Crop registries
and
Location data management

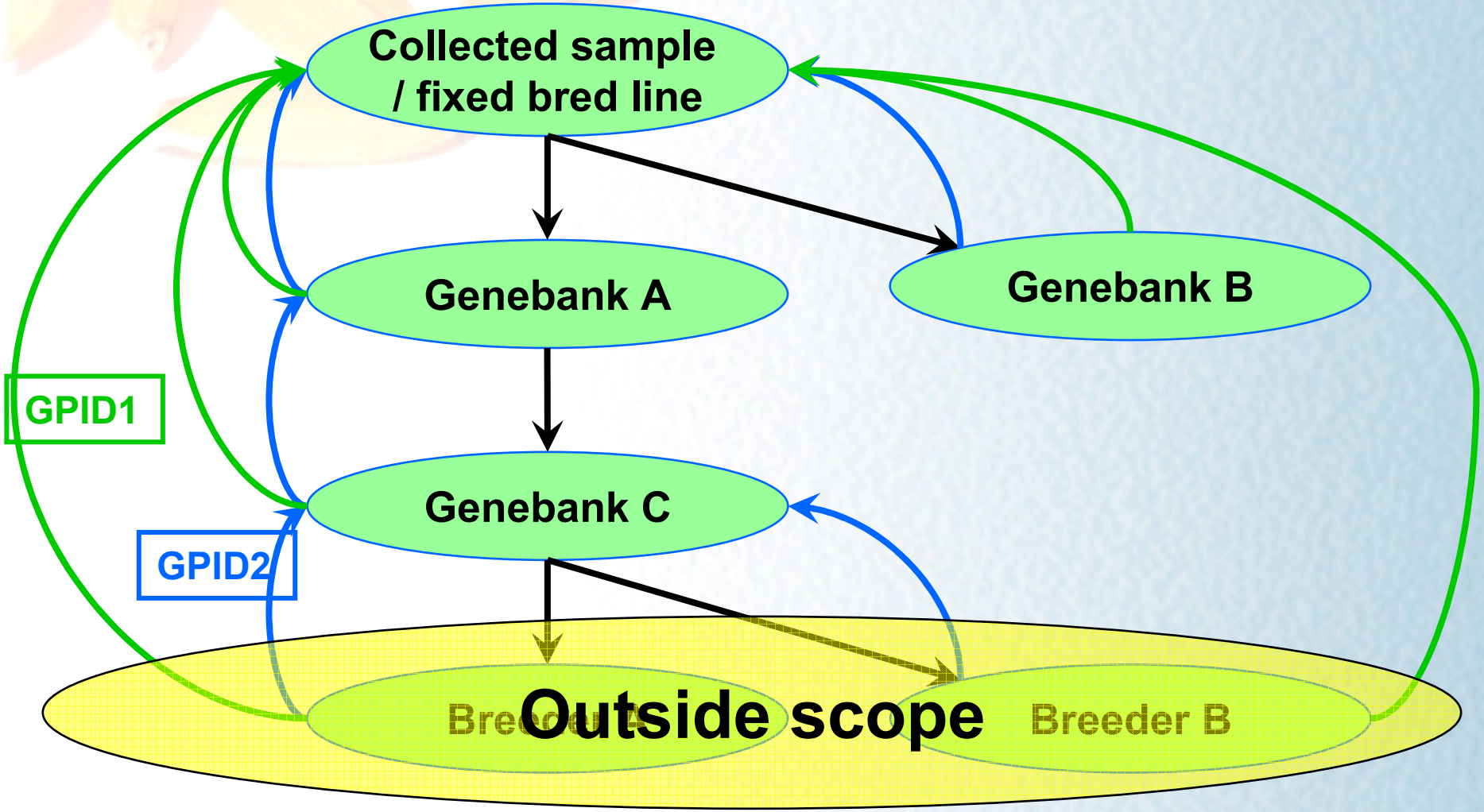
IRRI

Crop registries – "GPG2" project

- **Assemble data on genebank accessions from N wheat and rice genebanks**
  - Tom Hazekamp?
  - Harvesting data via web services where possible, or by any other possible means
- **Coverage:**
  - Original aim: ~80% of global holdings documented
  - Revised aim: ~25% of global holdings documented
- **Key feature missing from previous projects:**
  - Data curation to identify equivalent accessions
    - In GCP-speak = identify maintenance neighbourhoods

