ICTV load next MSL (#37/2021)

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

# Status/Next

Total MSL37 taxa 10,.431 (official MSL export)

# File locations

Proposal Merge <https://github.com/ICTV-Virus-Knowledgebase/MSL_merge>

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

Macbook:/Users/curtish/Documents/ICTV/ICTVonlineDbLoad

Genome-**ws-02**.ad.uab.edu: E:\ICTV\ICTVonlineDbLoad

Box :\ictvonline\taxonomy\ICTV\_update\2022\_updates\ ????

Server-local staging directory:

??? D:\Data transfer\taxonomy [Production]

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

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)

ICTVonline36prod backup: ec2-54-89-205-80.compute-1.amazonaws.com: D:\Backup\ICTVonline36\_022522.zip

# Protocol

## Setup/Backup: create ICTVonline37@dev.ictvonline.org

Get this year’s IP

<https://whatismyip.host/>

home (ATT-Fiber)

laptop @ UAB**: 138.26.210.250**

Curtis created ICTVonline36

* Elliot copied .bak file from production onto DEV server
* Curtis restored that backup into ICTVonline36
* Curtis created ICTVonline37
* Curtis restored the backup into ICTVonline37

Production Server:

## Data load & MSL Prep overview

N=7 master ZIP files for ratified proposals were downloaded from

<https://talk.ictvonline.org/files/ratification/>

R code

[merge\_proposal\_zips.Rmd](https://github.com/ICTV-Virus-Knowledgebase/MSL_merge/blob/main/merge_proposal_zips.Rmd)

Unzips, extracts those zips, and merges xlsx files to produce

[merged\_proposals.tsv](https://github.com/ICTV-Virus-Knowledgebase/MSL_merge/blob/main/merged_proposals.tsv)

was run in RStudio on Curtis’ MacBook.

When re-building zip files with changed .xlsx files – need to avoid internal path names

zip ../$(basename $PWD).zip \*

## Implementation – create load\_next\_msl

**Now** do a Replace of [X] => [] to clear the “done” checkboxes.

### [X] 0.b1. rename old load\_next\_msl.sql

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

DECLARE @OLD\_MSL varchar(32); SET @OLD\_MSL='36'

### [X] 0.b2.create\_table-load\_next\_msl-delta.sql

Edit MSL # in Default constraint at the bottom.

ALTER TABLE [dbo].[load\_next\_msl] ADD CONSTRAINT [DF\_load\_next\_msl\_\_msl\_release\_num] DEFAULT ((37)) FOR [dest\_msl\_release\_num]

Run script to create this year’s load table.

### [X] 0.b3.alter\_schema.sql

Add computed column, eg:

alter table taxonomy\_node add [\_numKids] AS ([right\_idx]-[left\_idx]-1)/2 PERSISTED

### [X] MSSQL: Tasks>import Data… (NOT import flat file)

Data Source:

General>Filename: c:\Users\curtish\Downloads\load\_next\_msl.37v1.txt.utf-16le.txt

General> check Unicode

Advanced>Suggest Types…

Number of Rows: 20000

Un-check Booleans

OK

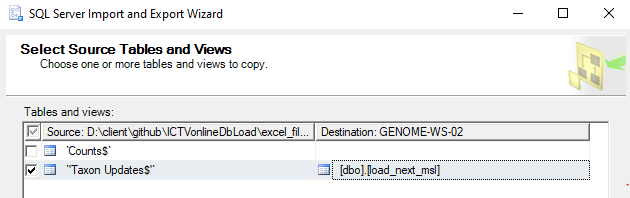
WARNING: if you forget to Suggest Types, columns default to STR(50) and [proposal] and other things get truncated.

