Data export

Overview

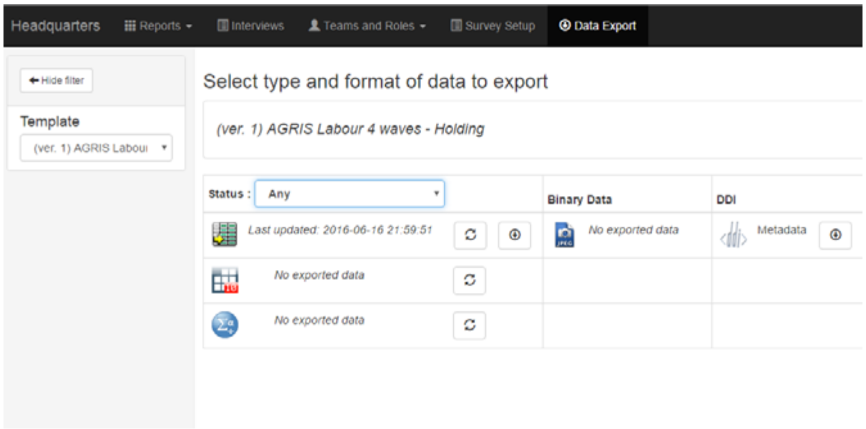
- When to export?
- How to export?
- What is exported?
- Structure of exported data files
- Interview Actions file

When to export?

- **FREQUENTLY!** Data export isn't just for exporting finalized data!
- WHY? Real time monitoring of data quality during collection can enable managers to detect and address problems immediately.
  - Detect fraudulent data, or enumerator mistakes.
  - Correct problems in the questionnaire.
  - Monitor precision.
  - If there's a listing exercise with CAPI, the list can be used as a sampling frame, and fed directly back into CAPI as pre-filled data.

When to export?

- Data can be exported at any time.
- It can be exported in .tab, .dta, or .sav.
- Binary and DDI compliant metadata separate.



Template: (ver. 1) AGRIS Labour

Select type and format of data to export

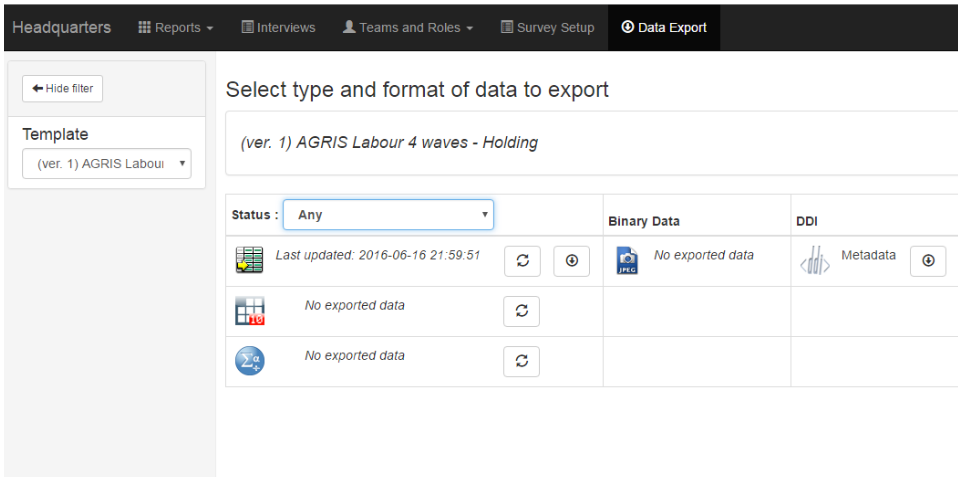
(ver. 1) AGRIS Labour 4 waves - Holding

Status : Any

	Binary Data	DDI
Last updated: 2016-06-16 21:59:51	*No exported data*	Metadata
No exported data		
No exported data		

How to Export?

- Data can only be exported by HQ or Admin.
- Select the template, click the arrows, then download.