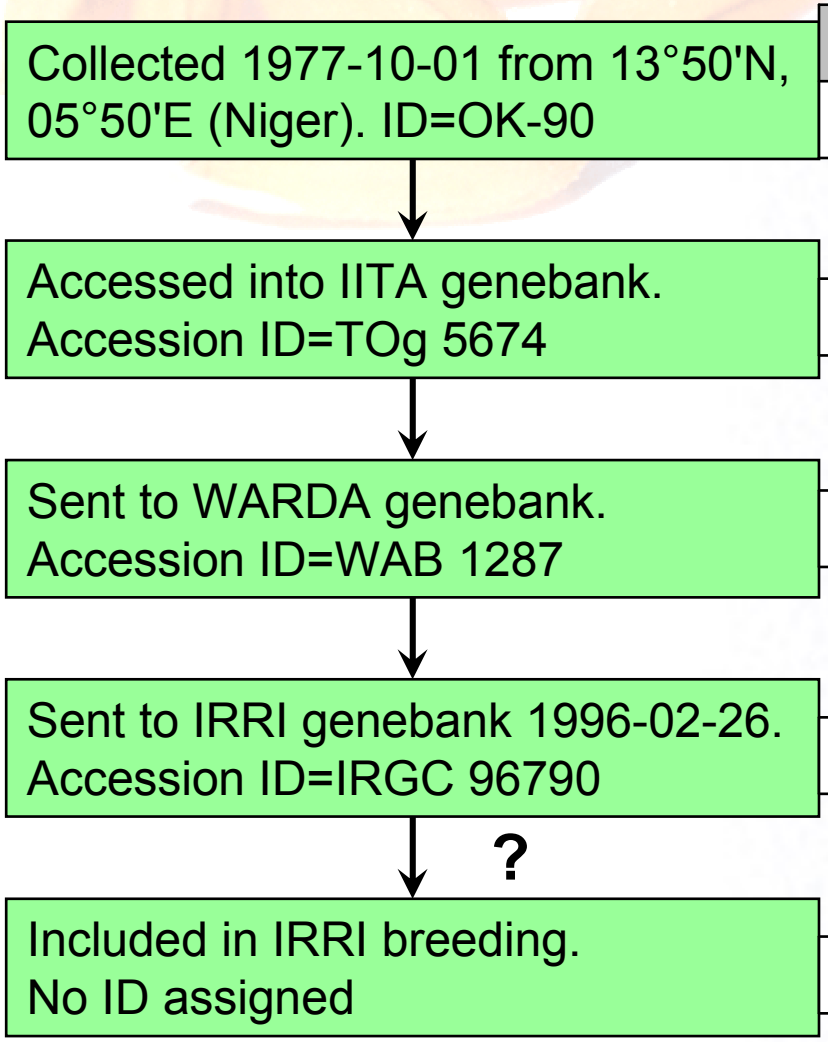
IRRI

Maintenance neighbourhoods in ICIS



IRRI

TOg 5674



GID	GPID1	GPID2	GLOCN	GDATE
-1	0	0	-10000	19771001
-2	-1	-1	9002	0
-3	-1	-2	9001	0
-4	-1	-3	9016	19960226
-5	-1	-2?-3?-4?	9000?	0

IRRI

Issues (1)
Rigour of GPID1 definition

- **GPID1 = cross for breeding lines**
- **GPID1 not clearly defined for germplasm that cannot be traced back to a cross**
  - Wrong tendency (continuing despite ICIS 2005) to create GIDs for landraces
    - e.g. 27 GIDs in IRIS with name “Azucena” all have same GPID1 2285994
- **Recommend**
  - GPID1 = cross or collected sample
  - Reclassify “collected sample” method as generative?
    - Recognizing that farmers are plant breeders

