GRC Location data:
Current status and next steps

Contents:

Two components:

1. Correcting and organizing IRGCIS location data into IRIS format
2. Correcting germplasm relationships through GID, GPID1 and GPID2, to provide the correct GIDs on which to hang the location data.

1. Correcting and organizing location data

1.1 IRIS design issues

The following are probably the minimal design changes required to enable IRIS to accommodate IRGCIS data:

- A new location type (LTYPE 413) for a populated place (city, town, village…). (For discussion: do we need more, e.g. to distinguish between cities, towns and villages? Or would that leave us with an impossible data curation task, best left to the geographer & GIS?)
- A new field (PPLID) in LOCATION to hold the locid of the nearest town or village to a location (for discussion: is this enough? IRGCIS has separate fields for town and village - should we have 2 PPLIDs, or shall we use the SNL3ID for one?)
- A new field (LLPrec) in GEOREF to indicate coordinate precision (i.e. the precision with which latitude/longitude coordinates are displayed by default, not to be confused with accuracy, which refers to how accurately the coordinates represent the true position / centre of a location). I have proposed valid values “D”, “DM”, “DMS” or “DMS.S”

1.2 Data already included in IRIS

Countries: all in IRIS, with lat-long data from CIA.

ADM1 divisions (“provinces”):

- all current regions with FIPS AAnn codes, with lat-long data from Geonames where available

- all historical regions with FIPS AAnn codes existing at the time IRIS locations were last updated

1.3 Data conversion completed but not in IRIS

\\netwin\cril\LDMS\GRC-IRIS\CreateFinalMap.mdb contains reorganized data. Key tables containing data for IRIS are:

- Table ***NewUDFLDS***. Contains the definition of the proposed new location type (413).
- Table ***NewLocations***: contains records for new locations to be added to LOCATION and GEOREF in IRIS.
- Table ***LocationChanges***: contains corrections to be made to existing LOCATION and GEOREF records in IRIS
- Table ***FinalMap*** lists all Accession IDs, with values, where available, for GLOCN, GDATE and METHN of their GPID1.
    - Non-null GLOCN means location data are clean for the accession (I hope!). Null GLOCN means location data not checked / converted
        - GLOCN=LOCID of ORI_COUNTRY for 49,217 accessions with no data on province, district, town, village, latitude, longitude or altitude. Includes values of ORI_COUNTRY that refer to continents or regions, e.g. Africa or Central America, in table NewLocations with ftype 401 or 402
        - GLOCN=LOCID of PROV for 16,506 accessions with validated data on ORI_COUNTRY and PROV but with all missing data for district, town, village, latitude, longitude and altitude.
            - Mostly = locid of ADM1 already in IRIS
            - Some are regions not ADM1 divisions (e.g. Manchuria, Mindanao, in table NewLocations with ftype 403)
            - Some are ADM2 divisions (e.g. Bontoc, in table NewLocations with ftype 407)
            - Some are populated places (e.g. Biratnagar, in table NewLocations with ftype 413)
        - GLOCN=LOCID of collecting locations assigned to 2,350 IRGC accessions when first loaded into IRIS. These are locations where (a) IRGCIS has non-missing data for at least one of latitude, longitude and altitude and (b) there is an exact match between the April 2005 IRGCIS and 1999 IRIS: ALT = GEOREF.ALT and (VILLAGE.TOWN.DISTRICT matches LNAME and PROV matches SNL1ID.LNAME; or VILLAGE.TOWN.DISTRICT;PROV matches LNAME)
    - METHN = 69 if IRGCIS contains any data showing the original sample was collected from the field. Null METHN means there's no such data in IRGCIS and we have to look outside IRGCIS for a METHN (e.g. IRIS: in the case of breeding lines, hopefully the GPID1 in IRIS already points to the correct cross).
    - GDATE = collecting date if recorded in IRIS (= a subset of accessions with METHN=69). Null METHN means we have to look outside IRGCIS for a GDATE e.g. IRIS: in the case of breeding lines, hopefully the GPID1 in IRIS already points to the correct cross).

In addition to the above tables, there are queries to check / verify:

- Query ***VerifyLocations*** shows IRGCIS location data beside IRIS data, to check whether IRGCIS and IRIS are showing the same information. Note: it takes IRIS data only from the live IRIS, not from the location changes and additions in NewLocations and LocationChanges
- Query ***VerifyGPID1Data*** tabulates the number of accessions that have same or different values for METHN and IRIS in FinalMap and the live IRIS

1.4 Next steps

Who	What
Graham / Robert / Ella / Adel	Discuss design proposals: • LTYPE 413=populated place. • Field PPLID in LOCATION • Field LLPrec in Georef
Arllet? William?	Implement above changes if agreed. • Add record in NewUDFLDS to UDFLDS • Add field PPLID to LOCATION • Add field LLPrec to GEOREF
Adel	Check my corrections, particularly where I think IRGCIS is wrong, and particularly for "province" that I have not been able to find. Where necessary, cross-reference against original documents
William?	After Adel has finished checking, add new locations to IRIS using table NewLocations
William?	After Adel has finished checking, correct existing locations using table LocationChanges
Adel	Separate place names from locality descriptions in IRGCIS: • Create new field LOCALITY in IRGCIS. • Populate LOCALITY field with concatenated locality descriptions (e.g. "LAI, 10KM S. OF BERE", "NEAR LAKE OF LERE") currently in DISTRICT, TOWN, VILLAGE • Replace locality descriptions in DISTRICT, TOWN and VILLAGE with just place names
Robert's GPG2 PDF?	Assemble reference data on "districts" (ADM level 2) – from www.statoids.com? (the Geonames gazetteer assembled by Yunlong is missing too many ADM2 locations to be useful). Add all ADM2 divisions to IRIS? Or just for countries where we have district data to be verified?
?	Check and validate Isaiah's corrections and Yunlong's corrections for all other accessions, i.e. those with non-missing values for at least one of district, town, village, locality, latitude, longitude and altitude.
?	Check which locations already have locids defined in IRIS
	Correct location definitions of existing locations in IRIS
?	Add location definitions for locations not yet defined in IRIS
?	Add GLOCN to the corresponding accessions

2. Correcting germplasm relationships through GID, GPID1 and GPID2

For each genebank accession,

- GPID2 should point to the GID representing the donor's sample
- For accessions derived from samples originally collected from the field (farm, market or wild population), GPID1 should point to the GID representing the collected sample.

IRIS contains many errors in these linkages. We must correct those linkages before we can store donor details and collection details in IRIS.

I corrected a few thousand in 2007 (mainly correcting wrong linkages between INGER and IRGC samples), but most of this task is pending.