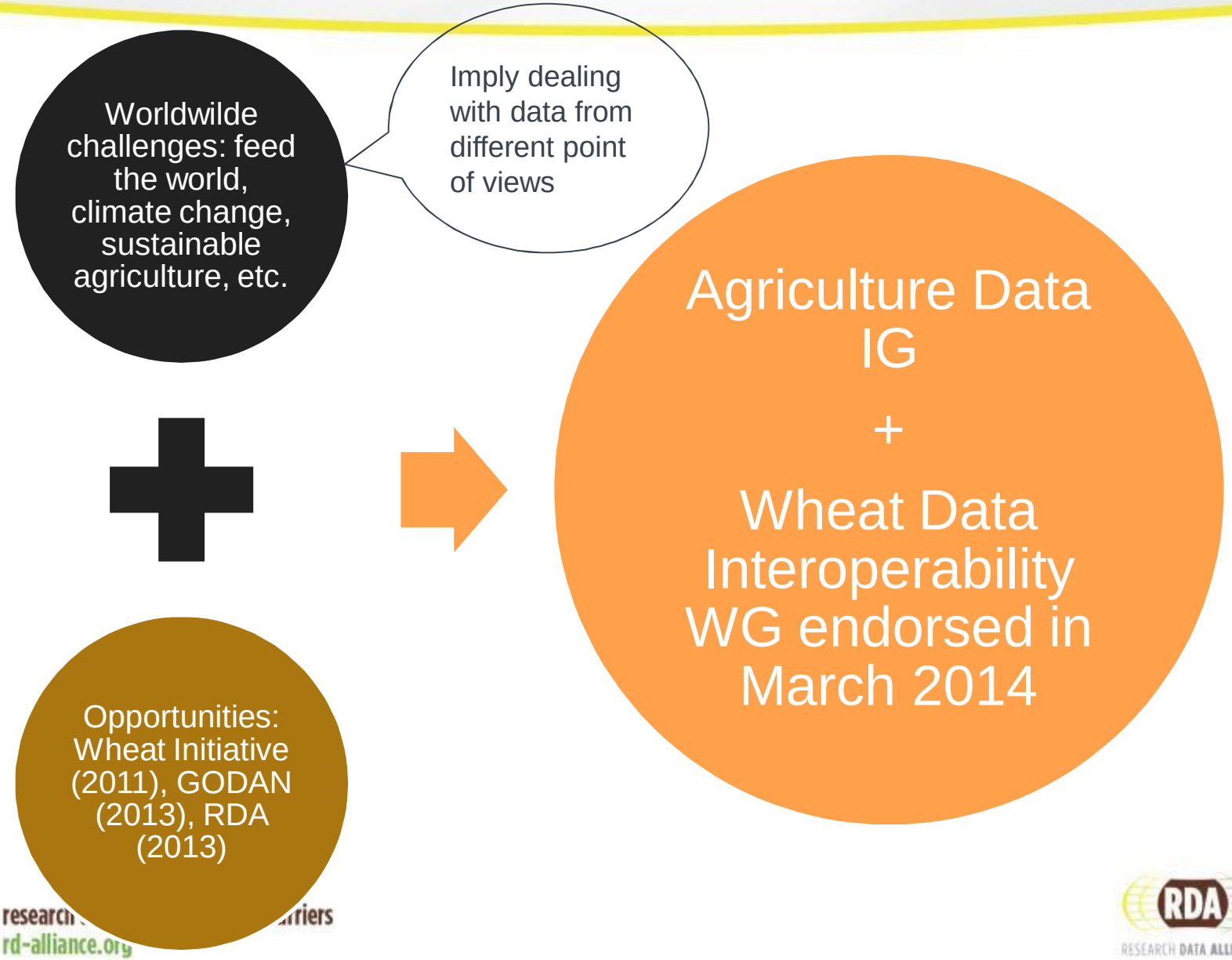
RESEARCH DATA ALLIANCE

Wheat Data Interoperability

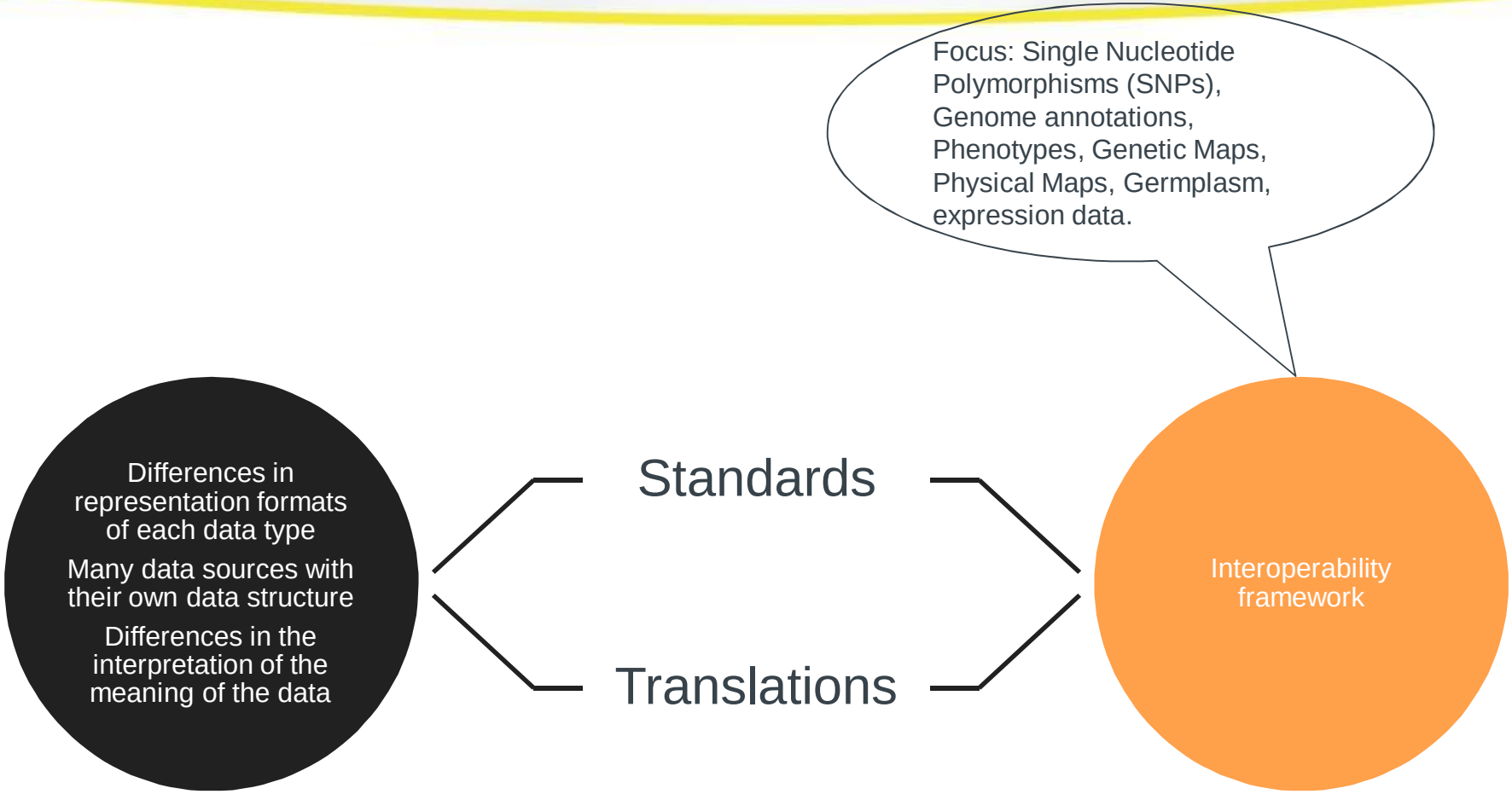
research data sharing without barriers

rd-alliance.org

Context in which the WG has been created



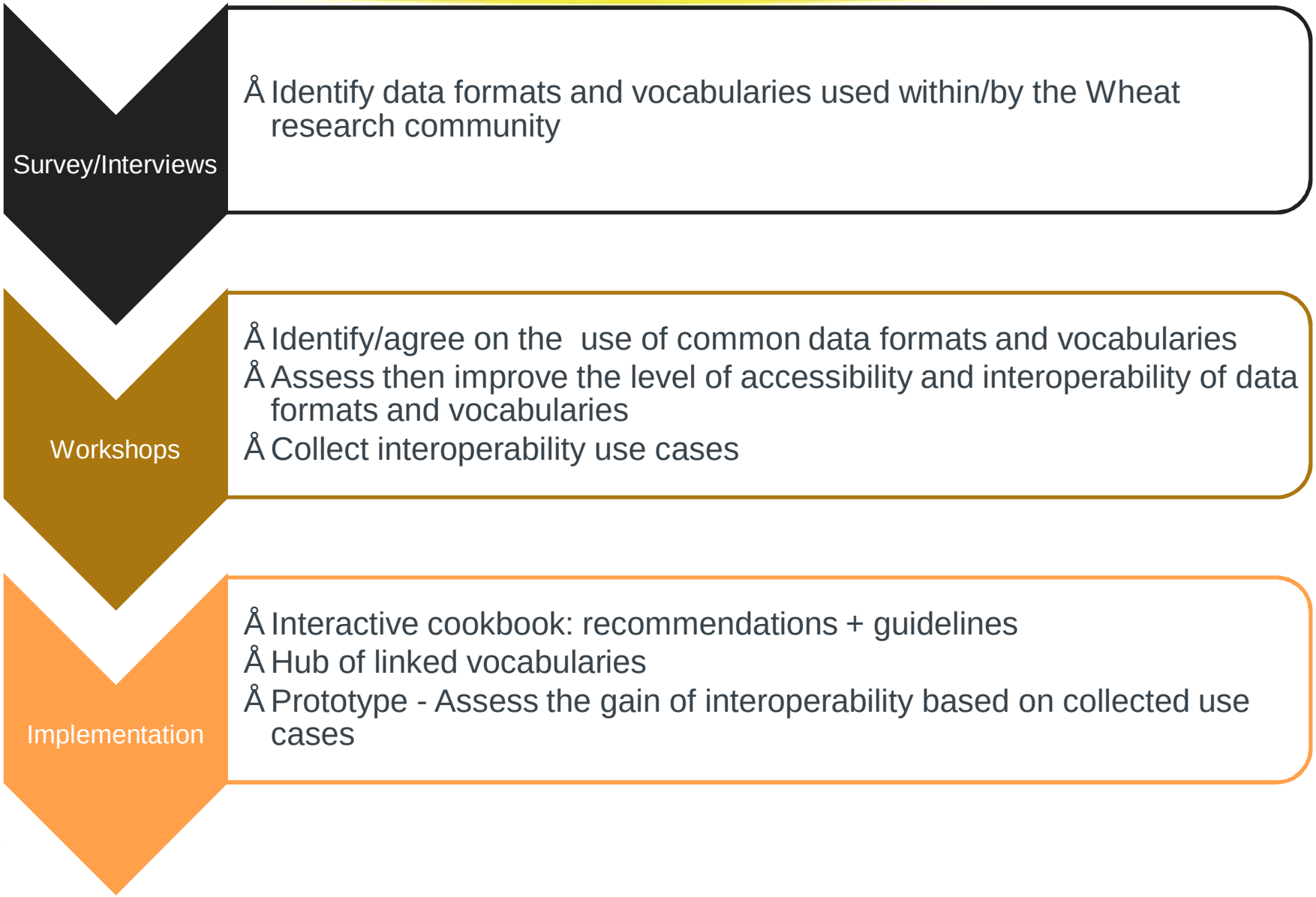
The WG main objective



Achieving semantic interoperability

- Two paths towards semantic interoperability
  - Make everyone speak the same language (common standards: data formats, metadata, vocabularies)
    - Based on the Wheat research community practices
    - Leverage existing data formats and vocabularies
  - Provide ñtranslationsòamong the existing metadata, data formats and vocabularies
    - Leverage existing translation rules
- Our strategy
  - Address both standards and translations
  - Use the W3C Resource Description Framework (RDF) as common information representation format

