

GPG2

Cross referencing of Rice and Wheat Germplasm

Global Public Goods 2

(GPG2)

The Project, implemented under the aegis of SGRP, is a comprehensive programme of work to **upgrade the CGIAR Centre genebanks and the standards of management of the collections.**

This will ensure that the **CGIAR Centres can meet their in-trust commitments, manage the collections efficiently and sustainably into the future, and facilitate access by users.**

Crop Registries

Goal

*"To create crop registers
for CGIAR crops in common"*

(Source: Jan Konopka-ICARDA)

GPG2 Crop Registries

Objectives

- ◆ To consolidate a list of accessions and associated information for a 'virtual global collection of a crop'
- ◆ To identify **overlap** between collections & unique holdings
- ◆ To collate, whenever possible, selected characterization and evaluation (registry specific)
- ◆ To publish on-line the consolidated information (with indication of overlap) and summaries about crop accessions
- ◆ To establish linkages with GPG2 projects:
 - 3.2 (one-stop-shop),
 - 4.1.1 (CG collecting missions),
 - 4.1.2 (quality of location data & geo-referencing)

(Source: Jan Konopka-ICARDA)

Crop Registries

Status

Agreed registers

Wheat

Barley

Cassava

Forages

Rice

Chickpea

Musa

Potato

Rice Crop Registry

Phase I

Rice Collection	Accessions
IRRI	117,272
WARDA	19,058
CIAT	1,635
USDA	34,451
INGER	24,716
Total	197,132 (Approx 1/3 of global holdings)

Wheat Crop Registry

Phase I

Wheat Collection	Accessions
CIMMYT	94,576
ICARDA	34,612
USDA	61,382
Total	190,570 (Approx 1/5 of global holdings)

Methodology

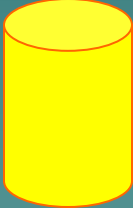
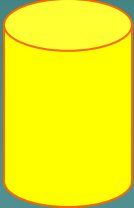

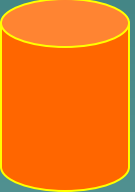
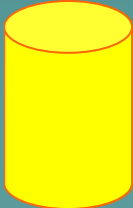
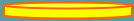


- ◆ Standardize selected passport data
- ◆ Select set of combinations to analyse
 - Germplasm exchange data
 - Similarity scores of selected passport descriptors

Methodology

◆ Similarity matrix

Descriptor	Comparison	Score
Species	Simple string comparison	0=no match, 1=match
Collecting no	Simple string comparison	0=no match, 1=match
Collecting date	Matching of year, month and day	For each matching part 0.33 is added to the score.
Sample status	Simple string comparison	0=no match, 1=match
Country of origin code	Simple string comparison	0=no match, 1=match
Province/state of origin	Levenshtein (edit distance)	Value between 0 and 1. Calculated as $1 - (\text{edit distance} / \text{Max edit distance})$
Latitude and longitude	Max difference between lat or longitude	Value between 0 and 1. $1 - (\text{Lat difference} / 180 +, \text{long difference} / 360) / 2$
Germplasm names	Levenshtein (edit distance)	Value between 0 and 1. Calculated as $1 - (\text{edit distance} / \text{Max edit distance})$
Pedigree	Levenshtein (edit distance)	Value between 0 and 1. Calculated as $1 - (\text{edit distance} / \text{Max edit distance})$

Methodology

Pairs	Descriptor	1	2	..	Average
A-B					
A-C					

Methodology

◆ DUP_DECISION (LEVELS)

SC1	Similarity Calculated Level 1: Reliable indication of similarity. E.g. direct linkages between germplasm accession numbers (e.g.linkage using the donor accession id).
SC2x	Similarity Calculated Level 2: Medium strength evidence for similarity. E.g. one institute indicating it had received an accession from the other institute, but the exact donor accession identifier was missing. The accession was matched on germplasm name instead
SC3x	Similarity Calculated Level 3: More circumstantial evidence for similarity such as germplasm with equal names, country of origin and pedigree.
SMAN	Combinations which were spotted during a manual inspection of the data and thought to be similar. These are combinations which were not assigned in any of the calculated sets.

Tools developed

Form1
version 13 Oct 2008

Cross Referencing Tool

Select New Dataset: C:\GLOBREG_2\WHEAT\USDA_WHEAT.mdb ...

Select Master Dataset: C:\GLOBREG_2\WHEAT\ICARDA_wheat.mdb ...

