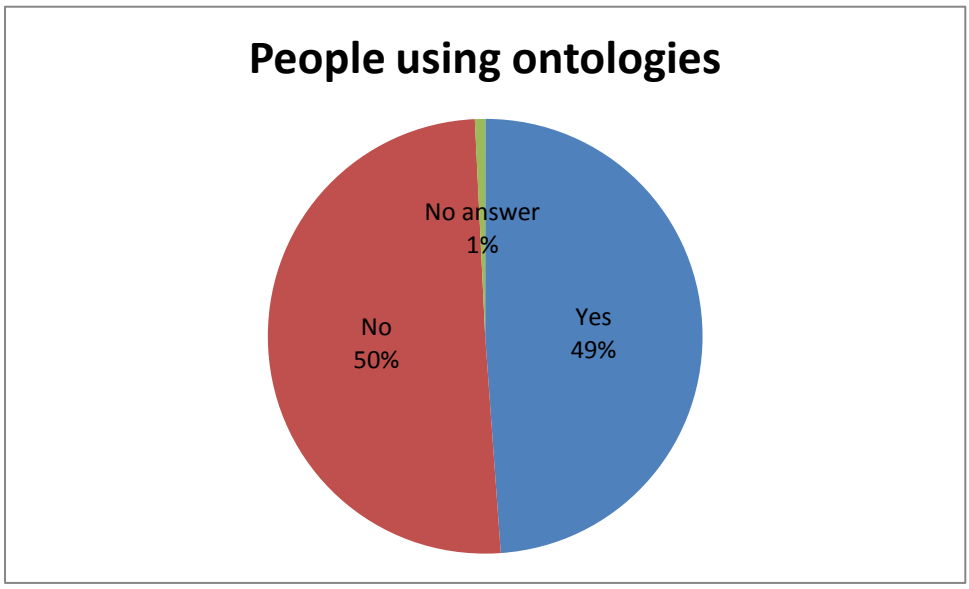
What other type(s) of Wheat data are you working with?

- phizical cold tolerans
- A number of ontologies coded as RDF triples within PODD
- Agronomic data and disease scoring data
- biotic and abiotic stress resistant
- Coordination Between Enstitues
- Distributional information
- Enzyme function
- Gene annotation
- genome assemblies using Gap5
- genome references/resequence data
- Genomic, transcriptomic, annotations
- Have and will work with any data that could be useful in identifying novel genetic variation for breeding.
- Heat stress tolerant traits, spad, ndvi
- High density molecular marker data such as GBS, usually from HDF5 format. Images generated by remote sensing, mosaic images with geo-references (usually large size)
- I have done many field and greenhouse studies on basic wheat physiology/agronomy/soil aspects for different management practices, environments, and peripherally in relation to breeding.
- Images
- images (plant phenotypes)
- Introgression
- Ionomics
- IWGSC survey sequences GBS data
- Molecular markers primer information
- Mostly data relevant to cropping system models from breeder trials, field experiments, farm surveys, etc. Data are converted and stored in AgMIP harmonized format (compressed JSON) and translated to model-specific formats used by multiple crop models (DSSAT, APSIM, STICS, CropSyst, WOFOST, etc.)
- Pedigree/Geneology, Coefficient of Parentage, non-SNP genotypic information, e.g. PAVs and low-density marker information (for particular rust resistance genes, etc.), remote sensing data (e.g. plot canopy temperature),
- phenotypic data
- Phenotypic data - disease reaction and end-use quality; text format
- Phenotypic scores for cereal rust evaluation as well as pre-harvest sprouting
- Proteomic / proteogenomic datasets Metabolomic datasets
- Proteomics
- public field trial data, looking at Phosphate use and uptake
- sequence data-genome browser
- sequences (FASTA) reads (FASTQ)
- Simulation, input output

- Validation of functional markers for heat and drought stress
- Variety registration data.
- Weather, crop management, soils ICASA standards are my preferred format

