Data Quality Review (DQR) Data Verification and System Assessment Workshop

Session 13

Data Compilation, Cleaning, and Quality Assurance

World Health Organization

USAID
FROM THE AMERICAN PEOPLE

MEASURE Evaluation

The Global Fund

Learning Objective

To understand the importance of, and techniques for, data cleaning, compilation, and quality assurance, and;

- Understand the mechanisms and resource requirements for data entry and cleaning of survey data (for paper-based data collection)
- Understand the mechanisms and resource requirements for compiling and cleaning electronic data in CSPro (and other software systems)
- Understand the purpose and use of tools in CSPro for assessing data completeness and quality (e.g. CSPro batch files for assessing completeness and comparing data collection and supervisor validation files)
- Be aware of other techniques for assessing the quality of survey implementation, e.g. re-assessment and comparison of collected data for a 5% sample.

SESSION 13

Data Compilation, Cleaning, and Quality Assurance

Compilation

SESSION 13

Data Compilation, Cleaning, and Quality Assurance

Data Compilation

- Once the data have been collected, they need to be compiled into a master data file for analysis to understand the results of the data quality review.
- If the data have been collected on paper forms, they should now be entered in the CSPro database.
- If they have been entered on the tablet computers, the data from the different data collection teams should be merged/downloaded into one data file.
- The data should be reviewed for quality and corrected, if necessary.
- Automated tools in CSPro can help identify gaps and inconsistencies in the collected data.

Data Entry

SESSION 13

Data Compilation, Cleaning, and Quality Assurance

Data Entry in CSPro

- Depending on the size of the health facility assessment (sample size) and the number of people working, data entry make require several weeks if data is collected on paper.
- Data for each facility should be entered twice, and the records compared, to assess and ensure good quality (double data entry).
- Once the data are in the computer, the quality of data entry should be checked by sampling records and reviewing the accuracy by comparing the electronic record to the paper-based form.

Data Cleaning

SESSION 13

Data Compilation, Cleaning, and Quality Assurance

Concatenate

- If the data have been entered into tablets in the field, they need to be merged into one data file.
- There are two potential processes for this depending on how the survey was deployed.
  - If the data was collected on Android tablets, the Data Viewer tool will be used to download the complete dataset.
  - If the data was collected on Windows computers, the Concatenate tool will be used to merge the data files into a single master file.

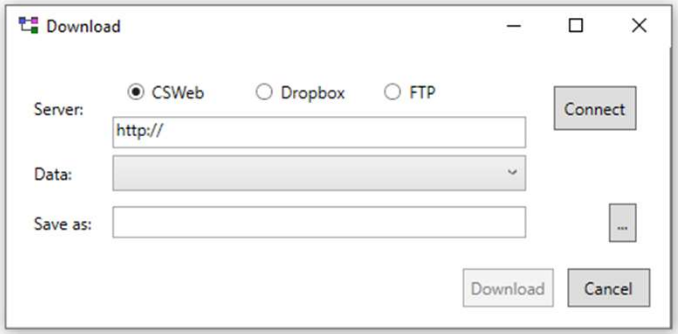
Data Cleaning

SESSION 13

Data Compilation, Cleaning, and Quality Assurance

Concatenate – Data viewer tool

- When CSPro stores data in on the server it stores the data in a format specific to CSPro. CSPro uses this format to allow synchronization at the case level.
- In order to access, the data, the Data Viewer tool must be used to concatenate all the data files from the tablets and convert the file to a CSPro database file.
- Open CSPro and navigate to the Tools menu and select Data Viewer
- From the file menu of the Data Viewer select "Download"
- Select the server you used, enter your login credentials, select the data file you wish to download, select the location where you would like to save the dataset.



Download

Server: ● CSWeb ○ Dropbox ○ FTP

http://

Data:

Save as:

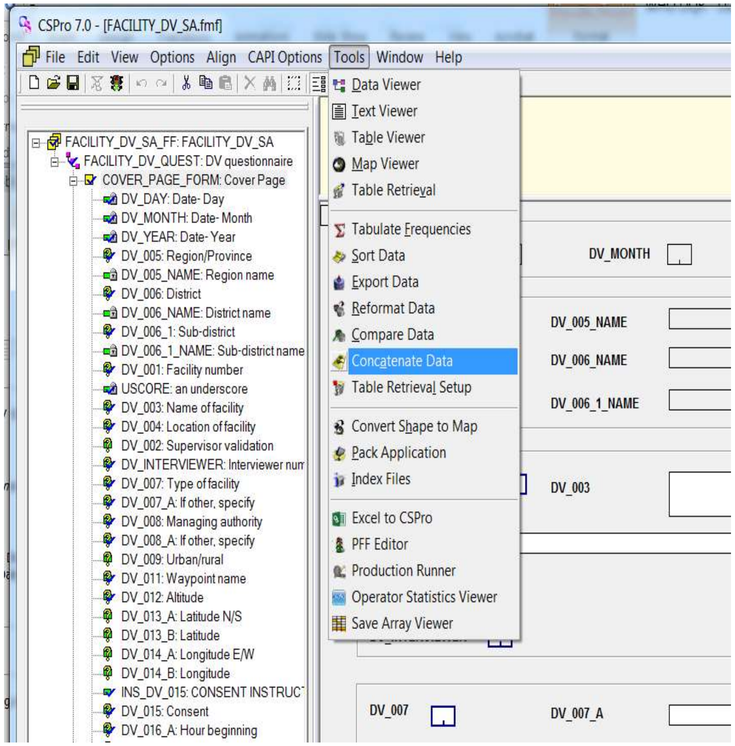
Data Cleaning

SESSION 13

Data Compilation, Cleaning, and Quality Assurance

Concatenate – Concatenate tool

- If the data have been entered on non-android tablets in the field, they need to be merged into one data file.
- Download the CSPro data files from the server. There should be one per data collection team. Put them in all in the same folder on your computer.



- Open CSPro and navigate to the Tools menu and select Concatenate Data

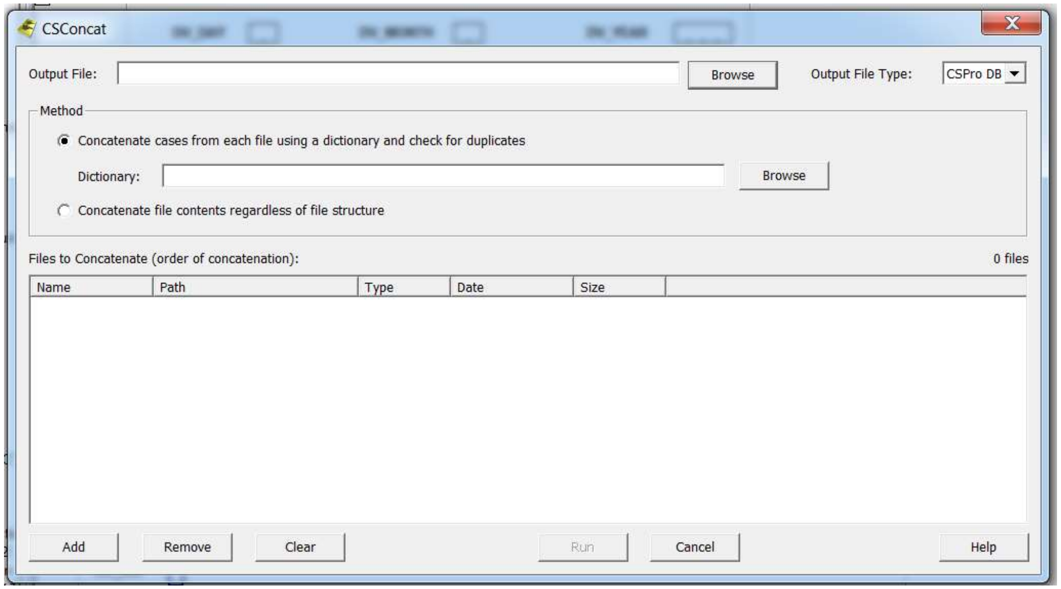
Data Cleaning

SESSION 13

Data Compilation, Cleaning, and Quality Assurance

Concatenate

- Specify the output file name in "Output file:"
- Specify the CSPro data dictionary FACILITY_DV_SA.dcf
- Click add and navigate to the folder containing the data files.
- Select all the data files and click "run".



Output File:

Output File Type: CSPro DB

Method

● Concatenate cases from each file using a dictionary and check for duplicates

Dictionary:

○ Concatenate file contents regardless of file structure

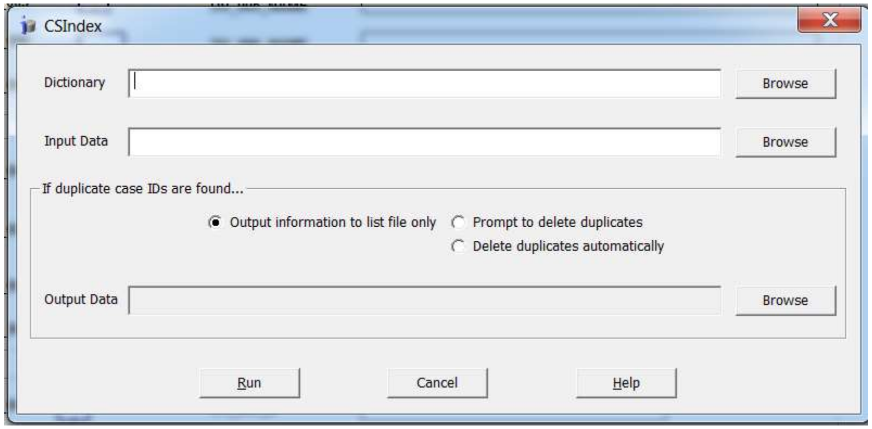
Files to Concatenate (order of concatenation): 0 files

Name	Path	Type	Date	Size

Data Cleaning

Index function

- Often when data files are compiled from the field, they contain duplicates. Run the Index function to remove the duplicates.
- Specify the CSPro data dictionary FACILITY_DV_SA.dcf.
- Specify the input data file, select "Output information to list file only" and select "run".



Dictionary

Input Data

If duplicate case IDs are found...

● Output information to list file only ○ Prompt to delete duplicates
○ Delete duplicates automatically

Output Data

- Examine the output list file (CSIndex.lst file in the default folder) and identify the duplicates.
- From the CSPro data entry module reconcile the duplicates into one record (update missing values in one record, if applicable, and delete the other record.

SESSION 13

Data Compilation, Cleaning, and Quality Assurance

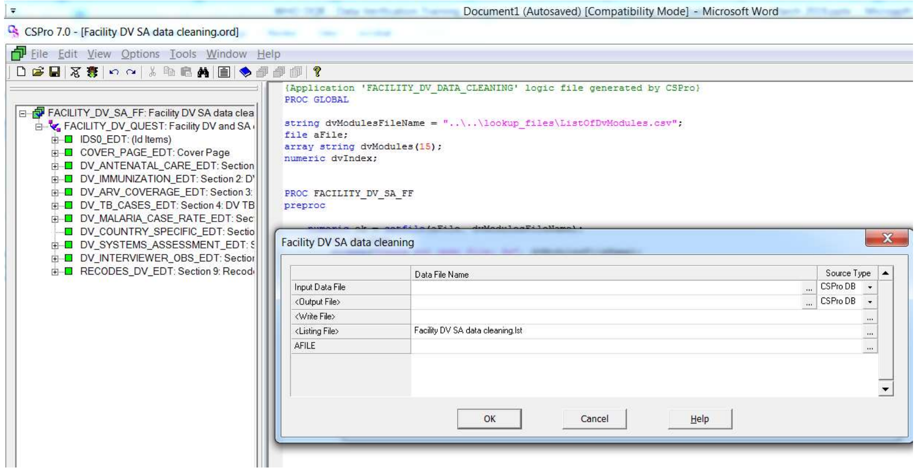
Data Cleaning

SESSION 13

Data Compilation, Cleaning, and Quality Assurance

Cleaning Batch File

- The Cleaning batch file identifies gaps (missing values) in the data file.
- This tool should be used regularly during data collection to examine the completeness of data collection.
- Feedback can then be provided to data collection teams who can potentially update the missing values.


Document1 (Autosaved) [Compatibility Mode] - Microsoft Word
CSPro 7.0 - [Facility DV SA data cleaning.ord]
File
Edit
View
Options
Tools
Window
Help
Facility DV SA data cleaning
OK
Cancel
Help

FACILITY_DV_SA_FF: Facility DV SA data clea
FACILITY_DV_QUEST: Facility DV and SA
IDS0_EDT: (Id Items)
COVER_PAGE_EDT: Cover Page
DV_ANTENATAL_CARE_EDT: Section
DV_IMMUNIZATION_EDT: Section 2: D
DV_ARV_COVERAGE_EDT: Section 3:
DV_TB_CASES_EDT: Section 4: DV TB
DV_MALARIA_CASE_RATE_EDT: Sec
DV_COUNTRY_SPECIFIC_EDT: Sectio
DV_SYSTEMS_ASSESSMENT_EDT: S
DV_INTERVIEWER_OBS_EDT: Sectior
RECODES_DV_EDT: Section 9: Recod

```
{Application 'FACILITY_DV_DATA_CLEANING' logic file generated by CSPro}
PROC GLOBAL

string dvModulesFileName = "..\..\lookup_files\ListOfDvModules.csv";
file aFile;
array string dvModules(15);
numeric dvIndex;


PROC FACILITY_DV_SA_FF
preproc
```

	Data File Name	Source Type
Input Data File		CSPro DB
<Output File>		CSPro DB
<Write File>		
<Listing File>	Facility DV SA data cleaning.lst	
AFILE		

- Click on the "Facility DV SA data cleaning.bch" in the "Batch 1" subfolder in the FACILITY_DV_SA folder.
- Specify the compiled data file and the output list file.
- The output file shows a list of facilities and missing data elements.