Template

(ver. 1) AGRIS Laboui

Select type and format of data to export

(ver. 1) AGRIS Labour 4 waves - Holding

Status : Any	Binary Data	DDI
Last updated: 2016-06-16 21:59:51	*No exported data*	Metadata
No exported data		
No exported data		

What is exported?

- A zip file is exported from HQ containing 3 file types:
  - Microdata files
  - Interview_actions.tab
  - Comments file
- Each data file represents a different level of data.
  - Example: HH member roster, and questions about each HH member would be stored in separate files.

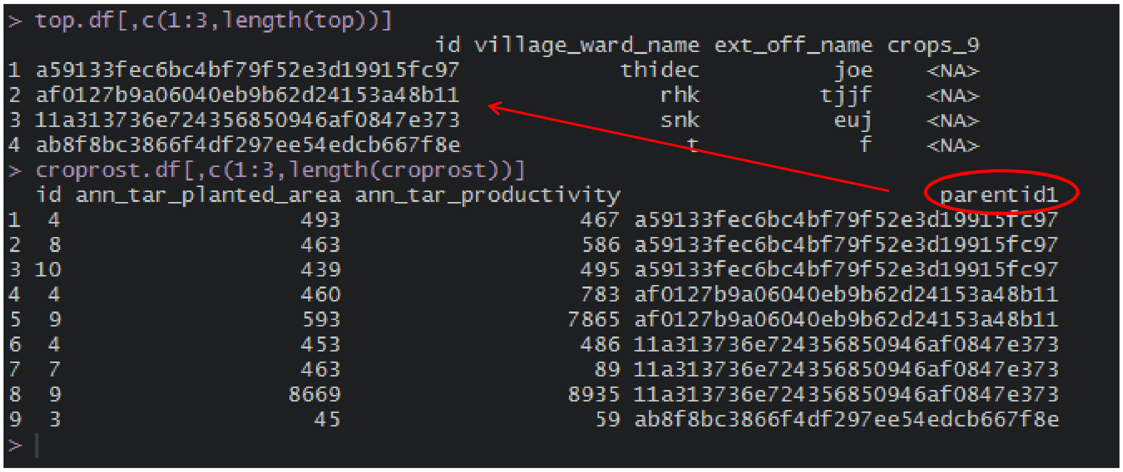
What is exported?

- For R users,
  – You can still take advantage of the categorical variable labels and coding contained in .dta and .sav files by reading them into R with the *foreign* package.

What is exported?

- More about levels of data files…
  – Often it is interesting to analyze datasets by different levels (i.e. urban/rural, household, individual). This is why the data is stored at different levels.
  – It is often necessary to merge these levels to have one aggregated dataset. Accordingly a unique Id is required that can facilitate the merge.

Structure of exported data files



```
> top.df[,c(1:3,length(top))]
                                id village_ward_name ext_off_name crops_9
1 a59133fec6bc4bf79f52e3d19915fc97            thidec          joe    <NA>
2 af0127b9a06040eb9b62d24153a48b11               rhk         tjjf    <NA>
3 11a313736e724356850946af0847e373               snk          euj    <NA>
4 ab8f8bc3866f4df297ee54edcb667f8e                 t            f    <NA>
> croprost.df[,c(1:3,length(croprost))]
  id ann_tar_planted_area ann_tar_productivity                        parentid1
1  4                  493                  467 a59133fec6bc4bf79f52e3d19915fc97
2  8                  463                  586 a59133fec6bc4bf79f52e3d19915fc97
3 10                  439                  495 a59133fec6bc4bf79f52e3d19915fc97
4  4                  460                  783 af0127b9a06040eb9b62d24153a48b11
5  9                  593                 7865 af0127b9a06040eb9b62d24153a48b11
6  4                  453                  486 11a313736e724356850946af0847e373
7  7                  463                   89 11a313736e724356850946af0847e373
8  9                 8669                 8935 11a313736e724356850946af0847e373
9  3                   45                   59 ab8f8bc3866f4df297ee54edcb667f8e
> 
```

