IRTP 456:

IRRI

- 15 GIDs in one derivative neighbourhood
- 48 errors in GERMPLSM, preferred name and preferred ID
- 9 missing GIDs

GERMPLASM												Preferred name					
GID	METHN	GNPGS	GPID1	GPID2	GERMUID	LGID	GLOCN	GDATE	GREF	GRPLCE	MGID	NID	N_GID	NTYPE	NSTAT	NUSER	PrefName
-1	Unknown	-1	0	0			New Delhi	0	GRIN	0		-1	-1				
521463	255	-1	83862	83862	30	-4459	101	19510701	26	0	0	652665	521463	6	1	30	T 3
	62		-1	-1			NPGS										
675489	62	-1	83862	1487	10	-675489	9016	0	1	0	675489	840141	675489	6	1	112	T 3
			-1	521463													
83862	203	-1	0	0	10	-83862	101	19510000	2	0	0	513726	83862	6	1	8	T 3
	62		-1	675489			INGER	19750000			83862						
-2	Unknown	-1	-1				9000			0		-2	-2	6			T 3
-3	62	-1	-1	-2			INGER	19881110		0	83862	-3	-3	6			T 3
-4	Unknown	-1	-1				10516			0		-5	-4	6			T 3
-5	62	-1	-1	-4			INGER	19890111		0	83862	-6	-5	6			T 3
-6	Unknown	-1	-1				9000			0		-8	-6	6			T 3
-7	62	-1	-1	-6			INGER	20020527		0	83862	-9	-7	6			T 3
513408	62	-1	83862	1487	30	-19186	102	19751113	26	0	0	636951	513408	6	1	30	T 3
			-1	675489			NPGS										
-8	Unknown		-1				10519					-11	-8	6			T 3
347943	62	-1	83862	83862	6	-99881	9016	19770321	12	0	347943	513729	347943	6	1	112	T 3
			-1	-8													
-9	Unknown		-1				10225					-13	-9	6			T 3
351334	62	-1	83862	83862	6	-106669	9016	19620407	12	0	351334	513730	351334	6	1	112	T 3
			-1	-9													
351466	255	-1	83862	83862	4	-106993	101	0	4	0	0	513731	351466	6	1	4	T 3
1487	62	-1	83862	521463	10	-1487	102	19760000	1	0	0	513725	1487	6	1	1	T 3
			-1	675489			INGER	19900103									
257064	62	-1	83862	521463	13	-20099	2295	19510000	11	0	0	625575	257064	6	1	1	T 3
	206		-1				NPGS										T 3 (Selection)
339650	62	-1	0	0	6	-81802	9016	0	12	0	339650	1592107	339650	22	0	112	PI 197612-1
			-1	257064													

IRRI

Errors in IRIS?

- **Error rate**
  - 1,000,000 - 10,000,000 errors in 2 GMS tables GERMPLSM and NAMES
  - 100 - 1,000 per day since IRRI started computing 30 years ago
- **Correcting errors**
  - 5 hours to identify 50 corrections for IRTP 456 neighbourhood
  - 50-500 person-years to identify corrections for current errors
- **What next?**
  - Tom H to speed up data correction rate
  - Improve validation of new data input
  - Lessons for other ICIS implementations?

IRRI

Validation of IRIS data

- **Setting the environment**
- **Retroactive validation and correction**
  - Finding out what's wrong
  - Deciding what is correct
  - Correcting the error
- **Validation of new data entry**
  - Documenting standards
  - Training data encoders
  - Software validated data entry

IRRI

Setting the environment

Done (more or less):

- **Unambiguous definitions of terms**
- **Consistent SOPs designed and enforced (*?*) across all users**
- **Data validation rules**

Still to come:

- **Data validation procedures**
- **ICIS conceptual problem:**
  - Designed recognizing that names aren't IDs
  - Connects assuming names are IDs
  - Doesn't document what germplasm each GID represents

IRRI

What does a GID represent?

- **Cross – usually no living seed**
  - MethN=cross
- **Collected sample – usually no living seed**
  - MethN=collected, associated data on collecting process
- **Accession conserved in genebank GLOCN**
  - GID has one name with NTYPE=1, NSTAT=8, NLocN=GLocN; no other GID has that NVAL at that GLocN
  - MethN usually =imported with GLocN≠GPID2.GLocN, sometimes = selected with GLocN=GPID2.GLocN
- **…**
- **But rules usually aren't followed**

IRRI

Can a name be a germplasm ID?

- **NSTAT = 8: NVAL uniquely identifies at GLOCN**
- **SOP: germplasm → new organizational unit:**
  - new GID
  - GDate=transfer date
  - GLocN=LocID of receiving organizational unit

  → GLocN = where germplasm is managed from birth
- **Confirmation:**
  - By type of MethN?
  - By "Represents"?

- **→ unique identification by NVAL + NSTAT=8 + GLOCN + METHN/Represents**

IRRI

What does a GID represent?

