ICTV load next MSL (#32/2017)

“EC 49, Budapest, Hungary, August 2017; Email ratification 2018”

# File locations

Scripts, code and original documents:

Genome-**ws-02**.ad.uab.edu: D:\Box Sync\ictvonline\ICTV\_update\2018\_updates\2018 MSL32.ICTV2017

Server-local staging directory:

Genome-**testdb**.ad.uab.edu D:\Data transfer\taxonomy [Production]

Genome-**bmidb**.ad.uab.edu D:\DB Backup [Dev]

# Protocol

## Setup/Backup: create ICTVonline32@genome-testdb

Production Server: genome-testdb.ad.uab.edu

Backup to D:\Microsoft SQL Server\MSSQL12.MSSQLSERVER\MSSQL\Backup\ICTVonline.20180204.1824.pre-msl32.bak

Restore that to ICTVonline32 for development.

Backup Server: genome-ictv.ad.uab.edu

Do nothing with that for now. Copy final db there.

## Data load & MSL Prep overview

***Return to older protocol!***

*Elliot started with a dump of previous year that I prepared in excel, and entered all the deltas.*

1. I have uploaded the file  ICTV2017\_MSL31\_to\_MSL32.v2.xlsx that contains all of the MSL updates to the ictvonline share in the 2018 MSL32.ICTV2017 folder.
2. The file ‘Ratification\_TP\_list\_AMQ\_v5.xlsx’ is in the \_ICTV ratification 2018 documents folder. This is the original list of taxonomic changes compiled by Andrew King, with updates to correct errors that I have found.
3. I added an ‘in’ change: ‘new\_type’ for new species that are type species.
4. I have highlighted the duplicated row for the ‘split’ genus Tectivirus.
5. We will no longer be entering any metadata (accession numbers, virus name, abbreviation, etc.) into the taxonomy tables and MSL. In fact, at some point we should think about removing this data/columns from the table since it is outdated. But for now, just leave them blank. We are now using the virus\_isolates table for this data.
6. The new order Ortervirales does not get a molecule type, since that needs to be assigned on the family level (molecule type differs by family).
7. Be sure to add the file extension to all existing filename records.
8. Be sure to keep the extension of the .zip file filenames when loaded into the table. I have already uploaded all of these to the production web server.
9. Don will be updating the web code to provide the filename including the extension when creating the url to download the proposal.

In previous year we used an exceptional protocol of loading the spread sheet and extracting the changes.

This year we return to our previous protocol of entering changes into ICTVXXXX\_MSLYY\_to\_MSLZZ.xlsx and loading that:

* exported a spreadsheet from current MSL
* Annotated & added taxa changes
* Used excel to assign new taxon\_ids
* Re-imported final sheet into dbo.load\_next\_msl
* Ran scripts to create the new MSL in taxonomy\_node from the load\_next\_msl table.
* Renamed load\_next\_msl to load\_next\_msl\_## to keep for posterity

## Implementation – create load\_next\_msl

### SCRIPT: 1.0 create load\_next\_msl table.sql

Edit script to rename current table(s).

exec sp\_rename 'load\_next\_msl', 'load\_next\_msl\_31'

exec sp\_rename 'load\_next\_msl\_taxonomy', 'load\_next\_msl\_taxonomy\_31' -– only used for 31

Edit create table to remove ID from dest\_taxnode\_id, which was added last year for the other protocol.

Run script to re-create.