Select Table: USDA_ACC1 Show fields

Select Table: ICARDA_ALL_ACC Show fields

Select fields to compare:

Field 1	COMPARE	Field 2	Algorithm	Threshold value
GPG2_PEDIGREE	Add to Set	ANCEST	Levenshtein Distance	.5

Set of fields to compare: USDA_ACC1.GPG2_PEDIGREE vs ICARDA_ALL_ACC.ANCEST [Levenshtein Distance: .5]

Clear All Sets

Select which additional columns you want to include in results from New Dataset:

- ACID
- ACIMPT
- ADD1
- ADD2
- ADD3
- CITY
- CNO
- DATE_COLLECTED
- DATE_DEVELOPED
- DATE_DONATED
- FNAME
- GENUS
- GEONO
- GID

Select which additional columns you want to include in results from Master Dataset:

- ACQDATE
- ANCEST
- COLLDATE
- COLLSRC
- DONORCODE
- DUPLSITE
- GPG2_ACQDATE
- GPG2_ANCEST
- GPG2_COLLDATE
- GPG2_DONORCODE
- GPG2_INSTCODE
- GPG2_ORICTY
- INSTCODE
- LOCALACCID

Start

Output of cross referencing tool

	CIA_BCF	ACC_GPG2_LOCA	COMPARED_FLD1	VALUE_FLD1	COMPARED_FLD2	VALUE_FLD2	ALGORITHM	THRES	SIMSCORE
	BCF 48	WAB 11382	CIAT_Passport.name_standard_v2	C 46-15	ACC_MAIN.GPG2_NAME_ICIS_V2	A 4-158	Levenshtein Distance	0.5	0.57142857143
	BCF 48	WAB 11383	CIAT_Passport.name_standard_v2	C 46-15	ACC_MAIN.GPG2_NAME_ICIS_V2	A 4-159	Levenshtein Distance	0.5	0.57142857143
	BCF 48	WAB 11389	CIAT_Passport.name_standard_v2	C 46-15	ACC_MAIN.GPG2_NAME_ICIS_V2	A 4-165	Levenshtein Distance	0.5	0.57142857143
	BCF 48	WAB 11405	CIAT_Passport.name_standard_v2	C 46-15	ACC_MAIN.GPG2_NAME_ICIS_V2	A 4-185	Levenshtein Distance	0.5	0.57142857143
	BCF 48	WAB 11415	CIAT_Passport.name_standard_v2	C 46-15	ACC_MAIN.GPG2_NAME_ICIS_V2	A 4-195	Levenshtein Distance	0.5	0.57142857143
	BCF 48	WAB 12385	CIAT_Passport.name_standard_v2	C 46-15	ACC_MAIN.GPG2_NAME_ICIS_V2	C 46-15	Levenshtein Distance	0.5	1
	BCF 48	WAB 13815	CIAT_Passport.name_standard_v2	C 46-15	ACC_MAIN.GPG2_NAME_ICIS_V2	CR 94-13	Levenshtein Distance	0.5	0.5
	BCF 48	WAB 15339	CIAT_Passport.name_standard_v2	C 46-15	ACC_MAIN.GPG2_NAME_ICIS_V2	C 1016-1	Levenshtein Distance	0.5	0.5
	BCF 48	WAB 15344	CIAT_Passport.name_standard_v2	C 46-15	ACC_MAIN.GPG2_NAME_ICIS_V2	C 1414-1	Levenshtein Distance	0.5	0.5
	BCF 48	WAB 15345	CIAT_Passport.name_standard_v2	C 46-15	ACC_MAIN.GPG2_NAME_ICIS_V2	C 15	Levenshtein Distance	0.5	0.57142857143
	BCF 48	WAB 15346	CIAT_Passport.name_standard_v2	C 46-15	ACC_MAIN.GPG2_NAME_ICIS_V2	C 35-1	Levenshtein Distance	0.5	0.57142857143