Data Cleaning

SESSION 13

Data Compilation, Cleaning, and Quality Assurance

Data entry interface for data review and cleaning

- The following steps should be taken to review the data in CSPro:
  - Step 1: Open the concatenated data set using the data entry application
  - Step 2: Review cases for key fields
    - Facility name and ID number correspond to each other
    - Facility ID information is correct (facility type, managing authority)
    - Interviewer ID information is present and correct
    - GPS coordinates are valid (if applicable)
    - Identify "other" responses for recoding as applicable
  - Step 3: Check for any duplicate facility codes
    - Export data to excel
    - Use conditional formatting to identify duplicate facility IDs

- Step 4: Delete any empty or duplicate cases
  - Duplicate cases are cases with the same facility code. If two cases appear to be duplicates according to facility name, but do not contain the same data, a list of criteria must be used to determine if it is a true duplicate.
  - The following data elements could be used as the criteria for determining duplicates. If these are all the same, it is safe to consider the cases as duplicates.
    - district
    - facility code/name
    - GPS coordinates (*if collected*)
    - facility type
    - managing authority
    - interviewer's code
- At this point, the most complete case should stay in the data set. If both cases are complete, the case with latest time stamp should be kept.

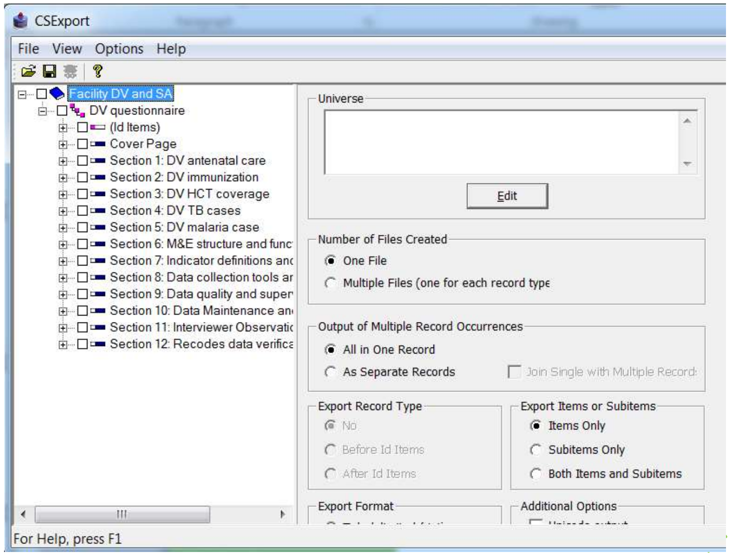
Data Export

SESSION 13

Data Compilation, Cleaning, and Quality Assurance

Data Export Function

- The data can be exported to MS Excel to assist with data cleaning.
- Open CSPro and select "Export Data" on the "Tools" menu.
- Specify the dictionary file when prompted.
- Check the box to the left of "Facility DV and SA" at the top of the tree in the left pane to select all data elements for output.
- Use the default settings of "one file", "all in one record", "items only.
- Select Tab Delimited and click on the stoplight icon on the menu bar to run the export function.
- The text file can then be imported into MS Excel.



- □ Facility DV and SA
  - □ DV questionnaire
    - □ (Id Items)
    - □ Cover Page
    - □ Section 1: DV antenatal care
    - □ Section 2: DV immunization
    - □ Section 3: DV HCT coverage
    - □ Section 4: DV TB cases
    - □ Section 5: DV malaria case
    - □ Section 6: M&E structure and func
    - □ Section 7: Indicator definitions anc
    - □ Section 8: Data collection tools ar
    - □ Section 9: Data quality and super
    - □ Section 10: Data Maintenance an
    - □ Section 11: Interviewer Observatic
    - □ Section 12: Recodes data verifica

Universe

Number of Files Created
● One File
○ Multiple Files (one for each record type

Output of Multiple Record Occurrences
● All in One Record
○ As Separate Records □ Join Single with Multiple Record:

Export Record Type
● No
○ Before Id Items
○ After Id Items

Export Items or Subitems
● Items Only
○ Subitems Only
○ Both Items and Subitems

Export Format

Additional Options

Compare data files

SESSION 13

Data Compilation, Cleaning, and Quality Assurance

Supervisor Split

- For quality assurance of data collection often the survey is implemented a second time at a small sample of facilities to compare the data collected. This is usually done by a supervisor, but sometimes by an external quality assurance team.
- Records in the database can be identified as either "interviewer" or "supervisor" by the value in the field Q002; 1 for interviewer, 2 for supervisor.
- To compare the records between interviewers and supervisors the records need to be split into two datasetsTo split the database, double click on the "Supervisorsplit.bch" batch file in the Batch 2 folder and click the stoplight.
- This will run a batch application to create two datasets: 2_FACILITY_DV_SA_FINALDATA.csdb which includes all the original data collector records and 3_FACILITY_DV_SA_SUPERVISORDATA.csdb which contains only the supervisor validations.
- When it is finished running, a report of the process will open. You can close this window when it is complete. Browse to FACILITY_DV_SA\Data and check to see that the two files are present.

Compare data files

SESSION 13

Data Compilation, Cleaning, and Quality Assurance

Compare Function

- The Compare Data tool is a CSPro tool that allows you to compare two data files and identify the differences.The data files must have the same structure, that is, be described by the same CSPro data dictionary. In addition, for comparisons to be made, the original case and the supervisor validation must have the same ID information and be in different data sets.
- In order to compare the supervisor validation record with the original facility record the two must have the same exact ID information. In order to accomplish this, please take the following steps.
- Press the Start button and navigate to the Programs -> CSPro7.0 -> Data tools -> Export data.
- The first screen in the CSPro Data Export application will ask you for the Data Dictionary File. Navigate to the FACILITY_DV_SA folder, select the FACILITY_DV_SA.dcf file, and click Open.

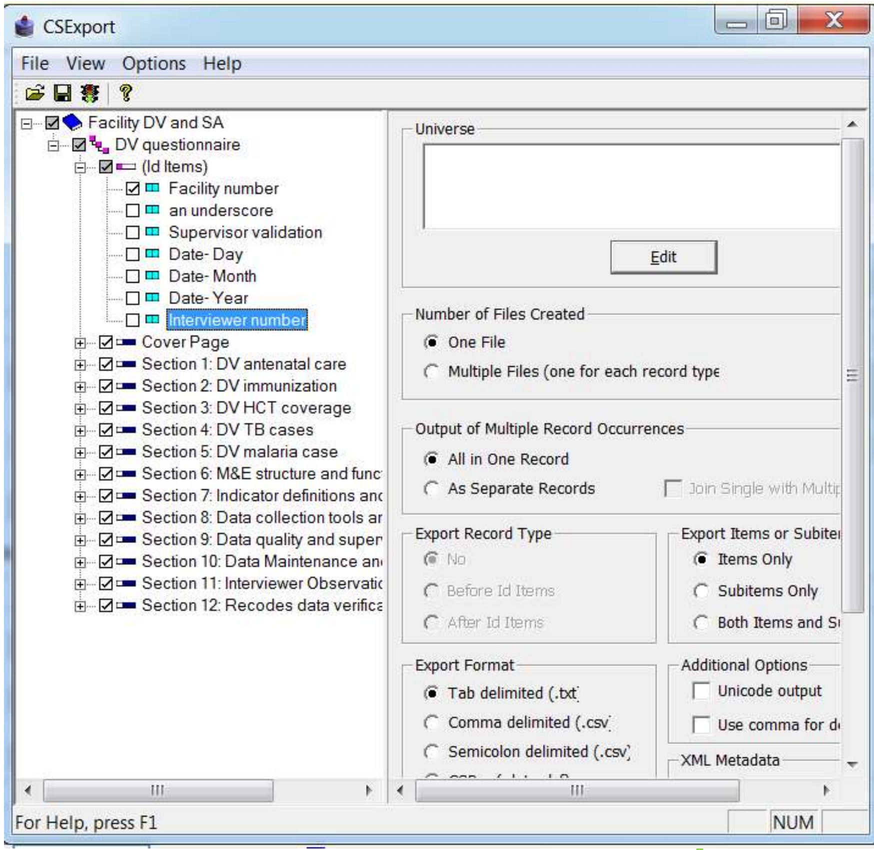
Compare data files

SESSION 13

Data Compilation, Cleaning, and Quality Assurance

Compare Function

- The panel on the left should now display the data dictionary's records and items in a selectable dictionary tree. From this dictionary tree you can select the data you would like to export. Click on the top box next to the dictionary icon to select all the data.
- Next, open the tree called ID items. De-select everything EXCEPT facility number. The tree should look like the image at right.