Destination:

Driver: SQL Server Native client 11

ICTVonline37

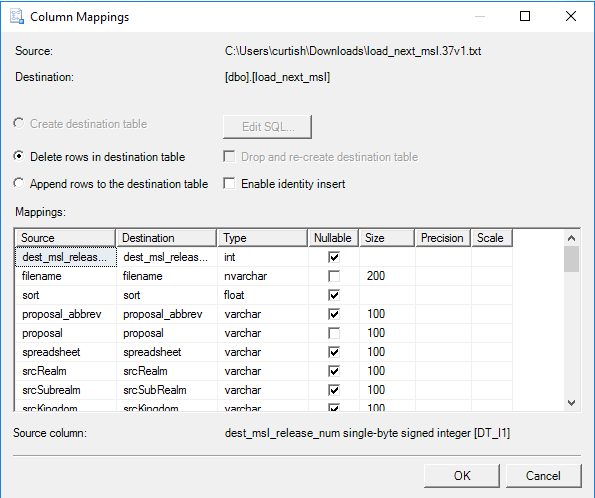
Map to table



> Edit mappings

\* check nothing unmapped

\* SELECT “Delete Rows in destination table”



### [X] 0.c.insert\_msl\_into\_taxonomy\_toc.sql

Create a row for new MS

Skip 50,000, rather than 100k, if 2nd MSL this year.

### [X] 0.d.drop\_obsolete\_tables.sql

Empty script this year

### [X] 1.0 load progress query.sql

Reports nothing yet; will be helpful as we add things to taxonomy\_node.

* + stats

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| dest\_in\_change | src\_out\_change | taxa\_ct | Done | in step |
| new | NULL |  |  | 4 |
| % | NULL |  |  | NA |
| NULL | type |  |  | 2 |
| NULL | rename |  |  | 2 |
| NULL | promote |  |  | 5 |
| NULL | move\_type |  |  | 5 |
| NULL | move\_rename |  |  | 5 |
| NULL | move |  |  | 5 |
| NULL | metadata |  |  | 2 |
| NULL | merge |  |  | 6 |
| NULL | abolish |  |  | 3 |
| NULL | % |  |  | NA |
| NULL | NULL |  |  | 2 |

* + scan for (and fix) src\_out\_change verb: rename -> rename\_move

### [X] 1.b.load\_next\_msl-convert\_blank\_to\_null.sql

Many found and fixed.

Must add any new columns…

### [] 1.X delete\_an\_MSL\_tree.sql

Need this only if we have to back out our inserts.

Instead, please make backups often!

### [X] 2.a. load\_next\_msl - data cleans.sql

-- remove leading and trailing spaces (update create in excel)

(144 rows affected)

COMMIT TRANSACTION

### [X] 2.a2.load\_next\_msl - data cleaning CSVs.sql

There are some a lot of “;”’s, but there are already a lot that have made it into taxonomy\_node.genbank\_accession\_csv.

Should we really fix these now? [nope]

### [X] 2.a3.QC\_check\_for\_illegal\_name\_suffixes

CODE NEEDS UPDATING, see: <https://ICTV.global/code>

These seem like typos for -virus, and Ops –these have been around since MSL 34/35.

2021.001B.A.v1.abolish\_Caudovirales.zip genus:Incheonvrus;

2021.010B.A.v1.Binomial\_names.zip genus:Tunggulviirus; genus:Incheonvrus;

Seem to violate rules for suffixes for ranks. Are these legal?

YES: new naming last year: <https://ICTV.global/code>

2021.006D.A.v1.Polydnaviriformidae\_1renfam\_3rensp.zip

Family Polydnaviridae => Polydnaviriformidae

Genus: Bracovirus => Bracoviriform

### [X] 2.b map actions.sql

Maps entry in the “change” column over to official verbs in the “\_action” column, with some QC

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **report** | **dest\_msl\_release\_num** | **\_action** | **row\_ct** | **max\_taxid** |
| action summary | load\_next\_msl.37v1.txt | 37 | abolish |  |
| action summary | load\_next\_msl.37v1.txt | 37 | move |  |
| action summary | load\_next\_msl.37v1.txt | 37 | new |  |
| action summary | load\_next\_msl.37v1.txt | 37 | promote |  |
| action summary | load\_next\_msl.37v1.txt | 37 | rename |  |

### [X] 2.c.load\_next\_msl-set-prev\_tax\_id.sql

Match up all to-be-changed taxa with their taxnode\_id’s from the previous MSL in taxonomy\_node

Correct by updating DB, and fixing .xlsx in parallel.

Faster than going back to 0 and re-loading.

### [X] 2.d.QC.check\_molecule\_types

The ones in the spreadsheet seem to have spaces, the ones in the database, don’t.

The script auto-corrects that problem.