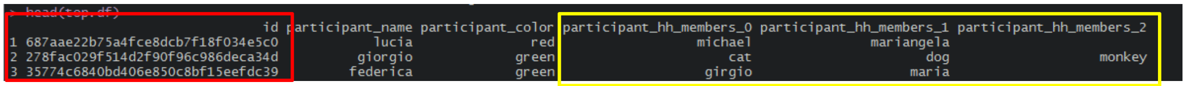
- *top.df* is the top level of data
- *croprost.df* is the second level of data coming from a crop roster.
- *top.df* and *croprost.df* can be merged on *croprost$parentid1* and *top$id.*

Structure of exported data files

- There will always be parentId, and ID variables allowing the user to merge datasets across different levels.
- Id is the unique identifier for that particular level.
- Parentid[#] relates that level of data to the one the next level up on the hierarchy starting with parentid1.

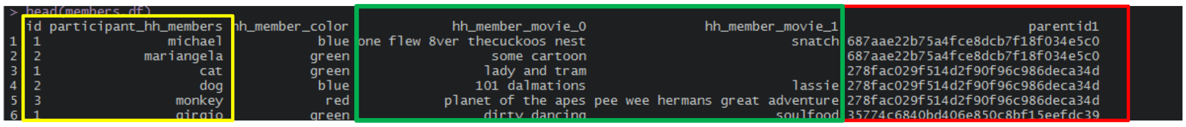
Structure of exported data files

Top-level data set, **id** = unique questionnaire id



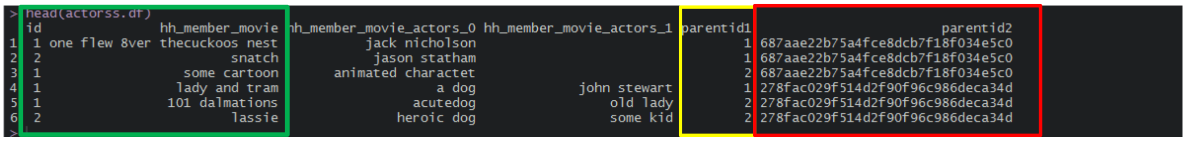
```
> head(top.df)
                                id participant_name participant_color participant_hh_members_0 participant_hh_members_1 participant_hh_members_2
1 687aae22b75a4fce8dcb7f18f034e5c0            lucia               red                  michael               mariangela
2 278fac029f514d2f90f96c986deca34d          giorgio             green                      cat                      dog                   monkey
3 35774c6840bd406e850c8bf15eefdc39         federica             green                   girgio                    maria
```

id = number of hh member, **parentid1** = unique questionnaire id



```
> head(members.df)
  id participant_hh_members hh_member_color             hh_member_movie_0                  hh_member_movie_1                        parentid1
1  1                michael            blue one flew 8ver thecuckoos nest                             snatch 687aae22b75a4fce8dcb7f18f034e5c0
2  2             mariangela           green                  some cartoon                                    687aae22b75a4fce8dcb7f18f034e5c0
3  1                    cat           green                  lady and tram                                   278fac029f514d2f90f96c986deca34d
4  2                    dog            blue                 101 dalmations                            lassie 278fac029f514d2f90f96c986deca34d
5  3                 monkey             red          planet of the apes pee wee hermans great adventure 278fac029f514d2f90f96c986deca34d
6  1                 girgio           green                 dirty dancing                           soulfood 35774c6840bd406e850c8bf15eefdc39
```

id = movie, **parentid1** = number of hh member, **parentid2** = unique questionnaire id



```
> head(actorss.df)
  id                hh_member_movie hh_member_movie_actors_0 hh_member_movie_actors_1 parentid1                        parentid2
1  1 one flew 8ver thecuckoos nest            jack nicholson                                  1 687aae22b75a4fce8dcb7f18f034e5c0
2  2                         snatch            jason statham                                  1 687aae22b75a4fce8dcb7f18f034e5c0
3  1                   some cartoon        animated charactet                                  2 687aae22b75a4fce8dcb7f18f034e5c0
4  1                  lady and tram                     a dog            john stewart         1 278fac029f514d2f90f96c986deca34d
5  1                 101 dalmations                 acutedog                old lady         2 278fac029f514d2f90f96c986deca34d
6  2                         lassie               heroic dog                some kid         2 278fac029f514d2f90f96c986deca34d
>
```