Represents	Comment	GID
Original sample		-1
Accession in NPGS		521463
Accession in IRGC		675489
Original IRTP line		83862
	Guess from GRC. Guess must be part of 1st year of IRTP	
Source of old IRTP line		-2
Old IRTP line		-3
Source of old IRTP line		-4
Old IRTP line		-5
Source of current IRTP line		-6
Current IRTP line		-7
Accession in NPGS		513408

Azucena

74 GIDs, 27 missing, 10 can't work out what they are, > 60% of GPID1-GPID2 values wrong

Represents	Observations
Nothing - stop using	Derivative single plant selection!!! Especially wrong for notional ancestor of all Azucena!
INGER entry	(a) INGER must have got it from GRC not the other way round because IRTP started 1975, IRGC328 is 1960 (b) note IRTP 4209 not in current INGER.
? Philippine sample? CIAT sample?	Must be error! Can't be original collection line and have GPID2=IRTP 806! NREF=198 is CIAT 1998 data.
?	GID has attribute Note Perlegen set. Why this extra GID? Is it the one Ken uses for Perlegen data?
Original collection line?!!!	Must be error! Can't be original collection line and have GPID2=IRTP 4209! Note 1955 is the year USDA got PI 223554 from UPLB - but unlikely to be year of collection. Delete? But it's documented as the source of IRTP 806!
INGER entry	Guess INGER must have got it from GRC. But note date 1986 - cannot be original date. Must be replacement. Replacement from where? Ed says INGER records do not show
GRC seed increase	Corrected to match GRC source
GQ lab's sample, purified stock obtained from Susan McCouch via Stuttgart	"GID represents" based on email from Grace. Better if it had an ID. Wrongly documented as IRGC 328 with NTYPE 1. Should be NTYPE 10.
K McNally's sample	"GID represents" based on email from Digna. Better if it had an ID
IRGC accession corresponding to Azucena used for Perlegen	IRGC accession corresponding to Perlegen set. Why parallel to 2389418, not in same lineage?

IRRI

What does a GID represent?

- **New table in IRIS-GRC local?**

Field	Description	Levels
GID		
Represents	Class of Germplasm	Unvalidated
		Inconsistent data
		Sample collected from field or market at GLocN
		Accession conserved in genebank at GLocN
		Cross made at GLocN
		Breeder's selection or other line produced at GLocN
		Sample maintained at GLocN for testing in nurseries
		Copy of a genebank accession or breeder's material held informally at GLocN, away from the original genebank/breeder/network
VReason	Text justification of classification	
VTimeStamp	Date-time validated	
Vref	Reference	
VUserID	User validating	

IRRI

Identifying errors

- **Go back to original source databases**
  - IRGCIS, INGERIS, GRIN, CIAT, WARDA (IITA)
  - Current updated originals or original imported datasets
- **Identify which existing GIDs represent the germplasm**
  - GRef / NRef / ARef
  - Fuzzy inspection of data
- **Cross reference source with IRIS**
  - Name / donor / origin data

IRRI

Simple changes

- **Error: 2849 NVAL are species name**
  - Most with NTYPE = cultivar name
  - 2848 with NSTAT = preferred name
    - → conflicting reports on taxonomy
- **Correcting:**
  - SQL to retrieve distinct users
    - Notify users they are not following the rules
  - SQL to retrieve distinct values
  - Manually construct table of species attribute (1141) values corresponding to each distinct value
  - SQL to delete names and add GID attribute

IRRI

Validating NLocN:
New table in IRIS-GRIMS local

Code	With hyphen?	Type	Name Location	Example
IRGC	No	Accession	IRRI GRC	IRGC 123
PI	No	Accession	USA NPGS	PI 123
TOS	No	Accession	IITA	TOS 123
TOG	No	Accession	IITA	TOG 123
IRTP	No	Test entry	IRRI INGER	IRTP 123
IR	No	Cross	IRRI PBGB	IR 8
IR	No	IRRI released line	IRRI PBGB	IR 8
IR	Yes	IRRI breeding line	IRRI PBGB	IR 8-2-3