### [X] 2.h.QC\_check\_for\_re-creates.sql

No problems

[X] **2.Z.insert\_root\_in\_taxonomy\_node.sql**

Set release name

MSL=37

NAME=2021

NOTES=EC 53, Online, July 2021; Email ratification March 2022 (MSL #37)

TREE\_ID=202100000

TR\_taxonomy\_node\_update\_indexes: tree[202100000] re-indexed (1 row affected)

### [X] 3.a.create\_next\_MSL\_by\_copying\_previous\_MSL.sql

Create new MSL by copying all nodes from prev MSL

SET @MSL=37

SET @TREE\_ID=202100000

TR\_taxonomy\_node\_update\_indexes: ROWS: DEL=0, INS=11870

(11870 rows affected)

### [X] 3.c.load\_next\_msl-set-dest\_taxnode\_id-dest\_parent\_id.sql

Errors:

|  |  |
| --- | --- |
| **ERROR\_report** | actions that should have dest\_parent\_id set, but do NOT |
| **sort** | 5658 |
| **\_action** | new |
| **\_dest\_taxon\_rank** | species |
| **rank** | species |
| **\_dest\_lineage** | Riboviria;Orthornavirae;Negarnaviricota;Haploviricotina;Monjiviricetes;Mononegavirales;Xinmoviridae;Doupovirus;Doupovirus australiaense |
| **\_dest\_parent\_name** | Doupovirus |
| **\_dest\_taxon\_name** | Doupovirus australiaense |
| **dest\_taxnode\_id** | 202112935 |
| **dest\_parent\_id** | NULL |
| **dest\_ictv\_id** | 202112935 |
| **FIX:** | Added “new genus” row to load\_next\_msl and .xlsx |

Fixed in script with an insert.

### [X] 4.a.apply\_create\_actions\_RANK\_high\_to\_low.sql

FIXES MSL37

merged N=62 “chained” changes – where a taxon is changed twice.

Hand fixes to all two-actions-on-same-dest\_taxon

Go back and re-run “2.b map actions.sql” for new lines created by this

BACKUP: ICTVonline37.20220308.3.c.done.bak (includes above fixes)

Edits current MSL in taxonomy\_node based on load\_next\_msl for (mostly) create actions

Merge (including destruction of merged nodes from current MSL)

Move (+rename +istype)

New/Split

Rename

Promote

Expected

|  |  |
| --- | --- |
| **ACTION** | **COUNT** |
| new | 1847 |
| move | 785 |
| abolish | 32 |
| rename | 3291 |
| promote | 3 |

Results

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| report | action | prevMSL | load\_next\_msl | nextMSL | (No column name) |
| counts by [ACTION] | abolish | NULL | 32 | NULL | ERROR (step 4.b) |
| counts by [ACTION] | demote | NULL | NULL | NULL | OK |
| counts by [ACTION] | merge | NULL | NULL | NULL | OK |
| counts by [ACTION] | move | 714 | 714 | NULL | OK |
| counts by [ACTION] | new | NULL | 1830 | 1830 | OK |
| counts by [ACTION] | promote | 3 | 3 | NULL | OK |
| counts by [ACTION] | rename | 3263 | 3263 | NULL | OK |
| counts by [ACTION] | split | NULL | NULL | NULL | OK |
| counts by [ACTION] | type | NULL | NULL | NULL | OK |

#### Investigating promotes 2<>3

Sort=3880 promote Peduovirinae(subfamily) to Peduoviridae(family), looks fine

Ok, in the xlsx, they set rank to “family” instead of “subfamily”. Need to fix code in step 4.a. to use \_dest\_taxon\_rank, NOT [rank].

Re-worked step 4.a to

* add a view “load\_next\_msl\_isOk” to bake in the “where isWrong is null” constraint.

Conclusion: data problem, not a code problem!

* Peduovirinae(2021.001B.A.v1.abolish\_Caudovirales: move) vs Peuoviridae (2021.063B.A.v1.Peduoviridae.zip: promote): I’ll need Elliot’s blessing, but I’m betting the promote is correct, as the abolish\_Caudovirales proposal has been rather high on the error count.

Also found 26 other cases where a \_src\_taxon\_name was modified by multiple proposals/actions.