Structure of exported data files

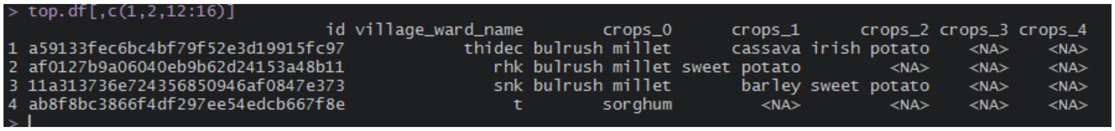
- Exported data follows the format of the question type.
  - Text -> exported as string
  - Numeric -> exported as string, dot is used as decimal separator.
  - Date -> UNIX: YYYY-MM-DDThh:mm:ss.s
  - Geo-location -> 4 separate columns
  - Categorical (1 answer) -> The numerical code is stored, and the label can be attached w/ do file.

Structure of exported data files

- Multi-select
  - Multiple variables created in dataset w/ indices 1,2, etc. For example, {variablename_1, variablename_2,…, variablename_n}.
  - For unordered questions, the value will be 1 for selected items, and 0 for unselected items.
  - For ordered questions, variable with index 1(item_1) will contain the first option selected, and index n (item_n) will contain the nth item selected.
  - For Y/N, each datapoint is a 0 or the number representing the order of selection or “Yes”.

Structure of exported data files

- Format continued…
  - Categorical: multiple answers:



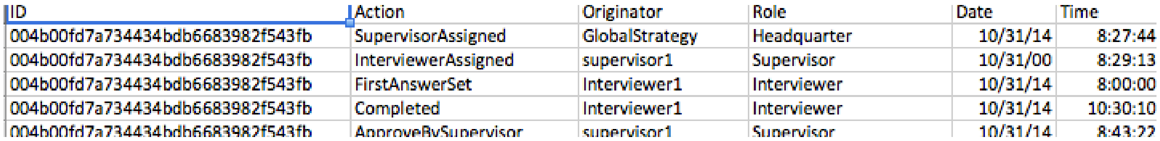
```
> top.df[,c(1,2,12:16)]
                                id village_ward_name        crops_0      crops_1      crops_2 crops_3 crops_4
1 a59133fec6bc4bf79f52e3d19915fc97            thidec bulrush millet      cassava irish potato    <NA>    <NA>
2 af0127b9a06040eb9b62d24153a48b11               rhk bulrush millet sweet potato         <NA>    <NA>    <NA>
3 11a313736e724356850946af0847e373               snk bulrush millet       barley sweet potato    <NA>    <NA>
4 ab8f8bc3866f4df297ee54edcb667f8e                 t        sorghum         <NA>         <NA>    <NA>    <NA>
> 
```

Structure of exported data files

- Format continued…
  - Lists -> Multiple variables are created in the export file with an index added at the end of the name. Example, if there multiple names {name_0, name_1, name_2,…,name_n}

Interview Actions file

- Each export zip file contains a Interview_actions.tab. This file contains a time and date stamp for each event in the life of a survey and the originator/role of originator.
- This information is very useful for monitoring data collection.