Towards semantic interoperability

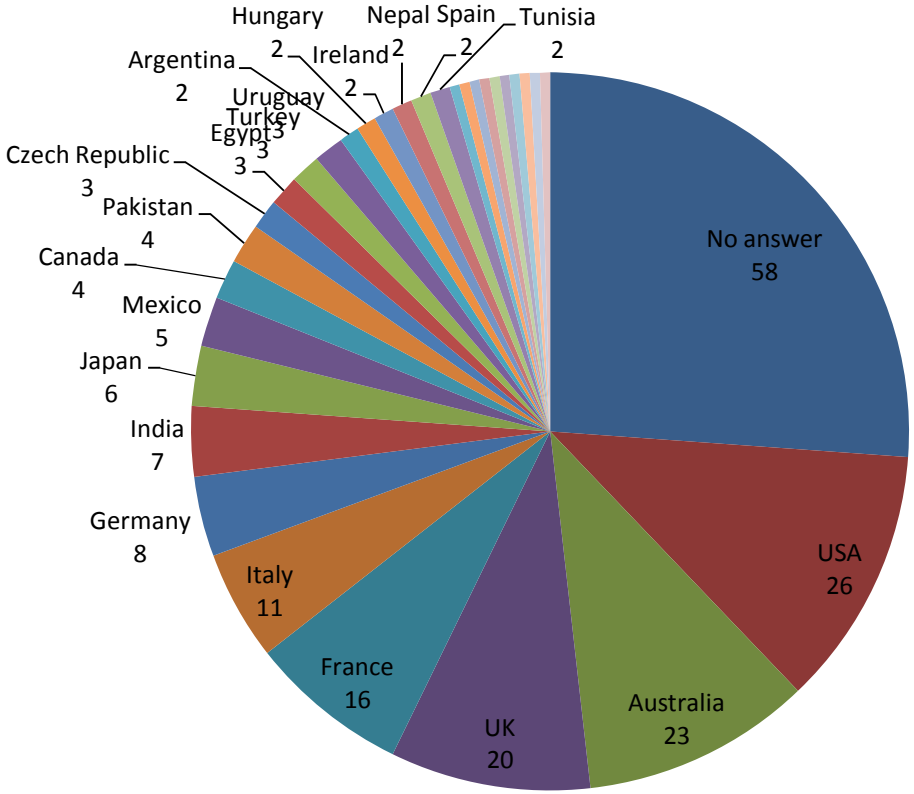


Where we are 1/2

Survey/Interviews

- A survey launched in April 2014
- Answers from more than 200 respondents from at least 31 different countries

The survey: participants



No answer	United States of America	Australia	United Kingdom		
France	Italy	Germany	India		
Japan	Mexico	Canada	Pakistan		
Czech Republic	Egypt	Turkey	Uruguay		
Argentina	Hungary	Ireland	Nepal		
Spain	Tunisia	Bangladesh	Brazil		
Croatia	Ecuador	Guyana	Iran		
Israel	Jordan	Kenya	Netherlands		

