ICTV load next MSL (#35/2019)

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

Protocol steps completed up through 2.Z. Running 3.a overnight.

# 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

### 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'

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

Edit MSL # in Default constraint at bottom.

Run script to create this years load table.

### 0.b3.alter\_schema.sql

Add computed column

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

### 0.b4.create\_view-view\_taxa\_level\_counts\_by\_release.sql

Create a view Don needs for the website: [view\_taxa\_level\_counts\_by\_release]

Tried an alternate approach that had simpler SQL, but it was much slower: [view\_taxa\_level\_counts\_by\_release\_simplified]

### 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

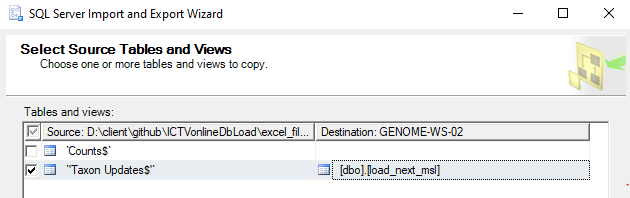
Destination:

Driver: SQL Server Native client 11

Genome-ws-02

ICTVonline35

Map to table



“Edit mappings”

### 0.c.insert\_msl\_into\_taxonomy\_toc.sql

Create row for MSL35

Skip 50,000, rather than 100k, as last year had 2 MSL.

### 0.d.drop\_obsolete\_tables.sql

Empty script this year

### 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

### 1.a.load\_next\_msl-delta-insert\_values.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 “1.a.load\_next\_msl-delta-insert\_values.sql”, save and run.

### ALTERNATE PATH:

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

Excel: save sheet as **Unicode** TXT

SQL: import via GUI

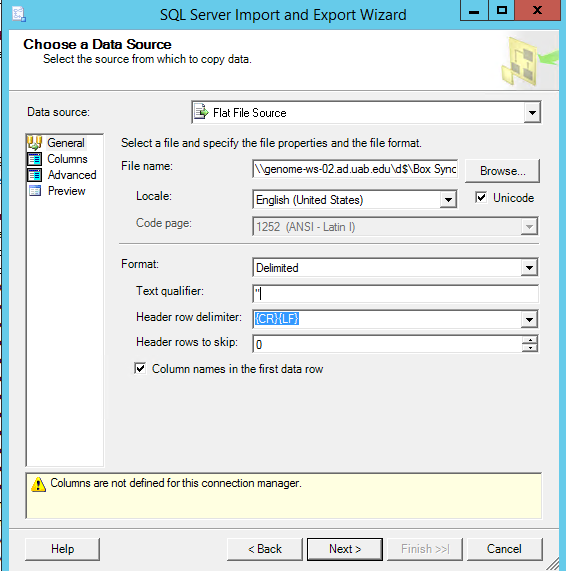
Drop the [load\_next\_msl\_unicode] first.

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]

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

None found/fixed.

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

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

Instead, please make backups often!

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

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

(171 rows affected)

COMMIT TRANSACTION

### 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? [review]

### 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 | 35 | abolish | 19 | 201900216 |
| action summary | 35 | merge | 1 | NULL |
| action summary | 35 | move | 224 | 201907070 |
| action summary | 35 | new | 1611 | NULL |
| action summary | 35 | rename | 52 | 201905366 |
| **action summary** | **35** | **all** | **1918** | **201907070** |

### 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

#### 2.c.load\_next\_msl-set-prev\_tax\_id.sql.xlsx

Report of all the problems found, and eventually corrected.

Correct by updating DB, faster than going back to 0 and re-loading.

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

1 error

* [FIXED] sort=1469; 2019.103B.zip; [Create new; assign as type species] taxon already exists: taxnode\_id=201850587, ictv\_id=20094401, lineage=Caudovirales;Podoviridae;Autographivirinae;Prochlorococcus virus PSSP7
  + Elliot: In the line before this, 1468, the species Prochlorococcus virus PSSP7 is abolished. But since sort 1469 creates it again in the new genus Tiamatvirus, this is really just a move. So I would delete sort line 1468, and change 1469 to a move.
  + Inserted row, sort=1468.5 , into load\_next\_msl
  + Set isWrong on both 1468 (abolish) and 1469 (new)

**2.Z.insert\_root\_in\_taxonomy\_node.sql**

Set release name

MSL=35

NAME=2019

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

TREE\_ID=201900000

*TR\_taxonomy\_node\_update\_indexes: tree[201900000] re-indexed*

(1 row affected)

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

Create new MSL by copying all nodes from prev MSL

SET @MSL=35

SET @TREE\_ID=201900000

(1 row affected)

(1 row affected)

TR\_taxonomy\_node\_update\_indexes: ROWS: DEL=0, INS=6898

TR\_taxonomy\_node\_update\_indexes: tree[201900000] re-indexed

(6898 rows affected)

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

-- SET load\_new\_msl: dest\_TAXNODE\_id, dest\_ICTV\_id and dest\_PARENT\_id

--

-- scan both taxa in taxonomy\_node(new MSL) and load\_next\_msl(new, split and renamed) things.