### SCRIPT: 1.0 drop obsolete tables.sql

Empty script this year

### SCRIPT: 1.01 load status query

Update tree\_ids

declare @new\_tree int; set @new\_tree = 20170000

### SCRIPT: 1.1 load\_next\_msl.sql

Excel: save sheet as **Unicode** TXT

SQL: import via GUI

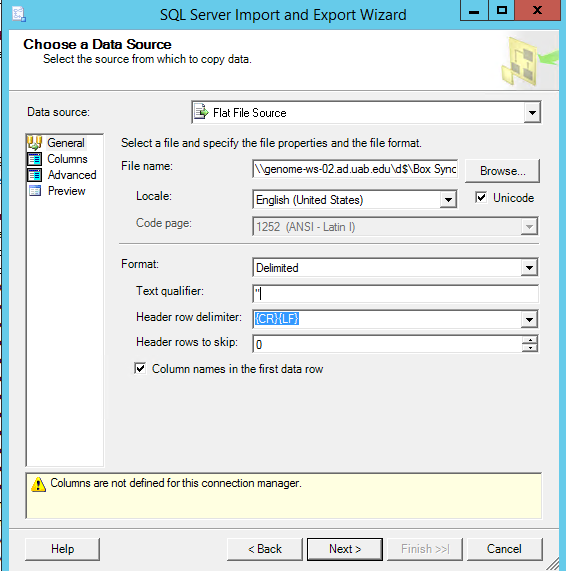
Drop the [load\_next\_msl\_unicode] first.

Database > tasks > import

Source: Flat file

Text qualifier: “

Advanced > change all non-int col widths to 500



Dest: SQL Server Native Client 11.0

Dest mapping: [load\_next\_msl\_unicode]

Find this code in “1.1…sql”

TRUNCATE TABLE [load\_next\_msl]

INSERT INTO [load\_next\_msl] SELECT \* FROM [load\_next\_msl\_unicode]

DROP TABLE [load\_next\_msl\_unicode]

(6013 rows transferred)

*Data checking/cleaning:*

### SCRIPT: 1.4 qc next\_msl.sql

Test 1.4.1 - some splits, new don't have dest\_level - is that set automatically later based on # of ; in dest\_target?

Test 1.4.6 - "WARNING: rename destination is a lineage - perhaps this should be move\_rename?" - do we need to fix that? Don't think so… next update auto-fixes…

Test 1.4.21 - missing internal nodes - will likely have to come back and fix these 33. Mostly Unassigned sub-families or genera.

* need to add subfamiliy Caudovirales;Siphoviridae;Nymbaxtervirinae that elliot had missed. Have fixed v3 excel, and in db (1.4 qc)
* rename family: Unassigned;Ophioviridae;Unassigned;Ophiovirus >> Unassigned;Aspiviridae;Unassigned;Ophiovirus
  + change the Unassigned subfamily to a move?
  + Have not fixed
  + Need to fix
* --------------------HERE----------------------QQQQ

## TODO

dump “differences” spreadsheet.

* 1.1 load\_next\_msl.sql
  + BULK LOAD into ‘load\_next\_msl’
  + check row count vs excel file
  + double check accents, smart quotes, etc
  + QC: check for non-existant implicit parents
  + run MSSQL profiler - QC query took > 3 hours. Profiler recommended additional indices on load\_next\_msl & taxonomy\_node
    - Addded load\_next\_msl indices to "1.0 create load\_next\_msl table.sql"
    - go back and re-load and re-run checks.
    - the QC check that finds non-existant move-to targets still gives false positives….
  + stats

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| dest\_in\_change | src\_out\_change | taxa\_ct | Done | in step |
| new | NULL | 692 | 692 | 4 |
| % | NULL | 692 | NA | NA |
| NULL | type | 1 | 1 | 2 |
| NULL | rename | 822 | 822 | 2 |
| NULL | promote | 1 | 1 | 5 |
| NULL | move\_type | 6 | 6 | 5 |
| NULL | move\_rename | 40 | 40 | 5 |
| NULL | move | 53 | 53 | 5 |
| NULL | metadata | 17 | 17 | 2 |
| NULL | merge | 2 | 2 | 6 |
| NULL | abolish | 44 | 44 | 3 |
| NULL | % | 986 | NA | NA |
| NULL | NULL | 2971 | 2971 | 2 |

* + scan for (and fix) src\_out\_change verb: rename -> rename\_move
* 2. xfer unchanged metadata and renamed.sql
  + update to support out\_change ="metadata"
  + transfering 824 src\_out\_change=rename records
  + transfering 2980 src\_out\_change=unchanged records
  + transfering 10 src\_out\_change=metadata records
* 3. xfer abolished.sql
  + transfering 45 abolished records
* 4. xfer new taxa and splits.sql
  + Error: Subquery returned more than 1 value. This is not permitted when the subquery follows =, !=, <, <= , >, >= or when the subquery is used as an expression.
    - fixed species being created in a renamed genus, refered to genus' old name
    - re-creation of a subfamily that already existed.
  + transfering 691 dest\_in\_change=new
* 5. xfer moves.sql
  + update to support move\_type: change to src\_out\_change like 'move%'
  + transfering 52 src\_out\_change=move
  + transfering 39 src\_out\_change=move\_rename
  + transfering 6 src\_out\_change=move\_type
* 6a xfer merges
* 6b. create new molecule types (move earlier? )
* 6c. xfer metadata (isolates, abbrevs, accession, molecule\_type)
* 7. fixups.sql
  + set MSL notes - need to get text from Elliot- moved that script 2, and defined a place in the excel file.
* 8a. check for duplicate taxa names
* 8ba. build delta nodes.sql
  + CONSIDER IF WE SHOULD USE SP [rebuild\_delta\_nodes]