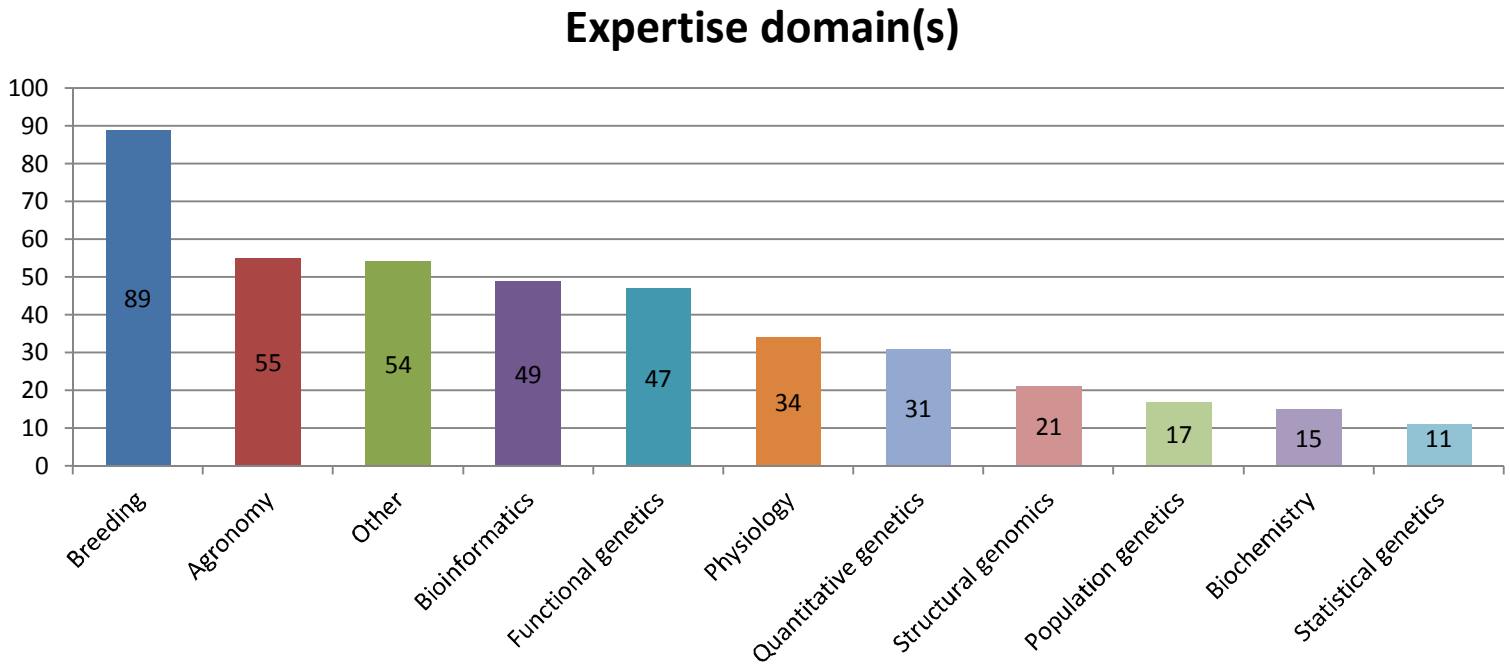
Top wheat producers(in million metric tons)

Rank	Country	2009	2010	2011	2012
Source: UN Food & Agriculture Organization [71]					
ð	*World*	686	651	704	675
1	China	115	115	117	126
2	India	80	80	86	95
3	United States	60	60	54	62
4	France	38	40	38	40
5	Russia	61	41	56	38
6	Australia	21	22	27	30
7	Canada	26	23	25	27
8	Pakistan	24	23	25	24
9	Germany	25	24	22	22
10	Turkey	20	19	21	20
11	Ukraine	20	16	22	16
12	Iran	13	13	13	14
13	Kazakhstan	17	9	22	13
14	United Kingdom	14	14	15	13
15	Argentina	9	15	14	11

Source : http://en.wikipedia.org/wiki/Wheat

RDA
RESEARCH DATA ALLIANCE

Research areas of survey participants



Additional domains (Other)

- ÅMolecular biology
- ÅGenetic resources
- ÅPhenotyping
- ÅMolecular plant pathology
- ÅEntomology
- ÅMolecular genetics
- ÅPlant biotechnology
- ÅGenomics
- ÅClimate
- ÅData management
- ÅModeling
- ÅOntologies, taxonomies
- Åé

The survey: participants

Also work with ↙	SNPs (76)	genome annotations (64)	Phenotypes (121)	Genetic maps (72)	Physical maps (41)	Germplasms (104)
SNPs		70,3%	50,4%	76,4%	75,6%	54,8%
genome annotations	59,2%		33,9%	54,2%	68,3%	37,5%
Phenotypes	80,3%	64,1%		81,9%	75,6%	89,4%
Genetic maps	72,4%	60,9%	48,8%		92,7%	51,9%
Physical maps	40,8%	43,8%	25,6%	52,8%		30,8%
Germplasms	75,0%	60,9%	76,9%	75,0%	78,0%	