VOCABULARIES AND ONTOLOGIES

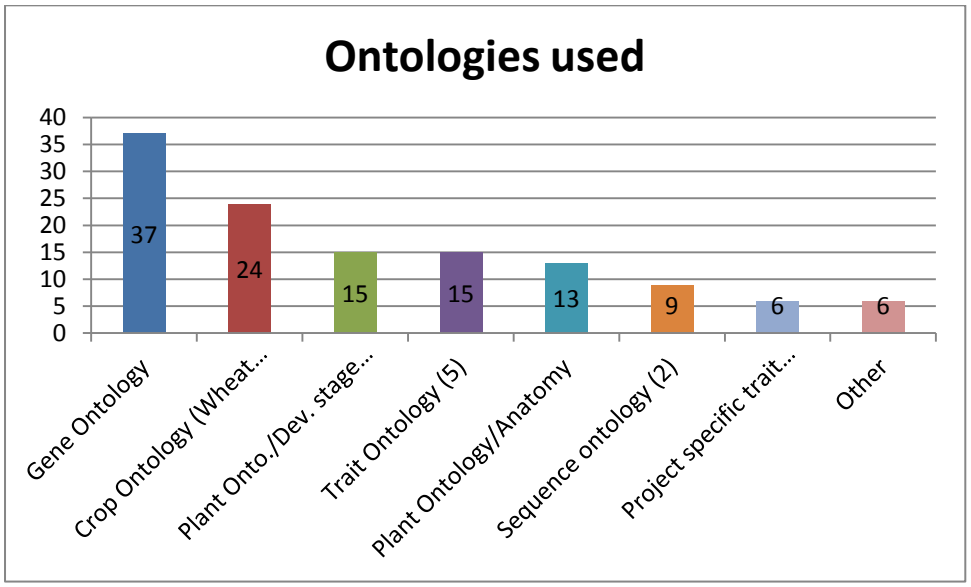
ONTOLOGIES



Why not ?

- do not know
- Don't trust the annotation
- Have concluded that ontologies are more suited for qualitative data. I DO use a vocabulary/dictionary, which is the ICASA master list
- I do not know what these are
- I DO NOT NEED FOR THE TYPE OF PROJECTS I HAVE A THE MOMENT
- I don't know
- I don't know
- I find use of the term "ontology" rather difficult to interpret as it is being used nowadays
- i found it not useful
- I want to, but don't know how to get started (partly my fault, of course).
- In development. Needs restructuring or mapping with current database development in our Institute.
- In progress.
- It not appropriate to the work I am doing at the moment.
- Lots of talk about their development, but little/no implementation.
- n/a
- NA
- No agreement, standards, incomplete
- No apparent need.
- No need
- No need in my breeding program
- no required to at present

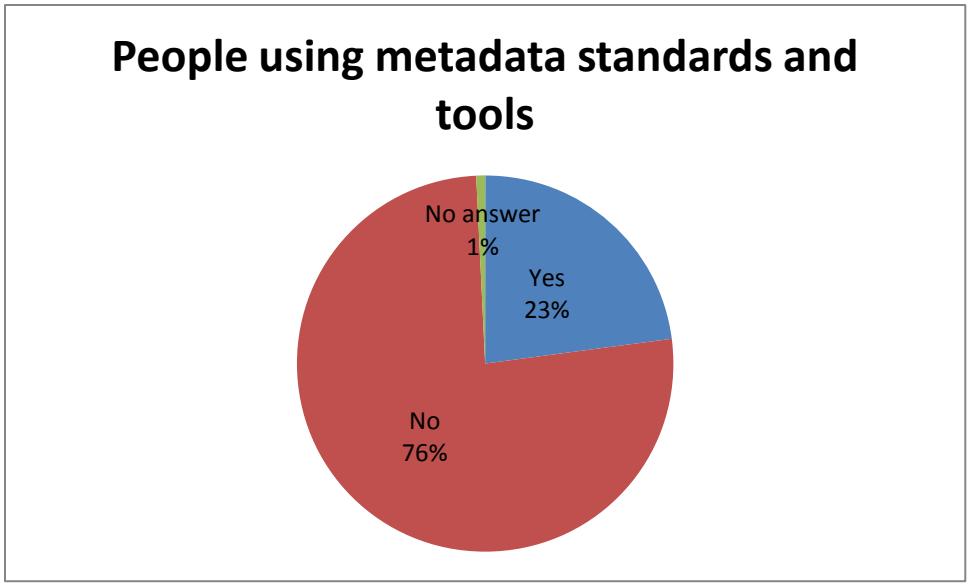
- not applicable
- not applicable
- Not capacitate
- Not initiated
- Not known
- not relevant for type of research (yet)
- Not yet feeling the need, but probably will adopt one soon
- There is no standard in my organisation
- too complicated
- Unknown to me in this context.
- We are using the data vocabulary established by the International Consortium of Agricultural Systems Applications (ICASA).
- wheat quality ontologies are not available



Other ontologies mentioned are:

- ECPGR
- Ontologies to develop conceptual ABM
- PATO, XEML
- Plant Environmental conditions ontology
- plant pathogens:: http://www.pathoplant.de/; PLEXdb;
- QUDT

