ESTABLISHMENT OF AN EUROPEAN PLANT GENETIC RESOURCES INFORMATION INFRA-STRUCTURE

EURISCO Uploading Mechanism – Technical notes

Draft, July 4, 2002

This paper will discuss the EPGRIS uploading mechanism, in particular it will concentrate on the following topics:

- Import file data format
- Preliminary operations
- Checks

This paper reflects the EURISCO descriptors version July 3 2002.It is assumed that an accession can uniquely be identified by the three descriptors INSTCODE, ACCENUMB and GENUS.

Import file data format

It is expected that the import file be TAB-delimited text file. Fields are separated by a TAB character (ASCII 9) and records are terminated by either a DOS end of line (ASCII 13 and ASCII 10) or the UNIX end of line (ASCII 10).

The first line of the import file must contain the list of descriptors that are expected in the import file, this means that the file should be viewed as a table where the column label is the descriptor and the rows are the actual data. All lines should have the same number of columns as the first line. If a line has fewer columns, we assume missing elements to be NULL, this would generate a message status. If a line has more columns than provided descriptors in the first line, it is assumed an error. All fields with no data (this would be like <TAB><TAB>) are assumed NULL.

Preliminary operations

The NICODE descriptor is used to identify the national inventory: prior to importing data from an import file, all database records associated with the national code will be deleted. This means that each time a file is submitted it should represent the whole list.

This procedure is necessary to simplify deletions and modifications. This procedure is necessary to simplify deletions and modifications, and avoids any ambiguity.

Checks

There are two main types of checks:

1. **Preliminary checks.** These are performed at the beginning of the process and cover data elements common to the whole import file. Errors caught in this process will generally abort the whole import process.
2. **Line checks.** Each line is checked and there can be errors or warnings generated. A warning will not prevent the record from being added, an error will.

The check sequence is as follows:

1. **Data format.** The file is opened and the first line is read (the line containing the descriptors). The first thing that is checked is the line terminator. If it doesn't correspond to either the DOS format (ASCII 13 + ASCII 10) or the UNIX format (ASCII 10), the file will be discarded and the whole import process aborted.

2. **Descriptors**. Each descriptor is matched against known descriptors, if one is not recognised, it will be ignored for all import records; one warning will be issued to describe this problem. This allows using files containing more data, without having to strip out additional descriptors.
3. **Required descriptors.** It is checked that among the provided descriptors we find all those that are required. In particular: INSTCODE, ACCENUMB, GENUS and NICODE will have to be provided. If any of these descriptors are not provided an error is issued and the whole import process is aborted.
4. Among these required descriptors NICODE refers to the national inventory and must match an entry in the countries lookup table: since this code will be the same for all entries, if there is no match with the lookup table it can be assumed that all entries will have an error, so the whole import process is aborted.
5. If all these preliminary tests are passed (points 1 to 4) we go on to check individual fields.
    a. **Required descriptors.** Although this check is done at the beginning, it is repeated to ensure that the required descriptors are not NULL, in particular, INSTCODE, ACCENUMB, GENUS and NICODE need to have a value. If this check fails an error is raised and the record is rejected.
    b. **INSTCODE.** The institute code must match one of the entries of the FAO/VIEWS institutes lookup table, if it is not so an error is posted and the record is rejected. It is not possible to ignore this error, because the institute code is part of the unique key for the accession.
    c. **NICODE.** The national inventory code refers to the countries lookup table, if the provided code does not match any entry the record is rejected.
    d. **Taxonomy -** GENUS, SPECIES, etc. We check if there is an entry in the taxonomy table corresponding to the non NULL provided fields, if so, we link it to the entry, if not, we create a new entry in the taxonomy table.
    e. **ACQDATE, COLLDATE:** dates are expected in the YYYYMMDD format where missing elements, day or day and month, are marked as ‘—‘. So one can have 1989----, 198906—or 19890615. If the date is fully formed we check if it is valid. If one provides 6 characters we assume the date is YYYYMM, if 4 characters are provided we assume it is a year: in these cases the date is normalized (by adding the appropriate ‘—‘) and a message status is issued. If the date is not valid a warning is issued, the field will become NULL, but the record is still accepted.
    f. **LATITUDE, LONGITUDE.** Latitude is expected in the ‘DDMMSSX’ format where DD stands for degrees, MM stands for minutes, SS for seconds and X for hemisphere. Longitude has the ‘DDDMMSSX’ format. Here also, missing values are expected to be expressed with ‘—‘. We accept and normalize the following cases: for latitude ‘DDX’, DDMMX’, and for longitude ‘DDDX’, ‘DDDMMX’ and ‘DDDMMSSX’. Other cases will be treated as warnings. Note that providing coordinates in a ‘DDX’ format is rather useless when showing maps, because the precision might be so low that the point falls outside the actual country. If either the latitude or the longitude fail the check, both will become NULL and a warning will be issued, the record is still accepted.
    g. **SAMPSTAT, COLLSRC, STORAGE.** These are codes that refer to lookup tables. If the value is missing we assume the default ‘UNKNOWN’ entry, if the value is present it must match an entry, if it does not, a warning will be issued and the code will be changed to the default ‘UNKNOWN’ one, the record is still accepted.
    h. **BREDCODE, DONORCODE, COLLCODE and DUPLSITE.** These all refer to the FAO/VIEWS institutes lookup table. If no data is provided the default unknown ‘UNK000’ institute code will be used, if the data is present it must match an entry in the institutes table or a warning is issued. These fields are related to the respective BREDDESCR, DONORDESCR, COLLDESCR and DUPLDESCR: these fields should be used to provide institute information if an appropriate code cannot be found in the institutes lookup table, or if the information found in the lookup table is not up to date.

i. Finally, we insert the record and if there is a duplicate we issue an error. Accessions are uniquely identified by the combination of INSTCODE, ACCENUMB and GENUS.

From the above discussion it should be noted that there are three levels of errors that can be generated:

- **Fatal errors:** Such errors, as a bad file format or an incorrect national inventory code, will prevent any record from being inserted or even processed. Such errors are an indication that the import file cannot be processed and that action must be taken before resuming the import process.
- **Errors:** These are errors, such as an incorrect INSTCODE, that prevent a record from being added. These occur when there is not enough information to identify an accession, making it impossible to add it to the database.
- **Warnings:** These indicate that a particular field has incorrect data and that its contents cannot be accepted. Usually this would mean that the descriptor will take the database default value, but the record will still be accepted.
- **Message status:** These are issued when a piece of information does not have a correct format, but its contents can be implied in an obvious way. An example would be a date such as 198707: the value is missing the trailing '—', we fix the value and issue a status message.