	BCF 121	WAB 7868	CIAT_Passport.name_standard_v2	FOKUNISHIKI	ACC_MAIN.GPG2_NAME_ICIS_V2	FUKUNISHIKI	Levenshtein Distance	0.5	0.90909090909
	BCF 121	WAB 12392	CIAT_Passport.name_standard_v2	FOKUNISHIKI	ACC_MAIN.GPG2_NAME_ICIS_V2	FOKUNISHIKI	Levenshtein Distance	0.5	1
	BCF 124	WAB 3887	CIAT_Passport.name_standard_v2	GLORIA	ACC_MAIN.GPG2_NAME_ICIS_V2	ELONI	Levenshtein Distance	0.5	0.5
	BCF 124	WAB 8099	CIAT_Passport.name_standard_v2	GLORIA	ACC_MAIN.GPG2_NAME_ICIS_V2	LERI	Levenshtein Distance	0.5	0.5
	BCF 124	WAB 8454	CIAT_Passport.name_standard_v2	GLORIA	ACC_MAIN.GPG2_NAME_ICIS_V2	MASRIA	Levenshtein Distance	0.5	0.5
	BCF 124	WAB 8701	CIAT_Passport.name_standard_v2	GLORIA	ACC_MAIN.GPG2_NAME_ICIS_V2	GORPU	Levenshtein Distance	0.5	0.5
	BCF 124	WAB 8864	CIAT_Passport.name_standard_v2	GLORIA	ACC_MAIN.GPG2_NAME_ICIS_V2	GOKOLIA	Levenshtein Distance	0.5	0.57142857143
	BCF 124	WAB 8965	CIAT_Passport.name_standard_v2	GLORIA	ACC_MAIN.GPG2_NAME_ICIS_V2	GWARI	Levenshtein Distance	0.5	0.5
	BCF 124	WAB 9223	CIAT_Passport.name_standard_v2	GLORIA	ACC_MAIN.GPG2_NAME_ICIS_V2	GORLO	Levenshtein Distance	0.5	0.5
	BCF 124	WAB 9580	CIAT_Passport.name_standard_v2	GLORIA	ACC_MAIN.GPG2_NAME_ICIS_V2	GBLODI	Levenshtein Distance	0.5	0.5
	BCF 124	WAB 9635	CIAT_Passport.name_standard_v2	GLORIA	ACC_MAIN.GPG2_NAME_ICIS_V2	KLOUMA	Levenshtein Distance	0.5	0.5
	BCF 124	WAB 14264	CIAT_Passport.name_standard_v2	GLORIA	ACC_MAIN.GPG2_NAME_ICIS_V2	DORI	Levenshtein Distance	0.5	0.5
	BCF 124	WAB 14320	CIAT_Passport.name_standard_v2	GLORIA	ACC_MAIN.GPG2_NAME_ICIS_V2	BORIO	Levenshtein Distance	0.5	0.5
	BCF 124	WAB 14424	CIAT_Passport.name_standard_v2	GLORIA	ACC_MAIN.GPG2_NAME_ICIS_V2	SARIA	Levenshtein Distance	0.5	0.5
	BCF 124	WAB 14487	CIAT_Passport.name_standard_v2	GLORIA	ACC_MAIN.GPG2_NAME_ICIS_V2	TIORI	Levenshtein Distance	0.5	0.5
	BCF 124	WAB 14489	CIAT_Passport.name_standard_v2	GLORIA	ACC_MAIN.GPG2_NAME_ICIS_V2	GNORE	Levenshtein Distance	0.5	0.5
	BCF 124	WAB 14504	CIAT_Passport.name_standard_v2	GLORIA	ACC_MAIN.GPG2_NAME_ICIS_V2	GLABERRIMA	Levenshtein Distance	0.5	0.5
	BCF 371	WAB 8099	CIAT_Passport.name_standard_v2	NEGRIN	ACC_MAIN.GPG2_NAME_ICIS_V2	LERI	Levenshtein Distance	0.5	0.5
	BCF 371	WAB 8162	CIAT_Passport.name_standard_v2	NEGRIN	ACC_MAIN.GPG2_NAME_ICIS_V2	TEMERIN	Levenshtein Distance	0.5	0.57142857143
	BCF 371	WAB 9274	CIAT_Passport.name_standard_v2	NEGRIN	ACC_MAIN.GPG2_NAME_ICIS_V2	EX GURIN	Levenshtein Distance	0.5	0.5

