ICTV load next MSL (#35/2019)

“EC 51, Berlin, Germany, July 2019; Email ratification March 2020 (MSL #35)”

# File locations

Using Genome-ws-02 as ETL DB server

ISO image from Elliot in Box.com: [SW\_DVD9\_NTRL\_SQL\_Svr\_Standard\_Edtn\_2017\_64Bit\_English\_OEM\_VL\_X21-56945.ISO](https://uab.box.com/s/hd790t74xcj157jmrjqvoc6tmq06axak)

ICTVonline34prod backup: box:/ICTVonline/taxonomy/ICTV\_update/2019\_updates/ICTVonline34\_033120.bak.zip

Genome-**ws-02**.ad.uab.edu: D:\client\github\ICTVonlineDbLoad

<https://github.com/rusalkaguy/ICTVonlineDbLoad>

Final copy of scripts will also be copied to box :

C:\Users\curtish\Box\ictvonline\taxonomy\ICTV\_update\2019\_updates\[20200408\_MSL35](https://uab.box.com/s/8feqpnqq0ivr6bbvo7mvwl6hwjy4qhc2)

Server-local staging directory:

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

dev.ictvonline.org D:\DB Backup [Dev]

# Protocol

## Setup/Backup: create ICTVonline35@genome-ws-02

Production Server:

box:/ICTVonline/taxonomy/ICTV\_update/2019\_updates/ICTVonline34\_033120.bak.zip

Restore that to ICTVonline35 for development.

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

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

## Data load & MSL Prep overview

On Mar 26, 2020, at 5:25 PM, Lefkowitz, Elliot J <ElliotL@uab.edu> wrote:

Hi Curtis,

The next ICTV release is now in your hands. The necessary files are in Box:ictvonline:Taxonomy:ICTV\_update:2019\_updates.

The spreadsheet ‘ICTV MSL Release 35 2019 Changes.2.xlsx’ is the starting point and contains a row for each change from each one of the proposal spreadsheets in the folder ‘Proposals’.

I created a copy of the current production database in dev.ictvonline.org. It is call ICTVonline34prod. This can be the starting point for ICTVonline35. (Note that this dev server is pretty slow. If you wanted to develop this on your local PC, that might work better.)

The description of this 2019 release is:

“EC 51, Berlin, Germany, July 2019; Email ratification March 2020 (MSL #35)”

Thanks!

Elliot

New protocol.

* In past years (MSL32) we started with a dump of the previous year I would export, into which Elliot would enter all the deltas.
* This year Elliot created a deltas-only spreadsheet from whole cloth, so all the column names will be different.

The MSL32 protocol of entering changes into ICTVXXXX\_MSLYY\_to\_MSLZZ.xlsx and loading was:

* 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

However, MSL33, MSL34a and MSL34b were all delta-only loads, for which I don’t have notebook/protocols.

load\_next\_msl\_32 6014

load\_next\_msl\_33 390

load\_next\_msl\_34a 1079

load\_next\_msl\_34b 53

Ad hoc MSL35 protocol (reconstructed from MSL33-34):

* Elliot Built spreadsheet “ICTV MSL Release 35 2019 Changes.2.xlsx” of just changed taxa
* Curtis renames columns by pasting in header row from “col\_mapped.xls”

## Implementation – create load\_next\_msl

### SCRIPT: 0.b1. rename old load\_next\_msl.sql

Edit script to rename current table(s) and their default constraints

exec sp\_rename 'load\_next\_msl', 'load\_next\_msl\_34b'

### SCRIPT: 0.b2.create\_table-load\_next\_msl-delta.sql

Edit MSL # in Default constraint at bottom.

Run script to create this years load table.