☑ Facility DV and SA
☑ DV questionnaire
☑ (Id Items)
☑ Facility number
□ an underscore
□ Supervisor validation
□ Date- Day
□ Date- Month
□ Date- Year
□ Interviewer number
☑ Cover Page
☑ Section 1: DV antenatal care
☑ Section 2: DV immunization
☑ Section 3: DV HCT coverage
☑ Section 4: DV TB cases
☑ Section 5: DV malaria case
☑ Section 6: M&E structure and func
☑ Section 7: Indicator definitions anc
☑ Section 8: Data collection tools ar
☑ Section 9: Data quality and super
☑ Section 10: Data Maintenance an
☑ Section 11: Interviewer Observatic
☑ Section 12: Recodes data verifica

Universe

Number of Files Created
● One File
○ Multiple Files (one for each record type

Output of Multiple Record Occurrences
● All in One Record
○ As Separate Records
□ Join Single with Multip

Export Record Type
● No
○ Before Id Items
○ After Id Items

Export Items or Subiter
● Items Only
○ Subitems Only
○ Both Items and S

Export Format
● Tab delimited (.txt
○ Comma delimited (.csv
○ Semicolon delimited (.csv)

Additional Options
□ Unicode output
□ Use comma for d

XML Metadata

For Help, press F1 NUM

Compare data files

SESSION 13

Data Compilation, Cleaning, and Quality Assurance

Compare Function

- This screen will also display various export options, such as the export format, how many files you would like the application to create, and whether you want to include XML Metadata. We will keep the default options for everything except the Export Format. Please select CSPro (.dat, .dcf) from the export format options.
- To export the data, click on the stoplight on the toolbar, or select Run from the File menu.
- The next screen will ask you to select the data file you would like to export. Please browse to your concatenated, de-duplicated data file and click Open.
- CSPro will then ask you to specify the name of the output data file and dictionary file. Browse to the FACILITY_DV_SA\Data\Supervisor validation folder and name the output data file 4_FACILITY_DV_SA_FINALDATA_EXTRACT.csdb and save the dictionary in FACILITY_DV_SA\Data\Supervisor validation as FACILITY_DV_SA_data_compare.dcf. Then click ok to run.
- Repeat steps with the 3_FACILITY_DV_SA_SUPERVISORDATA.csdb dataset. Saving the file in the FACILITY_DV_SA\Data\Supervisor validation folder as 5_FACILITY_DV_SA_SUPERVISORDATA_EXTRACT.csdb.

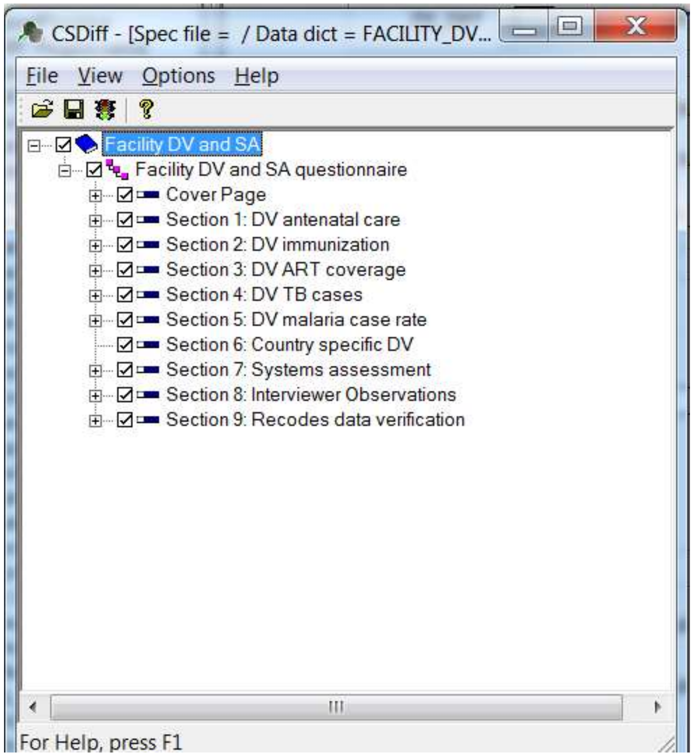
Compare data files

SESSION 13

Data Compilation, Cleaning, and Quality Assurance

The data files are now ready to use the compare tool.