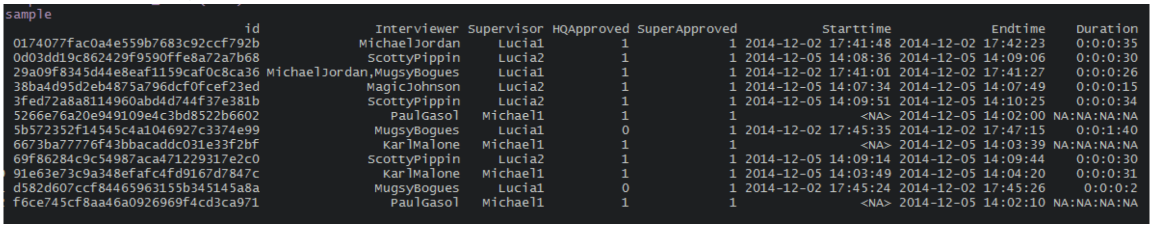
ID	Action	Originator	Role	Date	Time
004b00fd7a734434bdb6683982f543fb	SupervisorAssigned	GlobalStrategy	Headquarter	10/31/14	8:27:44
004b00fd7a734434bdb6683982f543fb	InterviewerAssigned	supervisor1	Supervisor	10/31/00	8:29:13
004b00fd7a734434bdb6683982f543fb	FirstAnswerSet	Interviewer1	Interviewer	10/31/14	8:00:00
004b00fd7a734434bdb6683982f543fb	Completed	Interviewer1	Interviewer	10/31/14	10:30:10
004b00fd7a734434bdb6683982f543fb	ApproveBySupervisor	supervisor1	Supervisor	10/31/14	8:43:22

Interview Actions file

- Tabulations of this data can provide insights about enumerator performance, supervisor performance, length of time of interviews, etc.
- I’ve written R functions to create tabulations by interview, enumerator, and supervisor. I will make these available through Github. Examples:

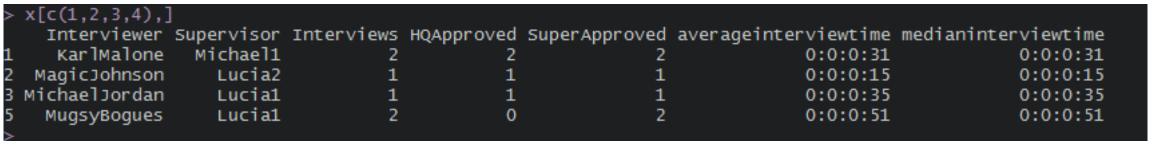
Interview Actions file

Tabulated by interview



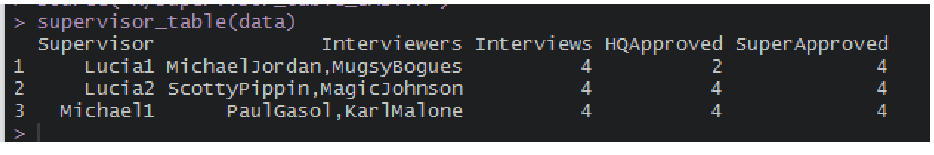
```
sample
                               id              Interviewer Supervisor HQApproved SuperApproved           Starttime             Endtime    Duration
 0174077fac0a4e559b7683c92ccf792b            MichaelJordan     Lucia1          1             1 2014-12-02 17:41:48 2014-12-02 17:42:23    0:0:0:35
 0d03dd19c862429f9590ffe8a72a7b68             ScottyPippin     Lucia2          1             1 2014-12-05 14:08:36 2014-12-05 14:09:06    0:0:0:30
 29a09f8345d44e8eaf1159caf0c8ca36 MichaelJordan,MugsyBogues     Lucia1          1             1 2014-12-02 17:41:01 2014-12-02 17:41:27    0:0:0:26
 38ba4d95d2eb4875a796dcf0fcef23ed             MagicJohnson     Lucia2          1             1 2014-12-05 14:07:34 2014-12-05 14:07:49    0:0:0:15
 3fed72a8a8114960abd4d744f37e381b             ScottyPippin     Lucia2          1             1 2014-12-05 14:09:51 2014-12-05 14:10:25    0:0:0:34
 5266e76a20e949109e4c3bd8522b6602                PaulGasol   Michael1          1             1                <NA> 2014-12-05 14:02:00 NA:NA:NA:NA
 5b572352f14545c4a1046927c3374e99              MugsyBogues     Lucia1          0             1 2014-12-02 17:45:35 2014-12-02 17:47:15    0:0:1:40
 6673ba77776f43bbacaddc031e33f2bf               KarlMalone   Michael1          1             1                <NA> 2014-12-05 14:03:39 NA:NA:NA:NA
 69f86284c9c54987aca471229317e2c0             ScottyPippin     Lucia2          1             1 2014-12-05 14:09:14 2014-12-05 14:09:44    0:0:0:30
 91e63e73c9a348efafc4fd9167d7847c               KarlMalone   Michael1          1             1 2014-12-05 14:03:49 2014-12-05 14:04:20    0:0:0:31
 d582d607ccf84465963155b345145a8a              MugsyBogues     Lucia1          0             1 2014-12-02 17:45:24 2014-12-02 17:45:26     0:0:0:2
 f6ce745cf8aa46a0926969f4cd3ca971                PaulGasol   Michael1          1             1                <NA> 2014-12-05 14:02:10 NA:NA:NA:NA
```