* Add columns [merged\_left\_sort, merged\_right\_sort, sort\_str]
* Resolved 2 with merge chain/dup query
* Resolved the rest with a query with a WHERE \_dest\_lineage in ( /\*review results from xlsx from EJL\*/)

Current status:

Restore backup “after 3.c”

Apply 4.a schema patches

START TRANSACTION

Run 4.a fixes

Run 4.a WORK

Run 4.a QC after work.

Stats now all happy!

Created backup: ICTVonline.20220311.4.a.done.bak

Proceed to 4.b

### [X] 4.b.apply\_delete\_actions\_RANK\_low\_to\_high.sql

Edits current MSL in taxonomy\_node based on load\_next\_msl fro destructive actions

Abolish

#### Problems

Several genera can’t be abolished, because they still contain species.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **taxnode\_id** | **Rank** | **lineage** | **\_numKids** |  |
| 202100190 | Order | Duplodnaviria;Heunggongvirae;Uroviricota;Caudoviricetes;Caudovirales | 16 | Abolish order |
| 202100540 | Family | Duplodnaviria;Heunggongvirae;Uroviricota;Caudoviricetes;Caudovirales;Podoviridae   * Genus Giessenvirus; renamed to Hungoriovirus in Caudovricetes in 2021.069B   + Species Hungariovirus C1302; moved-renamed into Hungoriovirus | 2 | Abolish family, move-rename genus |
| 202100713 | family | Duplodnaviria;Heunggongvirae;Uroviricota;Caudoviricetes;Caudovirales;Siphoviridae   * Nevevirus: move to Caudoviricetes, species already renamed-moved [4a;2c;3c] * Pharaohvirus: new in Caudoviricetes 2021.064B instead of here [4a2c;] * Refugevirus: new in Caudoviricetes 2021.064B instead of here [4a;2c;]   Fixed in 4.a; several times.  Issues were fixing dest\_parent\_id – not updated automagically when changed dest lineage!  Fixes in: 2021.097B.R.error\_correction\_Caudoviricetes.xlsx | 11 | Abolish family, move genera  Edited 2021.064B to create Pharaoh and Refuge in Caudoviricetes |
| 202108168 | Genus | Duplodnaviria;Heunggongvirae;Uroviricota;Caudoviricetes;Drexlerviridae;Gyeonggidovirus   * Species: Cronobacter virus PhiCS01; abolish along with genus [4a;2c] | 1 | Abolish genus AND species  Note:  Duplication of exemplar Cronobacter phage CS01 in genus Kyungwonvirus. |
| 202103641 | genus | Guttaviridae;Alphaguttavirus   * Species Sulfolobus newzealandicus droplet-shaped virus; abolish | 1 | Already abolished in 2021.004A; bug in code? |
|  |  |  |  |  |

#### Results

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| report | action | prevMSL | load\_next\_msl | nextMSL | STATUS |
| counts by [ACTION] | abolish | 33 | 33 | 33 | OK |
| counts by [ACTION] | merge | 1 | 1 |  | OK |
| counts by [ACTION] | move | 224 | 224 |  | OK |
| counts by [ACTION] | new |  | 1610 | 1610 | OK |
| counts by [ACTION] | promote |  |  |  | OK |
| counts by [ACTION] | rename | 51 | 51 |  | OK |
| counts by [ACTION] | split |  | 11 | 11 | OK |

MSL37: Corrections all done in 4.a

MSL37: After successful abolish (N=33), create D:\MSSQL\Backup\ICTVonline37.20220317.4.b.done.bak

MSL37: After adding 2021.097B.R.error\_correction\_Caudoviricetes.zip proposal; create ICTVonline37.20220318.4.b.done.bak

### 4.c.qc\_actions

Doublecheck the above.

MSL37: pass

### 7a.qc\_empty\_taxa.sql

* MSL37: Duplodnaviria;Heunggongvirae;Uroviricota;Caudoviricetes;Mesyanzhinovviridae;Keylargovirus empty
  + Should have species [Keylargovirus JL001], 2021.051B.A.v1.Mesyanzhinovviridae.zip
  + Instead it was renamed and not moved:
    - Duplodnaviria;Heunggongvirae;Uroviricota;Caudoviricetes;Mesyanzhinovviridae;Rabinowitzvirinae;Yuavirus;Keylargovirus JL001
* Moved by hand in 7.a….sql