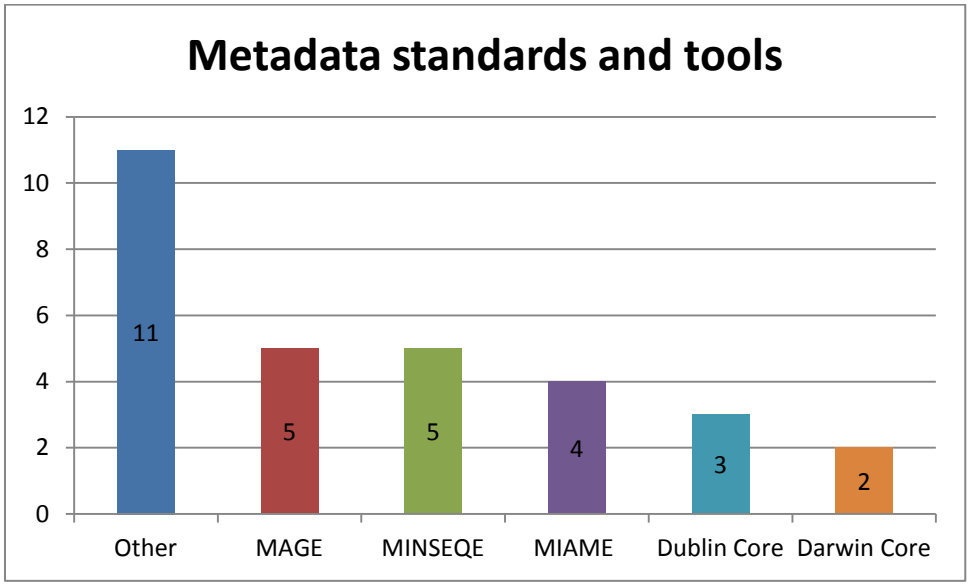
METADATA



Why not ?

- Because I am just a conventional breeder
- ditto
- do not know
- do not know these
- Do not what it is
- Have no idea what it is.
- have no more idea about this
- have no need
- I DO NOT NEED FOR THE TYPE OF PROJECTS I HAVE A THE MOMENT
- I do not work with metadata yet
- I don't know
- I don't know
- I don't know about this
- I don't know them or how to use them with fiability
- Just started to plan for our data
- Lots of talk about their development, but little/no implementation.
- n/a
- NA
- No capacity.
- no existing metadata standard for phenotyping data
- No idea what they are or how to use them
- no idea what this is
- No need
- no need
- No need
- No need as of now.
- no required to at present
- No requirement for current research
- not applicable
- not applicable

- not applicable to current datasets
- Not awared
- not creating new metadata
- not familiar with it
- not initiated
- Not introduced to
- Not known
- not needed
- Not yet feeling the need, but probably will adopt one soon
- Same ad above. In addition, to my knowledge, there are no supported ontologies in the public domain that cover 'experimental designs', 'environmental data', and experimental metadata in general (e.g., instruments and method used, etc) that are useful and consistent.
- there no oportinity
- Too complex
- Too much other stuff to worry about
- We are trying to make a move towards using MINSEQE
- We don't know about them



Other metadata standards and tools are:

- AgMIP
- AgMIP/ICASA
- custom use/produce
- custom xml based format used for my simulations
- FCDC
- ICASA
- mostly internally developed standards
- other
- Phenotypic metadata standards developed in Genesys
- PODD science ontology
- un-standardized

ADDITIONNAL QUESTIONS

COMMENTS ON THE SURVEY

- Data sharing in wheat research community open it will help the scientist.
- a bit generic, the purpose of the survey is not clear
- Capacity building components, gxe tools
- Fewer surveys, more action.
- Good start!
- Good work in addresseing these question
- I am mostly an end user so I am not deeply involved in data production and handling. I work with people that handle the data
- I think that the wheat sequnce data as presented in URGI server is not sufficant. For instance, many groups have information of BAC end sequnce- i think it should be presented in the physical maps of the genome brwoser. There is no sequnce information about the markers presented in the genome brwoser. Thank you
- I think this survey is a good start at getting an idea of what people use.
- i would be happy to be informed of news, announcements,working groups.
- It is good that you have started such type of survey which is promoting wheat data analysis .....
- It is great if you can create a data portal, though it must be a tough work.
- It would be great if you share the results of the survey and update your recent activities.
- Main interest is to ensure access to as much data relating to wheat (and other crops) as possible to facilitate discovery and use of germplasm. Believe that spatial data is of considerable use in this process. Future is probably in developing data 'conduits' between major systems aggregating and sharing all data (passport, characterization, phenotypic, environmental and genetic) to enable its effective and efficient use through novel discovery methods and subsequent deployment.
- More background information what you are planning to do would be useful.
- No comment
- Sure, its my pleasure.
- Suvey is Good it will be good if more annotation will be provided for wheat gene
- The genetic map question did not seem to work correctly, but its probably because I changed the answer.
- The survey almost covered the important topics related to data sharing. But there are topics like standardizing protocols, sharing un-published data and huge-data storage need to be discussed.
- The survey is a bit difficult to understand for me as a non Bioinformatic expert person, for example, I am not aware of different current formats for data storage in databases.
- To do anything meaningful with phenotypic data, you need to know about the environment and management. Recall that the phenotype = f(G x E x M), where G=genotype, E=environment and M = crop management.
- untill now, registered for one year with wheat research community without any efficiency
- Very useful and interactive survey
- We would be grateful to see the results of this survey once they have been collected.
- ya if you need any information about my Lab as well as our research, you can contact without any hesitation.