Methodology

◆ Similarity matrix

CS_ALL_IRRI_USDA : Table							
IRGC	USDA	SCORE_LATLON	IRRI_NAME	USDA_NAME	SCORE_NAME	IRRI_PEDIGREE	
IRGC 2467	PI 62526		MASHIBU	MESHIBU	0.857142857142857		
IRGC 2791	PI 154459		TAICHU TOKU 157-6	TAICHU TOKU 157	0.882352941176471		
IRGC 7948	PI 154705		TAICHU 65	TAICHU 65	1		
IRGC 2805	PI 154485		PASIRANKASU 10	PASIRANKASU	0.785714285714286		
IRGC 2818	PI 154521		KONKO TAIKEI TO 20	KONKO TAIKEI TO	0.833333333333333		
IRGC 2815	PI 154512		TANGO NAKETO 11	TANGO NAKETO	0.8		
IRGC 2811	PI 154498		FUKERUPAGAI 11	FUKERUPAGAI	0.785714285714286		
IRGC 2806	PI 154487		PASHAMU 12	POSHAMU	0.6		
IRGC 10406	PI 154469		PEHBI KINKSU	PEHBI KINKAU	0.916666666666667		
IRGC 2795	PI 154463		TAICHU MOCHI 46-10	TAICHU MOCHI 46	0.833333333333333		
IRGC 2794	PI 154462		TAICHU TOKU 162-9	TAICHU TOKU 162	0.882352941176471		
IRGC 3539	PI 154488		PERIIZU	PERRIZN	0.714285714285714		
IRGC 2807	PI 154488		PERIIZU 1	PERRIZN	0.555555555555556		
IRGC 2845	PI 154644		NO IKU MOCHI 116	NO IKU MOCHI 116	1		
▶ IRGC 2848	PI 154675		TAICHU 159	TAICHU 159	1		
IRGC 2821	PI 154532		TAKENARI 31	TAKENARI	0.727272727272727		
IRGC 10413	PI 154662		SHINSHIKU 7	SHINCHIKU 7	0.909090909090909		
IRGC 2847	PI 154662		SHINCHIKU 4	SHINCHIKU 7	0.909090909090909		
IRGC 11345	PI 154533		SHIGA SHIRO	MIZUHO	0.272727272727273		
IRGC 2843	PI 154631		NO IKU MOCHI 34	NO IKU MOCHI 34	1		
IRGC 2842	PI 154627		NO IKU MOCHI 20	NO IKU MOCHI 20	1		
IRGC 2876	PI 215828		SHINCHIKU IKU 19 SELN	SHINCHIKU IKU NO 19	0.714285714285714		
IRGC 2875	PI 215828		SHINCHIKU IKU 19 SELN	SHINCHIKU IKU NO 19	0.714285714285714		
IRGC 3957	CIOR 4875		CATORSA	CATORSA	1		

Methodology

◆ Similarity matrix

CS_ALL_IRRI_USDA : Table

	IRGC	USDA	SCORE_PEDIGREE	SCORE_AVG	SCORE_DEV	SCORE_N	AVG_DIV_DEV	DUP_SELECTIO
	IRGC 2467	PI 62526		0.92857142857	0.10101525446	2	9.19238815543	SC1A
	IRGC 2791	PI 154459		0.94117647059	0.08318903308	2	11.313708499	SC1A
	IRGC 7948	PI 154705		1	0	2	999	SC1A
	IRGC 2805	PI 154485		0.89285714286	0.15152288168	2	5.89255650989	SC1A
	IRGC 2818	PI 154521		0.91666666667	0.1178511302	2	7.77817459305	SC1A
	IRGC 2815	PI 154512		0.9	0.14142135624	2	6.36396103068	SC1A
	IRGC 2811	PI 154498		0.89285714286	0.15152288168	2	5.89255650989	SC1A
	IRGC 2806	PI 154487		0.8	0.28284271247	2	2.82842712475	SC1A
	IRGC 10406	PI 154469		0.95833333333	0.0589255651	2	16.2634559673	SC1A
	IRGC 2795	PI 154463		0.91666666667	0.1178511302	2	7.77817459305	SC1A
	IRGC 2794	PI 154462		0.94117647059	0.08318903308	2	11.313708499	SC1A
	IRGC 3539	PI 154488		0.85714285714	0.20203050891	2	4.24264068712	SC1A
	IRGC 2807	PI 154488		0.77777777778	0.31426968053	2	2.47487373415	SC1A
	IRGC 2845	PI 154644		1	0	2	999	SC1A
	IRGC 2848	PI 154675		1	0	2	999	SC1A
	IRGC 2821	PI 154532		0.86363636364	0.19284730396	2	4.47834294751	SC1A
	IRGC 10413	PI 154662		0.95454545455	0.06428243465	2	14.8492424049	SC1A
	IRGC 2847	PI 154662		0.95454545455	0.06428243465	2	14.8492424049	SC1A
	IRGC 11345	PI 154533		0.42424242424	0.51693097301	3	0.82069453446	SC1A_INCORR
	IRGC 2843	PI 154631		1	0	2	999	SC1A
	IRGC 2842	PI 154627		1	0	2	999	SC1A
	IRGC 2876	PI 215828		0.90476190476	0.16495721977	3	5.48482755730	
	IRGC 2875	PI 215828		0.57142857143	0.51507875364	3	1.10940039245	
	IRGC 3957	CIOR 4875		1	0	3	999	SC1A
	IRGC 3593	CIOR 7253		1	0	2	999	SC1A
	IRGC 10576	PI 208456		0.875	0.21650635095	3	4.04145188433	
	IRGC 3859	CIOR 3880		1	0	2	999	SC1A
	IRGC 4085	CIOR 5302		0.74074074074	0.44905020937	3	1.64957219768	
	IRGC 2473	CIOR 6441		1	0	3	999	SC1A
	IRGC 2475	CIOR 6646		1	0	3	999	SC1A