#### **BACKUP**: ICTVonline37.20220318.7.a.done.bak

### 7b. QC name suffixes

MSL37: added columns and checks for -viriform\*

#### **BACKUP**: ICTVonline37.20220318.7.b.done.bak

Pre-existing problems I’m not fixing:

|  |  |  |  |
| --- | --- | --- | --- |
| msl\_release\_num | name | out\_change | out\_target |
| 34 | Incheonvrus |  |  |
| 35 | Incheonvrus |  |  |
| 36 | Incheonvrus | move | Duplodnaviria;Heunggongvirae;Uroviricota;Caudoviricetes;Incheonvrus |
| 37 | Incheonvrus |  |  |
| 35 | Tunggulviirus |  |  |
| 36 | Tunggulviirus |  |  |
| 37 | Tunggulviirus |  |  |

### 7d. QC check for duplicate taxon names.sql

MSL37: no problems!

#### **~~BACKUP~~**~~: ICTVonline35.20200415.2150.step\_7d\_done.bak~~

### 7e.QC\_check\_for\_genera\_with\_type\_species\_issues.sql

N/A: Type species were abolished in MSL36.

### 7f.QC check for ictv\_id wo a new change-proposal.sql

MSL37: No new problems

### 7g.QC\_molecule\_id.sql

Trickle-up code is at

-- Runs for 8 minutes!

exec sp\_simplify\_molecule\_id\_settings

\* Start by opening XLSX report file

\* we don't have a way of flagging things we've reviewed, so many "problems" get re-reported every year.

**8a.rebuild\_delta\_nodes\_and\_merge-split\_table**

MSL37: worked

#### **BACKUP**: ICTVonline36.20220318.8a.done.bak

### 8b.QC\_merge-split\_table

MSL37: no issues

### 8c. pull isolates-accession-molecule

Accession number maintenance is sketchy

First: pull values from previous MSL in taxonomy\_node

Second: pull values from load\_next\_msl\_isOK view (this year’s load)

### MSL37: error species='Mykissvirus tructae', accession=”KX882061

select exemplarAccessions,\* from load\_next\_msl where exemplarAccessions like '% to %' or exemplarAccessions like '%-%'

update load\_next\_msl\_isok set

--select \*,

exemplarAccessions = '1:KX882061; 2:KX882062; 3:KX882063; 4:KX882064; 5:KX882065; 6:KX882066; 7:KX882067; 8:KX882068'

from load\_next\_msl\_isok where species='Mykissvirus tructae' and exemplarAccessions like 'KX882061%' --or genus='Mykissvirus'

update taxonomy\_node set

-- select \*,

genbank\_accession\_csv = '1:KX882061; 2:KX882062; 3:KX882063; 4:KX882064; 5:KX882065; 6:KX882066; 7:KX882067; 8:KX882068'

from taxonomy\_node

where msl\_release\_num=37 and name='Mykissvirus tructae' and genbank\_accession\_csv like 'KX882061%'

Added new file ‘2.a4.QC\_check\_for\_illegal\_accessions.sql’ to detect this earlier, next time

#### BACKUP: ICTVonline37.20220318.8.c.done.bak

### 9a.recompile views.sql

MSL37: ok

### 9x.QC\_deltas.sql

MSL37: Delta QC all clean!

#### BACKUP: ICTVonline37.20220318.9.x.done.bak

### 9y: copy proposal zips

Destination directory: D:\Web\CSProd\ICTV\proposals

Create zip with MSL\_merge/pro\_zips/zip\_me.sh

MSL37: copies up MSL\_merge/pro\_zips.20220318.zip

### 9z0.EXPORT\_MSL\_cheap\_and\_dirty.sql

MSL37: 10,433 rows, run time 02sec

### 9za.EXPORT\_MSL\_extended\_from\_taxonomy\_node.sql

MSL37:

### 9zc.export terse diff list.sql

Todo?\: script before and after whole-db-schema into git

* 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>

# Post Production fixes

## BACKUP: ICTVonline35.20200416.1731.step\_9x\_done.bak

### 7e.QC\_check\_for\_genera\_with\_type\_species\_issues.xFix\_only\_child\_species.sql

Set all only-child species to be type species.

Send new report to Elliot.

20200417.1700 Ran this on dev.ictvonline.org

### 7e.QC\_check\_for\_genera\_with\_type\_species\_issues.xFix\_designateTypeSpecies.sql

For the 3 remaining MSL35 genera with multiple type species, Elliot specified the correct one.