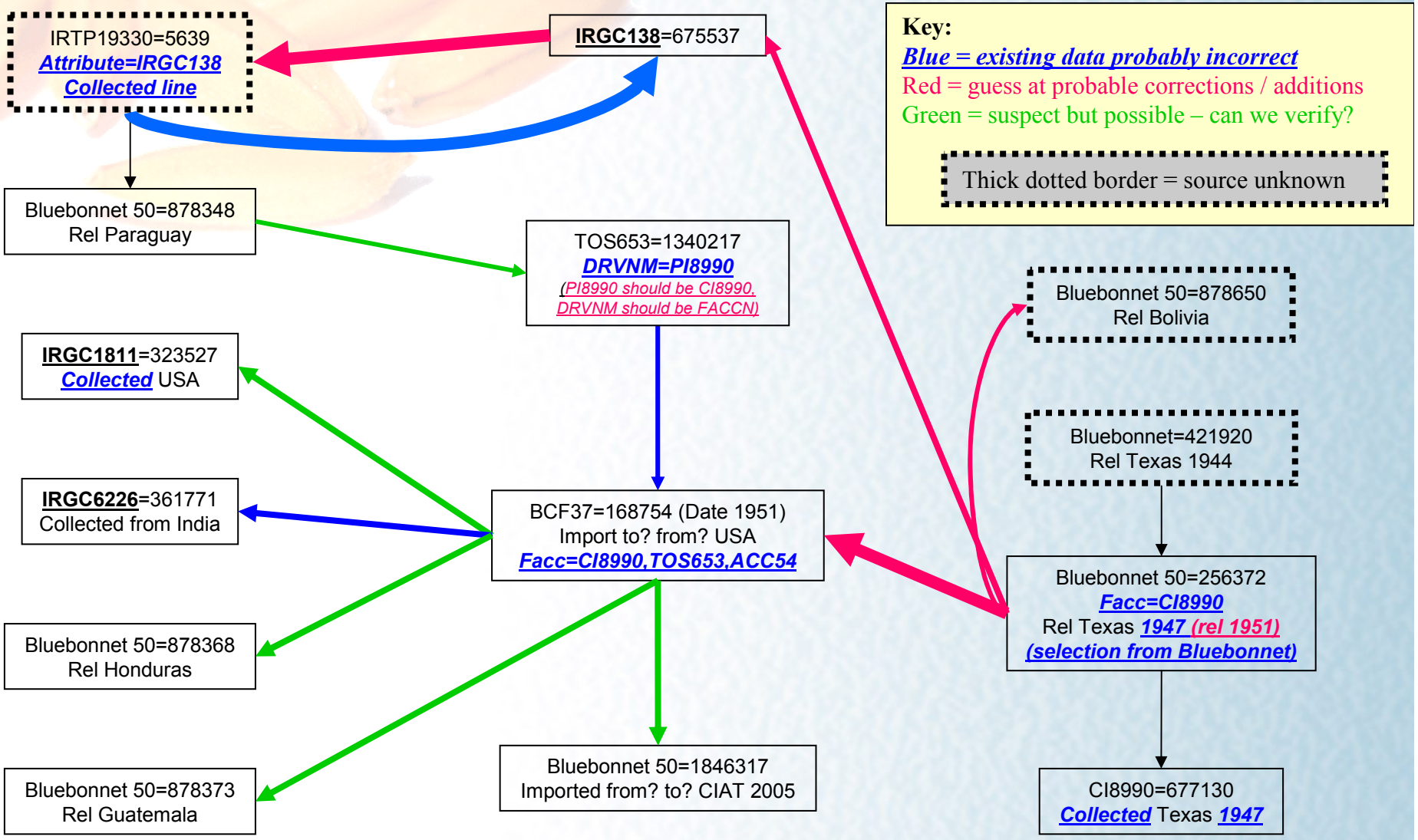
IRRI

Issues (2)
Handling uncertainty

- **Germplasm Creation Methods for "probably offspring of …"?**
- **Germplasm Creation Methods for "known maintenance neighbourhood, unknown donor"?**
  - e.g. PBGB sample of TOg 5674 from IITA, WARDA or GRC?
- **Structures to document multiple possibilities?**
  - e.g. PBGB sample of TOg 5674 from IITA, WARDA or GRC?

IRRI

Bluebonnet 50 in IRIS: Common Group = Rexoro/Fortuna=382072 : 98 GIDs



IRRI

Issues (3)

Need to identify what each GID represents

TOg 5674:

G Represents	GID	METHN	GPID1	GPID2	GERMUID	GLOCN	GDATE
Accession at IITA	1347273	255	0	0	77	9002	19771001
Accession in GRC	1204383	62	537163	537163	70	108	19960226
Line held in PBGB	537163	256	0	0	1	0	0

Not recorded in ICIS

IRRI

Corrected TOg 5674:

G Represents	GID	METHN	GPID1	GPID2	GERMUID	GLOCN	GDATE
Original collected sample	-1	69	0	0		-100	19771001
Accession at IITA	1347273	62	-1	-1	77	9002	0
Accession at WARDA	-2	62	-1	1347273		9001	0
Accession in GRC	1204383	62	-1	-2	70	9016	19960226
Line held in PBGB	537163	62	-1	1204383 ?	1	9000	0

IRRI

G_Represents:
What type of germplasm does the GID represent?

- **Accession securely conserved long-term at GLOCN**
- **Breeding line maintained at GLOCN as long as it's used for breeding / trials**
- **Historical germplasm now represented only by progeny**
  - Segregating population
  - Founding sample of Maintenance Neighbourhood
    - Seed sample collected from GLOCN
    - Cultivar released in/at GLOCN
  - Founding sample of Derivative Neighbourhood
    - Cross
  - Founding sample of Generative Neighbourhood