Tabulated by interviewer



```
> x[c(1,2,3,4),]
    Interviewer Supervisor Interviews HQApproved SuperApproved averageinterviewtime medianinterviewtime
1    KarlMalone   Michael1          2          2             2             0:0:0:31            0:0:0:31
2  MagicJohnson     Lucia2          1          1             1             0:0:0:15            0:0:0:15
3 MichaelJordan     Lucia1          1          1             1             0:0:0:35            0:0:0:35
5   MugsyBogues     Lucia1          2          0             2             0:0:0:51            0:0:0:51
>
```

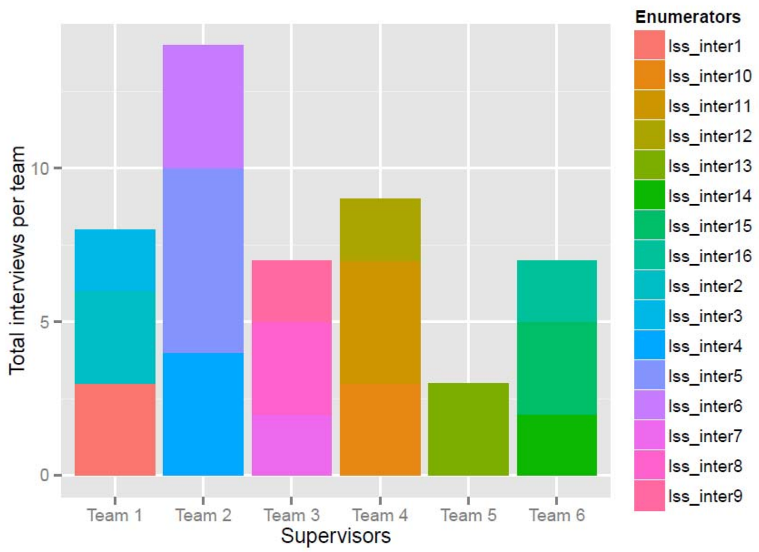
Tabulated by supervisor



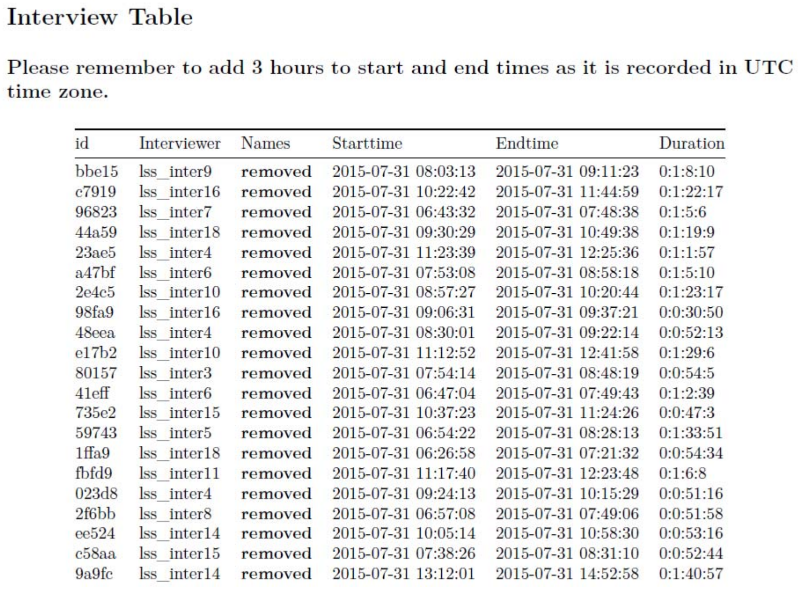
```
> supervisor_table(data)
  Supervisor               Interviewers Interviews HQApproved SuperApproved
1     Lucia1 MichaelJordan,MugsyBogues          4          2             4
2     Lucia2 ScottyPippin,MagicJohnson          4          4             4
3   Michael1       PaulGasol,KarlMalone          4          4             4
> |
```

Interview Actions file

Team Summary Plot



Interview Actions file



Interview Table

Please remember to add 3 hours to start and end times as it is recorded in UTC time zone.

id	Interviewer	Names	Starttime	Endtime	Duration
bbe15	lss_inter9	**removed**	2015-07-31 08:03:13	2015-07-31 09:11:23	0:1:8:10
c7919	lss_inter16	**removed**	2015-07-31 10:22:42	2015-07-31 11:44:59	0:1:22:17
96823	lss_inter7	**removed**	2015-07-31 06:43:32	2015-07-31 07:48:38	0:1:5:6