- Navigate to Programs -> CSPro7.0 -> Data tools -> Compare data.
- The first screen in the CSPro Compare data application will ask you for the Data Dictionary File. Navigate to FACILITY_DV_SA\Data\Supervisor validation, select the FACILITY_DV_SA_data_compare.dcf file, and click Open.
- The panel on the left should now display the data dictionary's records and items in a selectable dictionary tree.
- Click on the top box next to the dictionary icon to select all the data. The screen should look like the image to the right.



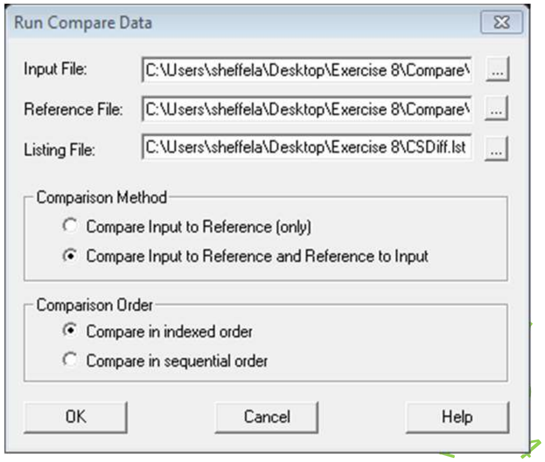
- ☑ Facility DV and SA
  - ☑ Facility DV and SA questionnaire
    - ☑ Cover Page
    - ☑ Section 1: DV antenatal care
    - ☑ Section 2: DV immunization
    - ☑ Section 3: DV ART coverage
    - ☑ Section 4: DV TB cases
    - ☑ Section 5: DV malaria case rate
    - ☑ Section 6: Country specific DV
    - ☑ Section 7: Systems assessment
    - ☑ Section 8: Interviewer Observations
    - ☑ Section 9: Recodes data verification

Compare data files

SESSION 13

Data Compilation, Cleaning, and Quality Assurance

- To run the Compare function, click "Run" on the toolbar; press Ctrl+R; or from the File menu, select Run.
- For the input file, select the 4_FACILITY_DV_SA_FINALDATA_EXTRACT.csdb file from the FACILITY_DV_SA\Data\Supervisor validation folder. For the reference file, select the 5_FACILITY_DV_SA_SUPERVISORDATA_EXTRACT.csdb from the FACILITY_DV_SA\Data\Supervisor validation folder.
- For the comparison method, make sure the "Compare Input to Reference and Reference to Input" box is selected.
- For the comparison method, make sure the "Compare in Indexed Order" box is selected. The screen should look like the image below.
- Click OK to run the Compare tool. An output summarizing the results of the file comparison will be shown.
- Examine the output. The output should look like the image on the following slide.



Run Compare Data

Input File: C:\Users\sheffela\Desktop\Exercise 8\Compare\

Reference File: C:\Users\sheffela\Desktop\Exercise 8\Compare\

Listing File: C:\Users\sheffela\Desktop\Exercise 8\CSDiff.lst

Comparison Method

○ Compare Input to Reference (only)