This script makes it so.

20200417.2050 Ran this on dev.ictvonline.org

### Improve speed of initial taxonomy (pre-expand to family) query.

Instead of computing child and descendant rank counts at query time, add that work to the taxonomy\_node\_compute\_indices SP.

Columns added to taxonomy\_node and computed by trigger/SP/UDF:

[rank]\_kid\_ct -- number of direct children of that rank

[rank]\_desc\_ct -- number of descendants (including direct children) of that rank

taxa\_kid\_cts -- English string listing non-zero direct, visible sub-taxa, aka udf\_getImmediateChildTaxaCounts()

taxa\_desc\_cts -- English string listing non-zero visible sub-taxa, aka udf\_getChildTaxaCounts()

\_numKids -- total number of children of any type

To implement the counts-to-English conversion, I created a new UDF (and deleted the two above): udf\_rankCountsToStringWithPurals()

Finally, if we ever need to update that function, or the indexer, there’s a convenience SP to re-index all MSLs: EXEC taxonomy\_node\_compute\_indexes\_ALL\_MSL

This means queries a MUCH faster and simpler, and this did not slow down insertion/updates noticeably (though the indexer code got a lot longer).

It also means that one can generate the table of historic releases and their counts much more easily as well, as the “root” node for each MSL has the counts for the whole MSL, so the page

<https://talk.ictvonline.org/taxonomy/p/taxonomy_releases>

can be generated with the simple and FAST query (and yes, I checked every value for every MSL – it matches exactly with PROD):

select name, notes, order\_desc\_ct, family\_desc\_ct, subfamily\_desc\_ct, genus\_desc\_ct, species\_desc\_ct

from taxonomy\_node

where msl\_release\_num is not null and level\_id = 100

order by tree\_id desc

and finally, Don’s “show pre-expanded nodes query”, 2.0: 9yb2.ddempsey\_show\_expanded\_nodes\_v2.sql

#### ALTER taxonomy\_node [hand-edited in Design GUI]

#### 0.b5.create\_udf\_rankCountsToStringWithPurals.sql

#### 0.b6.alter\_taxonomy\_node\_compute\_indexes.sql

#### 0.b7.create\_taxonomy\_node\_compute\_indexes\_ALL\_MSL.sql

#### 9yb2.ddempsey\_show\_expanded\_nodes\_v2.sql

### 9ya1.patch - remove Unyawo genus and species.sql

remove genus Unyawo & species "Xylella virus Paz"

### 9yc1.patch\_4genera\_into\_rhabdoviridae.sql

move 4 genera into family 'Rhabdoviridae'

* Barhavirus
* Lostrhavirus
* Sawgrhavirus
* Zarhavirus

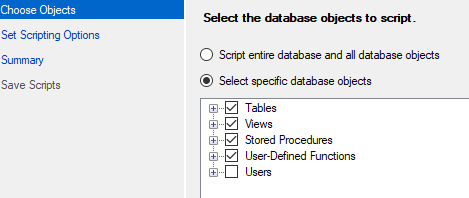
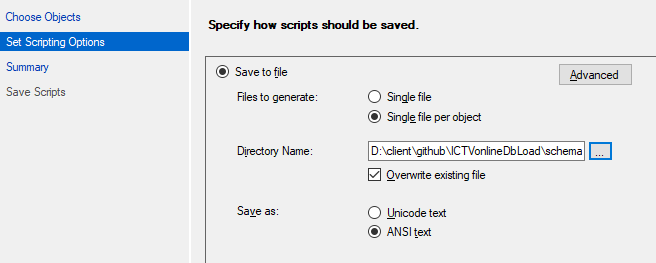
## BACKUP: ICTVonline35.20200422.0133.step\_9x\_done.rank\_cts.bak

Script database into « schema/ » directory

* First delete all SQL files in schema/, so we can « see » deletes.

cd schema

rm \*.sql

* Databases > ICTVonline35 > Tasks > Generate Scripts …
* Choose all objects, except users
* 
* One file per object, ANSI
* 
* Run schema/cleanup\_sql.sh to strip suffix from DB names : [ICTVonline35] -> [ICTVonline], to avoid uninteresting diff noise.

cd schema

./cleanup\_sql.sh