    - Presumed/virtual landrace progenitor
- **Germplasm without progeny persisting only for duration of study at GLOCN**
  - DNA/leaf/tissue/experimental sample

IRRI

Issues (4)
Location of what when?

- **IRIS**
  - A germplasm sample (GID) has a location and date
  - A germplasm name (NID) has a location and date
- **Reality**
  - A germplasm sample has a location and date where
    - it was harvested (biological genesis)
    - it was assigned a GID (ICIS genesis)
    - its man. neighbourhood was born (MCPD: holding institute)
    - its parent man. neighbourhood was born (MCPD: donor)
    - its maintenance or derivative neighbourhood was born (MCPD: origin)
  - A germplasm name has a location and date where
    - an inherited name was first given to an ancestor
    - a name was associated with the germplasm

IRRI

Locations and dates defined:

http://cropwiki.irri.org/icis/index.php/TDM_Genealogy_Management_System_5.4#GERMPLASM_TABLE_.28GERMPLSM.29

- **GLOCN**
  - Location where the germplasm was created as a distinct unit of management with a new GID
- **GDATE**
  - Date on which the germplasm was created as a distinct unit of management with a new GID
- **NLOCN**
  - Identifier for the location where the name was first assigned to the maintenance neighbourhood of which this GID is a member
- **NDATE**
  - Date on which the name was first assigned to the maintenance neighbourhood of which this GID is a member

IRRI

Types of germplasm location

- **Location (geographical) where sample was collected from field or market**
- **Location (institute) where sample was grown / managed**
- **Location (in cold store) where sample is stored**