44a59	lss_inter18	**removed**	2015-07-31 09:30:29	2015-07-31 10:49:38	0:1:19:9
23ae5	lss_inter4	**removed**	2015-07-31 11:23:39	2015-07-31 12:25:36	0:1:1:57
a47bf	lss_inter6	**removed**	2015-07-31 07:53:08	2015-07-31 08:58:18	0:1:5:10
2e4c5	lss_inter10	**removed**	2015-07-31 08:57:27	2015-07-31 10:20:44	0:1:23:17
98fa9	lss_inter16	**removed**	2015-07-31 09:06:31	2015-07-31 09:37:21	0:0:30:50
48eea	lss_inter4	**removed**	2015-07-31 08:30:01	2015-07-31 09:22:14	0:0:52:13
e17b2	lss_inter10	**removed**	2015-07-31 11:12:52	2015-07-31 12:41:58	0:1:29:6
80157	lss_inter3	**removed**	2015-07-31 07:54:14	2015-07-31 08:48:19	0:0:54:5
41eff	lss_inter6	**removed**	2015-07-31 06:47:04	2015-07-31 07:49:43	0:1:2:39
735e2	lss_inter15	**removed**	2015-07-31 10:37:23	2015-07-31 11:24:26	0:0:47:3
59743	lss_inter5	**removed**	2015-07-31 06:54:22	2015-07-31 08:28:13	0:1:33:51
1ffa9	lss_inter18	**removed**	2015-07-31 06:26:58	2015-07-31 07:21:32	0:0:54:34
fbfd9	lss_inter11	**removed**	2015-07-31 11:17:40	2015-07-31 12:23:48	0:1:6:8
023d8	lss_inter4	**removed**	2015-07-31 09:24:13	2015-07-31 10:15:29	0:0:51:16
2f6bb	lss_inter8	**removed**	2015-07-31 06:57:08	2015-07-31 07:49:06	0:0:51:58
ee524	lss_inter14	**removed**	2015-07-31 10:05:14	2015-07-31 10:58:30	0:0:53:16
c58aa	lss_inter15	**removed**	2015-07-31 07:38:26	2015-07-31 08:31:10	0:0:52:44
9a9fc	lss_inter14	**removed**	2015-07-31 13:12:01	2015-07-31 14:52:58	0:1:40:57

Questions??