The survey: use of data formats

Data type	Data formats		
	Standardized	Tool specific	Non standardized
SNPs	VCF	BAM/SAM, BED, VARSCAN, VEP	
genome annotations	Genbank Flat File, General Feature Format (GFF), EMBL		
Germplasms	MPCD, ABCD, Darwin Core, Darwin Core Germplasm	Grin Global	tabulated
Gene expression	Many format standards laid out by repositories such as NCBI (GEO) and EBI Array Express		
Physical maps	GFF	Cmap, fpc	
Genetic maps		Cmap, gnpmap	
Phenotypes		Drops, ped, isa-tab, ephesis	tabulated

Where we are 2/2

Survey/Interviews

- A survey launched in April 2014
- Answers from more than 200 respondents

Workshops

- 1-2 October 2014
- List of recommended data formats and vocabularies for each data type
- List of follow up actions for each data type (standardization e.g for traits, minimal set of metadata e.g for SNP files provenance or for markers, QTL and map handling, check for existing mapping tools, etc..)
- List of interoperability use cases

Examples of use cases

Title	**Searching for germplasm with specific traits**
Description	Example of searching for germplasm with specific traits - tagged with ontology terms?
Data types	Germplasm Phenotype
Challenges	ŏ Metadata very important ~ standardized format ŏ Association of genes to traits, linked to germplasm, marker information ŏ Need for quality controls- how confident are you of the data source? ŏ Provenance of the germplasm- pedigree, ownership, ŏ Standard system for tracking germplasm, names
Title	**Identification of wheat genes that control root growth**
Description	Requires: Annotated genes (Gene Ontology, PFam, and other functional annotation)
Data types	Genomic annotations? - Gene location ? (IWGS-SS ID or MIPS HCS link)
Challenges	Mapping between wheat genes and orthologs from other species (deduce function by seq. similarity); Access to RNASeq data (genes that are not expressed in roots may be irrelevant) ; mapping of wheat genes and information on their function based on literature
Title	**Query on trial data associated with varieties**
Data types	Phenotypic data, GIS data, (wheat economy/production data)
Description	To search wheat varieties with distribution maps, production figures, performances in wheat mega environments, associated projects worldwide plus layers of climatic data on specific wheat production areas and disease prevention information.
Challenges	Phenotypic data should be linked to GIS data. Using keywords or ontology terms a system or a tool should be able to pull out such information from different websites/systems developed by wheat community.

Workshop summary

	Recommendations	Follow up actions
SNPs	Use of VCF data format	Look at a metadata set to contextualize the provenance of SNP files
genome annotations	Use of GFF3 data format Use of ontologies to fill « Attributes » column (a list of feature attributes in the format tag=value), column 9	Provide description guidelines for filling in content for column 9
Germplasms	Use of MPCD and Darwin Core Germplasm formats	Check how to integrate with tool specific formats (Grin Global, Genesys) Provide a table like human readable format for DWC Germplasm
Gene expression	Follow existing format standards laid out by repositories (NCBI GEO, EBI Array Express	Check for mapping and conversion tools
Physical maps	Same as for genome annotations	Same as for genome annotations
Genetic maps	Data formats depend on tools that are used, rather concentrate on metadata harmonization	Look at a minimal metadata set to handle markers, QTL, maps Obtain details for linking requirements
Phenotypes	Use of isa-tab data format	Standardize the traits metadata Improve the reference to ontologies used for traits

Define metadata for data sets

	Recommendations	Follow up actions
SNPs	Use of VCF data format	Look at a metadata set to contextualize the provenance of SNP files
genome annotations	Use of GFF3 data format Use of ontologies to fill « Attributes » column (a list of feature attributes in the format tag=value), column 9	Provide description guidelines for filling in content for column 9
Germplasms	Use of MPCD and Darwin Core Germplasm formats	Check how to integrate with tool specific formats (Grin Global, Genesys) Provide a table like human readable format for DWC Germplasm
Gene expression	Follow existing format standards laid out by repositories (NCBI GEO, EBI Array Express	Check for mapping and conversion tools
Physical maps	Same as for genome annotations	Same as for genome annotations
Genetic maps	Data formats depend on tools that are used, rather concentrate on metadata harmonization	Look at a minimal metadata set to handle markers, QTL, maps Obtain details for linking requirements
Phenotypes	Use of isa-tab data format	Standardize the traits metadata Improve the reference to ontologies used for traits