IRRI

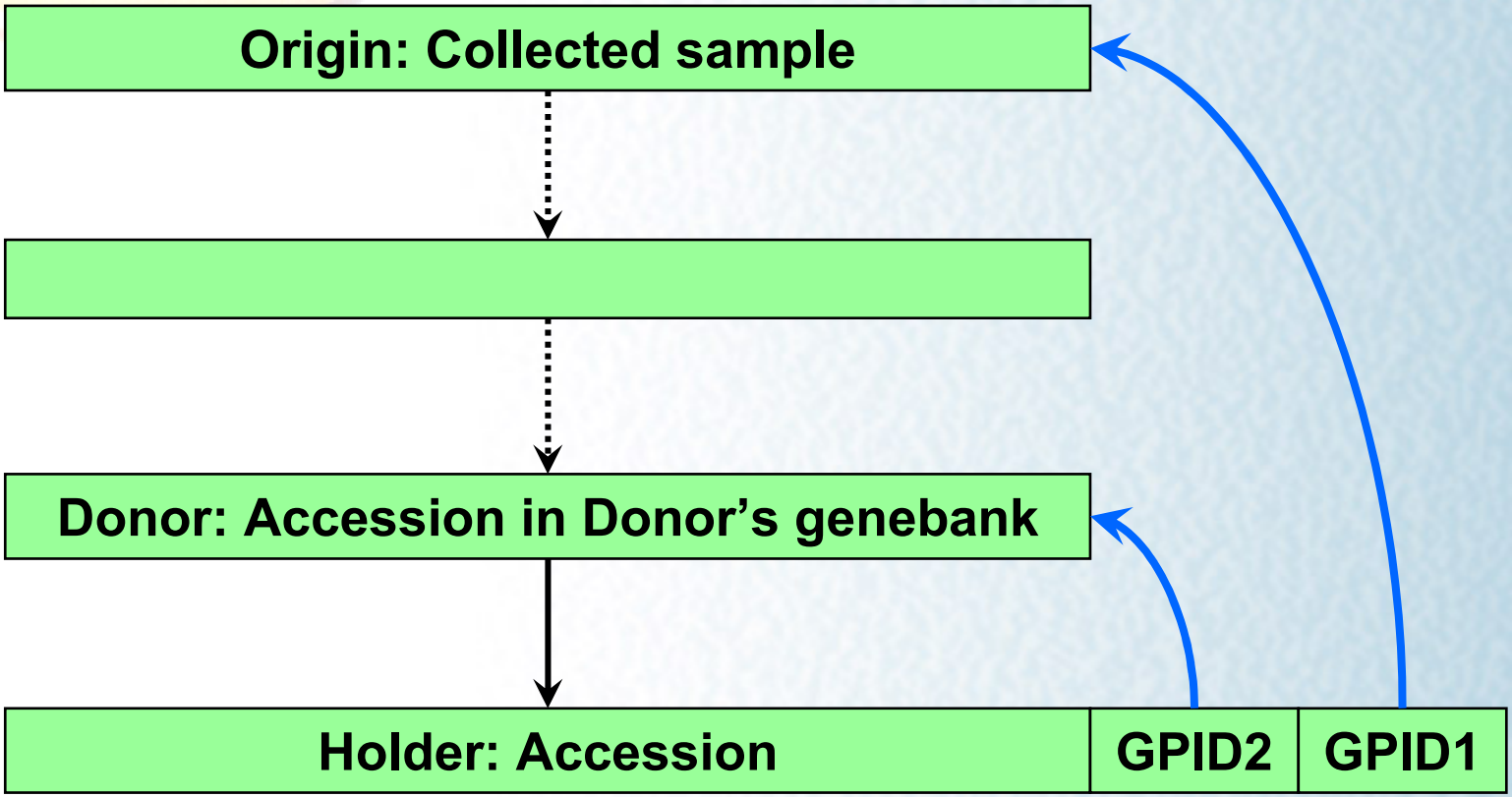
Issues (5)

How to enforce adherence to definitions?