--

--

-- this must happen after

-- \* prev MSL is copied to create new one

-- \* load\_next\_msl is loaded

-- and before we start applying changes

#### 3.c.load\_next\_msl-set-dest\_taxnode\_id-dest\_parent\_id.sql.QC.xlsx

Fixed 7 taxa with typos that prevented linking to new parent\_id

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

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 not yet implemented; no promotions this year]

Results

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| report | action | prevMSL | load\_next\_msl | nextMSL | STATUS |
| counts by [ACTION] | abolish |  |  |  | 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 |

### 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

Results

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| report | action | prevMSL | load\_next\_msl | nextMSL | STATUS |
| counts by [ACTION] | abolish | 18 | 18 |  | 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 |

### 4.c.qc\_actions

Doublecheck the above.

Found 2 double entries:

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

* QC query turned up a 9 non-species taxa with no species under them.
* About half of these come from problems with splits (in some cases, a taxon was split to A,B,C, but there were also separate lines creating A,B,C, sometimes in a different parent taxon).
  + Split code doesn’t delete original taxon if one of the splits is not that name. [Issue #1](https://github.com/rusalkaguy/ICTVonlineDbLoad/issues/1)
* There were some other random ones – species specified in proposals, but not entered on the excel, a phylum created 3 times.
* I think I know how to solve all of these.
* I’d forgotten how long it takes to puzzle through all the twisted bits. I’m definitely moving the duplicate taxon name checking code much earlier in the process…. That would have caught several of these when they would take less typing to fix.

Progress tracked in ‘7a.qc\_empty\_taxa.sql.QC.xlsx’

* 20200414 First 5 issues resolved, 4 left.
  + 7a.qc\_empty\_taxa.x35fix1\_Autographivirinae.sql
  + 7a.qc\_empty\_taxa.x35fix2\_nucleorhabdorvius.sql
  + 7a.qc\_empty\_taxa.x35fix3\_Ambidensovirus.sql
  + 7a.qc\_empty\_taxa.x35fix4\_Ambidensovirus\_split.sql

#### **BACKUP**: ICTVonline35.20200415.0106.step\_7a\_5of9\_done.bak

* 20200415 fix 4 remaining issues
  + 7a.qc\_empty\_taxa.x35fix5\_cossaviricota\_new3x.sql
  + 7a.qc\_empty\_taxa.x35fix6\_redondoviridae.sql
  + 7a.qc\_empty\_taxa.x35fix7\_Escherichia\_virus\_NBD2.sql
  + 7a.qc\_empty\_taxa.x35fix8\_Gammanucleorhabdovirus.sql

#### **BACKUP**: ICTVonline35.20200415.1544.step\_7a\_done.bak

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

6 duplicate taxa.

* 7d.QC check for duplicate taxon names.x35fixNewNew.sql
* 7d.QC check for duplicate taxon names.x35fixNewSplit.sql

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

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

NEEDS UPDATE: Note that only one type species per genus, including all subgenera, is allowed.

Looks like there are a fair number of issues with this.

Many are un-resolved issues from previous MSLs.

[unfinished] I’m going to move on w/o putting time into this aspect until I have things pretty much finished.

I looked into the first one,

                Duplodnaviria;Heunggongvirae;Uroviricota;Caudoviricetes;Caudovirales;Demerecviridae;Novosibvirus: 2 type species

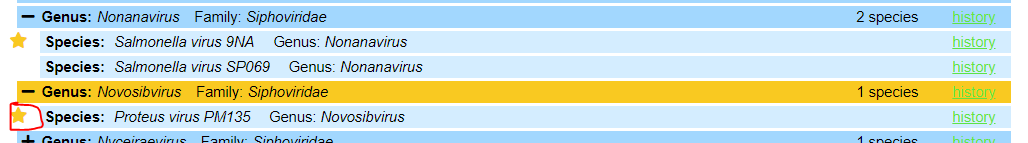
                MSL35 new species ‘Proteus virus Stubb’ isType=1

                2019.099B.Demerecviridae\_1fam3subfam6gen.xlsx

                However, MSL34 new species ‘[Proteus virus PM135’](https://talk.ictvonline.org/taxonomy/p/taxonomy-history?taxnode_id=201857060) isType=1

                                (side note, the history display doesn’t show that it was made type – didn’t it used to???)

                I suppose the most recent designation should win?



#### Came back after 9x.QC\_deltas.sql to fix

Changed query to define “exactly 1 type species under each genus (including all it’s subgenera, if any)”.

That cut down on the errors a lot.

Sent new report to Elliot.

##### 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

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

Added checks for correct transfer of load\_next\_msl.molecule into taxonomy\_node

Used update to fix problems.

[unfinished] There are many species with molecule\_id matching that of their parents, grandparents, but I don’t find code to push that up.

Updated REPORTS.

### 9x.QC\_deltas.sql

9x.QC\_deltas.xFix35.1\_split\_Ambidensovirussql.sql

9x.QC\_deltas.xFix35.2\_new\_wo\_ins.sql

Looks like we have to do this every year

9x.QC\_deltas.xFix35.3\_root\_out\_change.sql

Delta QC all clean!

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

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

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

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

Takes ## to run

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