|  |  |  |
| --- | --- | --- |
| MSL | change\_type | taxa count |
| 30 | UNCHANGED | 2782 |
| 30 | Abolished, | 44 |
| 30 | Merged,Moved, | 2 |
| 30 | Moved, | 164 |
| 30 | New, | 692 |
| 30 | Renamed, | 417 |
| 30 | Renamed,Moved, | 446 |
| 30 | Renamed,Moved,Promoted, | 1 |

* 8bb. rebuild merge-split table.sql
  + CONSIDER IF WE SHOULD USE SP [rebuild\_node\_merge\_split]
* 9x. QC queries.sql

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| prev\_msl | prev\_tags | msl | next\_tags | next\_msl | viz\_count | hidden\_count | total\_count | notes and errors |
| NULL | NULL | 30 | NULL | NULL | 0 | 101 | 101 |  |
| NULL | New, | 30 | NULL | NULL | 686 | 6 | 692 |  |
| 29 |  | 30 | NULL | NULL | 2782 | 0 | 2782 |  |
| 29 | Merged,Moved, | 30 | NULL | NULL | 2 | 0 | 2 |  |
| 29 | Moved, | 30 | NULL | NULL | 164 | 0 | 164 |  |
| 29 | Renamed, | 30 | NULL | NULL | 416 | 1 | 417 | OK: Tree root node is hidden and renamed |
| 29 | Renamed,Moved, | 30 | NULL | NULL | 446 | 0 | 446 |  |
| 29 | Renamed,Moved,Promoted, | 30 | NULL | NULL | 1 | 0 | 1 |  |

*these are correct in taxonomy\_node\_n (nvarchar/ntexT)*

|  |  |  |  |
| --- | --- | --- | --- |
| title | msl\_release\_num | status\_msg | count |
| accent\_check: | 30 | ERROR: ǎ missing | 1 |
| accent\_check: | 30 | ERROR: ā missing | 1 |
| accent\_check: | 30 | ERROR: ē and/or ō missing | 2 |
| accent\_check: | 30 | ERROR: ĭ and/or ì missing | 1 |
| accent\_check: | 30 | ERROR: ī missing | 1 |
| accent\_check: | 30 | OK: á | 2 |
| accent\_check: | 30 | OK: ë | 3 |
| accent\_check: | 30 | OK: í | 2 |

* push ICTVonlineDEV -> ICTVonline
  + MSSQL: ICTVonlineDEV > (right-click) > Tasks > Backup
  + Start > Administrative Tools > IIS > Application Pools > ICTVonline > STOP
  + MSSQL: ICTVonline> (right-click) > Tasks > Restore (dump of ICTVonlineDEV)
  + Start > Administrative Tools > IIS > Application Pools > ICTVonline > START
  + test <http://ictvonline.org/virusTaxonomy.asp>
* [20180305] 9za.export\_MSL\_extended\_from\_taxonomy\_node.sql
  + open “ICTV Master Species List 2016 v5.xlsx” in excel
  + Version tab - paste in results from first query (check dates!)
  + taxonomy tab - rename
  + taxonomy tab - Paste result from query
  + re-select the version tab
  + save as “ICTV Master Species List 2017 v1.xlsx”
  + email to elliotl@uab.edu
  + delete old version from the folder
* [20180306] 9zc. export terse diff list.sql
  + open “DeltaReport\_MSL30.xlsx” in excel
  + run query
  + copy/paste data w/o headers
  + save as “DeltaReport\_MSL32.xlsx”
  + delete old version from the folder
  + email to Elliotl@uab.edu

# Post Production fixes