NID	3861511	3840368
GID	2451818	2441088
NTYPE	6	1
NSTAT	1	0
NUID	231	146
NVAL	KAY NOI LEUANG	IRGC 10658
NLOCN	9000	9016
NDATE	20080124	0
NREF	0	0

IRRI

With good definitions / training / curation, ICIS can hold all locations with 1 loc/GID



IRRI

Germplasm collecting locations

- **Defined by one or more of:**
  - ***In*** an admin region / country
  - ***Near*** a populated place (or near a named place?)
  - Text description of locality
    - 2km E of Los Baños
    - 3km along highway from A to B
    - Near edge of lake X
  - Point coordinates (lat-long-alt) from
    - GPS / altimeter
    - Map / atlas
    - Interpreted directions + gazetteer

IRRI

Collecting location use cases

- **Case 1: Know validated country and/or snl1 and/or snl2 and/or town; no text description, no lat-long-alt**
  - GLOCN points to LOCID of smallest-scale non-missing place name (country, snl1, snl2, town)
  - LNAME = name of smallest-scale non-missing place name
  - Larger-scale place names through pointers in LOCATION
  - lat-long of LOCID from gazetteer
- **Case 2: Have text description or unverifiable place name and/or lat-long and/or alt**
  - GLOCN points to LOCID for collecting location
  - LNAME blank if no text description / name
  - Place names through pointers in LOCATION

IRRI

Collecting location issues (1) Near place

- **No field in Location for “nearest (populated?) place**
- **No location types for**
  - populated place (city, town, village)
  - other named features (lake, mountain, river)

IRRI

Collecting location issues (2) Georeferencing (GEOREF)

- Accuracy / certainty
  - How close to collecting location?
  - How close to field where harvested?
- Supplied display precision (d, dm, dms, dms.s)
  - Conversion to d.d loses information on display precision
- Source
  - Germplasm provider
  - Place name + gazetteer
  - Text description + gazetteer
  - GPS / altimeter
  - Map / atlas
- Datum
  - Should be / almost never is supplied with datum!
  - Assume WGS84?

IRRI

Collecting location issues (3) Variant names

- **IRIS provides for only one LNAME per LOCID**
  - No allowance for
    - Variant spellings
    - Language differences
    - Temporal changes
    - Errors by data provider
    - Conventions followed by data provider
- **Consequences**
  - Suitable only for display
  - Not suitable for searching or validation

IRRI

Collecting location issues
(4) Interpreted vs original location data

	NO.	VARIETY	TYPE	LOCATION
	179	SELASIH	W	BENUTAN
80217	180	PULUT KABANG	/	BRUNEI
88411	181	BUNTU	U	TEKALIT
	182	BANGGAN	U	KEBUBOK
78707	183	KUJAM	U	AMO A
	184	HANTU	W	SENUKOH
78712	185	MET	U	KEBUBOK
80212	186	MAYANG MERAH	U	KENUA
	187	SEBAKUT	W	PARIT
78726	188	TEMANI	U	KEBUBOK
80201	189	DAMIT	U	AMO A
80226	190	SAYU	U	KENUA

Accno 78712

Accno 78726

The only Kebubok known to gazetteers or Google is a stream in Brunei

IRRI

Chinese Rice Varieties Introduced by the
Visiting U.S. Plant Science Group
from Mainland China
September, 1974

Variety name	Synonym	Source (prov./city)	Description* given	IRRI Acc. No.
珍珠矮11 Chen-chu-ai 11	Pearl Dwarf 11	Kwangtung	Med., early crop	Acc. 28463
秋二早 Chiu-er-tsao		"	Early, late crop	
秋塘早 Chiu-tang-tsao		"	Early, late crop	Acc. 28472

Recognized variant spelling Guangdong province

IRRI

Collecting location issues
(4) Interpreted vs original location data

- **Interpreted / standardised data needed for:**
  - Normal users
  - Georeferencing
- **Need original data for (re-)validation**
- **LOCATION and GEOREF suitable for interpreted data**
- **Use LOCDES to record**
  - Original data
  - Reason for deviance
- **DTYPE?**
- **Format?**