Methodology

- ◆ Grouping algorithm
(equivalence relations)

$$A=B \text{ AND } C=B \Rightarrow C=A$$

Methodology

◆ Grouping algorithm (example)

GA3_2_ALL_GROUPS_FINISHED : Table

ID1	ID2	DUP_DECISION	GROUPINIT	GROUPEND	SOURCE_ACCI	D1_cod	D1_ISO	D1_org	D2_cod	D2_iso	D2_org
IRGC 6420	PI 459446	SC1A	487	487	IRGC 6420		IND	CENTRAL RICE	PHL001	PHL	International Ric
IRGC 7550	PI 459447	SC1A	488	488	IRGC 7550		IND		PHL001	PHL	International Ric
IRGC 9532	PI 459448	SC1A	489	489	IRGC 9532		ESP	ESTACION ARF	PHL001	PHL	International Ric
IRGC 8346	PI 459449	SC1A	490	490	IRGC 8346		BGD	BANGLADESH	PHL001	PHL	International Ric
▶ IRGC 3630	PI 459450	SC1A	491	491	IRGC 3630	USA007	USA	PLANT GENETI	PHL001	PHL	International Ric
IRGC 3630	PI 220749	SC1A	8059	491	PI 220749	USA007	USA	PLANT GENETI		IDN	Balai Penjelidik
▶ IRGC 3630	PI 459450	SC1A	491		491		IRGC 3630				
▶ IRGC 3630	PI 220749	SC1A	8059		491		PI 220749				
IRGC 8350	PI 459452	SC1A	493	493	IRGC 8350		BGD	DEEP WATER	PHL001	PHL	International Ric
IRGC 8350	PI 406083	SC1A	11217	493	IRGC 8350		BGD	DEEP WATER	PHL001	PHL	International Ric
IRGC 8350	PI 414232	SC1A	11218	493	IRGC 8350		BGD	DEEP WATER	PHL001	PHL	International Ric
IRGC 1397	PI 459453	SC1A	494	494	IRGC 1397	USA007	USA	PLANT GENETI	PHL001	PHL	International Ric
IRGC 1397	PI 160806	SC1A	4362	494	PI 160806	USA007	USA	PLANT GENETI			
IRGC 9225	PI 459454	SC1A	495	495	IRGC 9225		IDN	CEREALS RES	PHL001	PHL	International Ric
IRGC 8189	PI 459455	SC1A	496	496	IRGC 8189		TWN	TAIWAN PROV	PHL001	PHL	International Ric
IRGC 8227	PI 459456	SC1A	497	497	IRGC 8227		TWN	TAIWAN PROV	PHL001	PHL	International Ric
IRGC 8227	PI 389223	SC1A	11170	497	IRGC 8227		TWN	TAIWAN PROV	PHL001	VNM	International Ric
IRGC 8227	F 1051	SC1B	12139	497	IRGC 8227		TWN	TAIWAN PROV	PHL001	PHL	International Ric
IRGC 5014	PI 459457	SC1A	498	498	IRGC 5014		PHL	MALIGAYA RIC	PHL001	PHL	International Ric

Preliminary results

Rice

- ◆ Out of 200,000 accessions analysed from 5 collections, almost 40,000 (20%) had 1 or more similar accessions replicated in other collections

Next Steps

- ◆ Wheat: final analysis needs to be concluded followed by integration into IWIS 3
- ◆ Rice: preparation of data for integration in IRIS have just started

Integration of data into IRIS

Difficulties encountered sofar

- ◆ Recursive nature of GID-GPID1,2

A need for visualisation tools

- For data entry
- To correct errors

Integration of data into IRIS

Problems encountered sofar

◆ Errors in IRIS data

- Over a period of time data were entered by a lot of different individuals
- “Rules” might have changed over time, misunderstood
- Indicates a certain weakness in Data QC

Integration of data into IRIS

Data QC

- ◆ Review workflow
 - Procedures (how)
 - Decision points (where and who)
 - Reporting (feedback)
- ◆ How can the application help prevent errors (“making mistakes should require an extra effort”)
- ◆ What type of tools are useful to retro-actively check data
- ◆ Where exists a need to incorporate reporting facilities to ensure adequate feedback