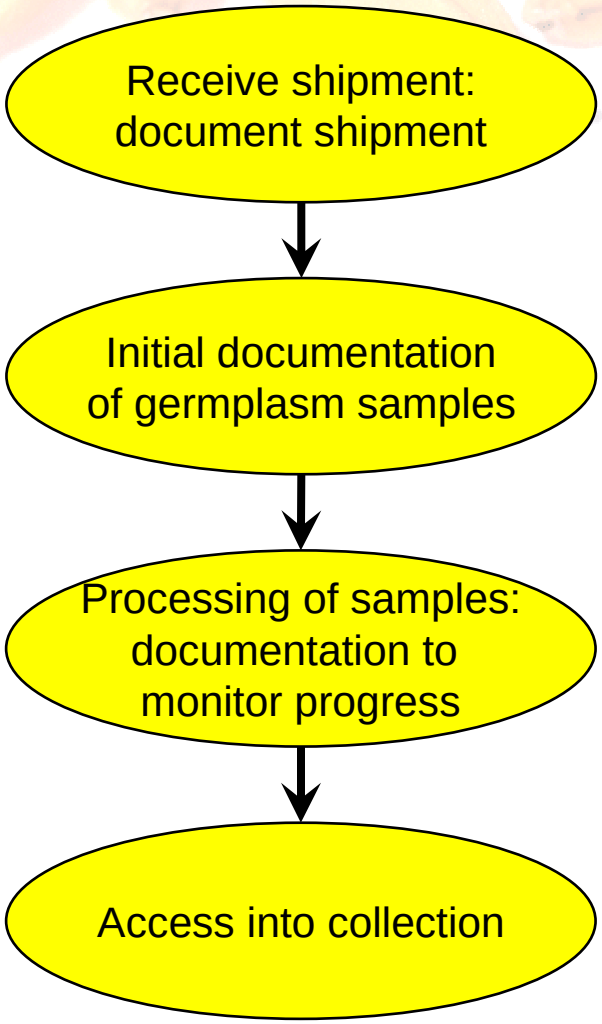
IRRI

Setgen??

- **Purpose =**
  advancing a generation within a breeding programme
  - Good
- **Purpose =**
  creating GIDs for germplasm received from others
  - Not good!
  - Generalise code further?
  - Start again?

IRRI

Workflow for incoming samples



Assign batch ID, date of receipt
Donor: person, institute, address
Admin: MTA, phytosanitary certificate

Create GIDs
Assign temporary ID
Enter passport data

Check data and seed for duplication
Quarantine, test health, germination
Prepare for conservation:
initial seed increase, test, clean, dry, pack, label

Assign IRGC accession number, MLS date

IRRI

Receiving shipment

Batch ID & Date

New list
- ID auto-assigned
- Date user editable, default=current

Donor
(person, institute, address)

User-guided assignment
- Drop-down selection of existing records
- Add new records
- Needs new table structures (Monday)
  - LOCID in INSTITUT
  - ADDRESS table

IP and other summary data

User-guided assignment

IRRI

Initial documentation of samples: GERMPLSM data

- **Create GID:**
  - GID = member of Batch list
  - GMethN = import
  - GNPGS=-1
  - GDate=Batch date = date of receipt
  - GLocN=Batch LocID = donor's LocID
  - GERMUID=
    - "User ID of the breeder of the current germplasm" (!)
  - GREF=?
  - GRPLCE=0
  - MGID=GID
  - GPID2 = ?
  - GPID1 = ?

IRRI

Initial documentation of samples: Assigning GPID2

- **Search for existing GID representing donor's sample**
  - Perfect match (can be accepted by encoder if data correct):
    - Donor provides own unique accession ID
    - IRIS has exactly 1 NID with
      - NVAL=donor's unique ID
      - NSTAT=8 (but …)
      - GLOCN=donor's locid (but …)
      - ? "GRepresents" says it represents the donor's sample
  - Fuzzy match (not acceptable by encoder)
    - Fuzzy name matches
    - Allowing for wrong data in most GIDs and NIDs
  - No match
    - Create GID for donor's sample based on donor's data
    - Search for GID representing original sample
      - Equivalent search methodology

IRRI

Initial documentation of samples: Fuzzy matching in search for GPID2/GPID1

- **Display key data for all names where NVAL = (exactly or approximately) donor-provided value**
  - Name data: NVal, Date, Type, State, User, Location, Reference
  - Preferred name data (if different): same fields
  - Preferred ID data (if different): same fields
  - Germplasm data: Creation Method, n parents, User, Location, Date, Reference
  - All attribute data for GID and NID
  - GPID1 - corresponding germplasm, name and attribute data

IRRI

Initial documentation of samples: Assisted creation of name data

- **IRGC name/ID:**
  - Assign automatically
- **Name data provided by donor**
  - Donor's sample ID, ID of original collected sample, preferred (e.g. cultivar) name, other IDs, abbreviations, non-English script
  - Name inherited from pre-existing GPID2 or GPID1:
    - Automatic assignment of NType, NStat
    - Inherited NLocN, NDate
    - Opportunity to edit if donor's data show something different
  - New name
    - Guided assignment of NType, NStat, NLocN, NDate

IRRI

Initial documentation of samples: Assisted entry of other passport data

- **ICIS categories:**
  - Collecting or breeding locations (→ GLocN of GPID1)
  - Collecting mission
  - Attributes:
    - Taxonomy
    - Info on nature of original sample and collecting site
- **New GIDs for GPID2 and/or GPID1**
  - Guided assignment of attribute data
  - Selection of existing or creation of new locations
- **Already existing GID for GPID2 or GPID1**
  - Display existing attributes
  - Opportunity to edit, correct, add