Define/standardize metadata for the data

	Recommendations	Follow up actions
SNPs	Use of VCF data format	Look at a metadata set to contextualize the provenance of SNP files
genome annotations	Use of GFF3 data format Use of ontologies to fill « Attributes » column (a list of feature attributes in the format tag=value), column 9	Provide description guidelines for filling in content for column 9
Germplasms	Use of MPCD and Darwin Core Germplasm formats	Check how to integrate with tool specific formats (Grin Global, Genesys) Provide a table like human readable format for DWC Germplasm
Gene expression	Follow existing format standards laid out by repositories (NCBI GEO, EBI Array Express	Check for mapping and conversion tools
Physical maps	Same as for genome annotations	Same as for genome annotations
Genetic maps	Data formats depend on tools that are used, rather concentrate on metadata harmonization	Look at a minimal metadata set to handle markers, QTL, maps Obtain details for linking requirements
Phenotypes	Use of isa-tab data format	Standardize the traits metadata Improve the reference to ontologies used for traits

Harmonize some practices and provide guidelines

	Recommendations	Follow up actions
SNPs	Use of VCF data format	Look at a metadata set to contextualize the provenance of SNP files
genome annotations	Use of GFF3 data format Use of ontologies to fill in « Attributes » column (a list of feature attributes in the format tag=value), column 9	Provide description guidelines for filling in content for column 9
Germplasms	Use of MPCD and Darwin Core Germplasm formats	Check how to integrate with tool specific formats (Grin Global, Genesys) Provide a table like human readable format for DWC Germplasm
Gene expression	Follow existing format standards laid out by repositories (NCBI GEO, EBI Array Express	Check for mapping and conversion tools
Physical maps	Same as for genome annotations	Same as for genome annotations
Genetic maps	Data formats depend on tools that are used, rather concentrate on metadata harmonization	Look at a minimal metadata set to handle markers, QTL, maps Obtain details for linking requirements
Phenotypes	Use of isa-tab data format	Standardize the traits metadata Improve the reference to ontologies used for traits

Tie in the existing resources

	Recommendations	Follow up actions
SNPs	Use of VCF data format	Look at a metadata set to contextualize the provenance of SNP files
genome annotations	Use of GFF3 data format Use of ontologies to fill « Attributes » column (a list of feature attributes in the format tag=value), column 9	Provide description guidelines for filling in content for column 9
Germplasms	Use of MPCD and Darwin Core Germplasm formats	Check how to integrate with tool specific formats (Grin Global, Genesys) Provide a table like human readable format for DWC Germplasm
Gene expression	Follow existing format standards laid out by repositories (NCBI GEO, EBI Array Express	Check for mapping and conversion tools
Physical maps	Same as for genome annotations	Same as for genome annotations
Genetic maps	Data formats depend on tools that are used, rather concentrate on metadata harmonization	Look at a minimal metadata set to handle markers, QTL, maps Obtain details for linking requirements
Phenotypes	Use of isa-tab data format	Standardize the traits metadata Improve the reference to ontologies used for traits

Possible interactions with other WGs

- Biosharing registries working group
  - Register the Wheat Data Interoperability WG outputs: guidelines, standards
- Data type registries working group
  - Register the definitions for each data format used within/by the Wheat research community plus, where useful, references to tools that can operate on them.
- Biodiversity Data Integration IG
  - Share experience with bioinformaticians/people with experience in biodiversity metadata, standards
- Metadata interest group
  - Specific advice:
    - what requirements should we consider when thinking of a minimal metadata set in general?
    - what standard should we use to build a specific minimal metadata set (for example to contextualize the provenance of SNP files)?

Next steps

- A first version of the cookbook (recommendations and guidelines) by December 2014 + Call for feedback from the Wheat community
- Work with experts to meet the identified needs
  - Metadata harmonization, minimal metadata sets
  - Mapping among metadata, formats and ontologies
- Two workshops in 2015
  - Refine the cookbook
  - Collect more interoperability use cases
- Implement the prototype

Thanks!