### XLS: map column names

Load & combine :

./excel\_files/

load\_next\_msl.col\_map.xls

* + - Copy first row

ICTV MSL Release 35 2019 Changes.2.xlsx

* + - Copy 1st row, insert as 2nd
    - Paste 1st row of load\_next\_msl.col\_map.xls over 1st row.
    - Add “filename” column in A:

=MID(CELL("filename",$A$1),FIND("[",CELL("filename",$A$1))+1,FIND("]", CELL("filename",$A$1))-FIND("[",CELL("filename",$A$1))-1)

* + - Add column C: “IsWrong”
    - Elliot added column F “Spreadsheet”
    - Elliot added columns Realm, subRealm, Kingdom, Subkingdom,Phylum,Subphylum, etc – which map together to [srcHigherTaxon]. Add mapping names (srcRealm, etc), but do not add those to the table.
    - Add column G “srcHigherTaxon”
      * Define as right most “higher taxa” [Realm…SubClass]
      * =IF(O3<>"",O3,IF(N3<>"",N3,IF(M3<>"",M3,IF(L3<>"",L3,IF(K3<>"",K3,IF(J3<>"",J3,IF(I3<>"",I3,IF(H3<>"",H3,""))))))))
    - Add column W “srcIsType”
    - Don’t map “srcAccessions”
    - Add mappings for realm, subrealm, kingdom, subkingdom, phylum, subphylum, class, subclass, suborder, subgenus
    - Don’t map “exemplarRefSeq”
    - Copy rows 1:2 “database columns” and “original headers” back to load\_next\_msl.col\_map.xls
    - Delete 2nd row (original headings)
    - Save to .xls

load\_next\_msl.col\_map.xls

Modify table schema:

### [spreadsheet] [varchar](100) NOT NULL, -- original (unzipped) spreadsheet filename

[srcRealm]...[srcSubClass] & [srcSubOrder] & [srcSubGenus] [varchar](100) NULL, -- MSL35: new

-- Update computed columns for rank, lineage, name (both src and dest to add new levels)

### MSSQL: Tasks>import Data… (failed; skip to SQL Insert)

Data Source:

Type: XLS

Filename: D:\client\github\ICTVonlineDbLoad\excel\_files\ICTV MSL Release 35 2019 Changes.2.col\_mapped.xls

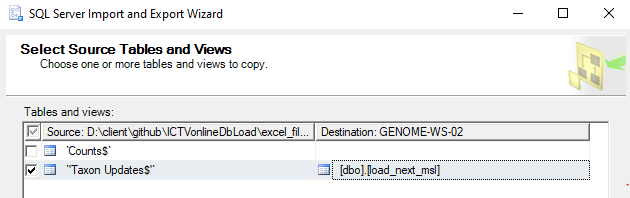
Destination:

Driver: SQL Server Native client 11

Genome-ws-02

ICTVonline35

Map to table



“Edit mappings”

### XSL->SCRIPT: 0.c.insert\_into-load\_next\_msl-delta.sql

Add worksheet “generateSQL” and add columns AW:BB to generate SQL inserts for each row in column AW

Paste all of column AW into “0.c.insert\_into-load\_next\_msl-delta.sql”, save and run.

# XXXXXXXX

### 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

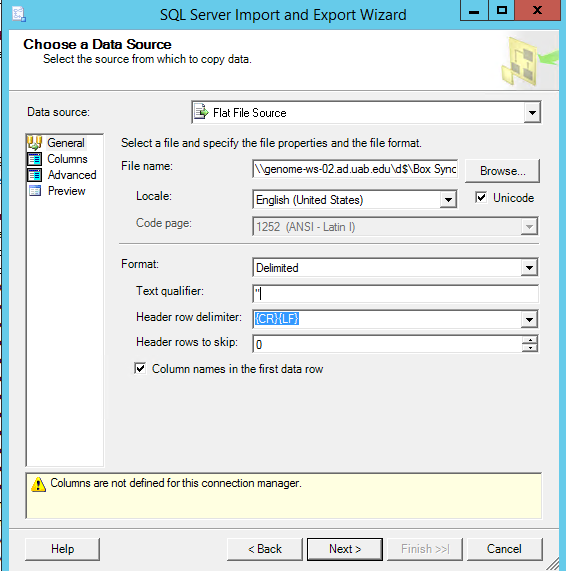
Drop the [load\_next\_msl\_unicode] first.

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