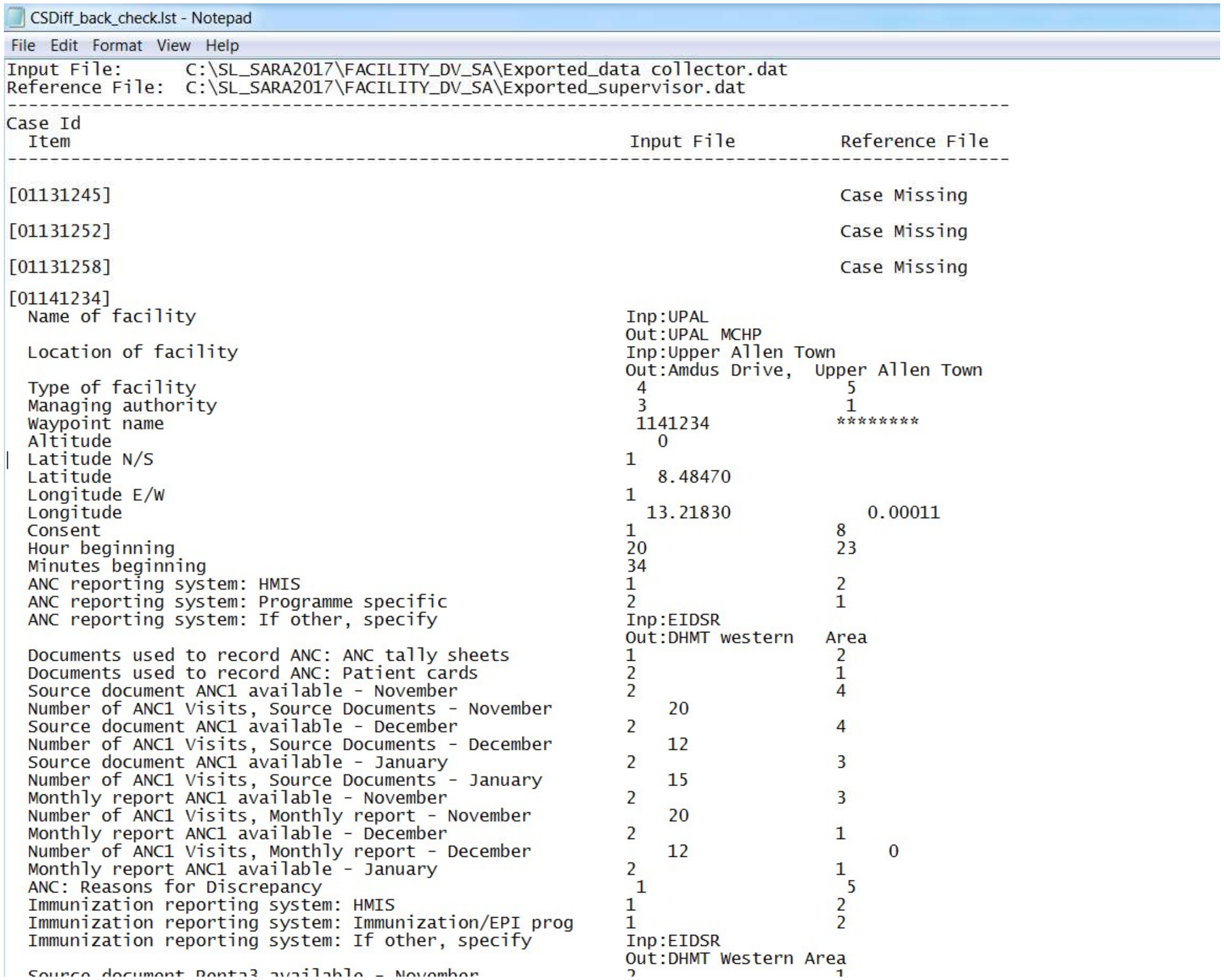
● Compare Input to Reference and Reference to Input

Comparison Order

● Compare in indexed order

○ Compare in sequential order

Compare data files



```
Input File:      C:\SL_SARA2017\FACILITY_DV_SA\Exported_data collector.dat
Reference File:  C:\SL_SARA2017\FACILITY_DV_SA\Exported_supervisor.dat
------------------------------------------------------------------------------------------------
Case Id
  Item                                                      Input File          Reference File
------------------------------------------------------------------------------------------------

[01131245]                                                                      Case Missing

[01131252]                                                                      Case Missing

[01131258]                                                                      Case Missing

[01141234]
  Name of facility                                          Inp:UPAL
                                                            Out:UPAL MCHP
  Location of facility                                      Inp:Upper Allen Town
                                                            Out:Amdus Drive,  Upper Allen Town
  Type of facility                                           4                   5
  Managing authority                                         3                   1
  Waypoint name                                              1141234            ********
  Altitude                                                     0
  Latitude N/S                                              1
  Latitude                                                     8.48470
  Longitude E/W                                             1
  Longitude                                                   13.21830             0.00011
  Consent                                                   1                   8
  Hour beginning                                            20                  23
  Minutes beginning                                         34
  ANC reporting system: HMIS                                1                   2
  ANC reporting system: Programme specific                  2                   1
  ANC reporting system: If other, specify                   Inp:EIDSR
                                                            Out:DHMT western   Area

  Documents used to record ANC: ANC tally sheets            1                   2
  Documents used to record ANC: Patient cards               2                   1
  Source document ANC1 available - November                 2                   4
  Number of ANC1 Visits, Source Documents - November            20
  Source document ANC1 available - December                 2                   4
  Number of ANC1 Visits, Source Documents - December            12
  Source document ANC1 available - January                  2                   3
  Number of ANC1 Visits, Source Documents - January             15
  Monthly report ANC1 available - November                  2                   3
  Number of ANC1 Visits, Monthly report - November              20
  Monthly report ANC1 available - December                  2                   1
  Number of ANC1 Visits, Monthly report - December              12                   0
  Monthly report ANC1 available - January                   2                   1
  ANC: Reasons for Discrepancy                               1                   5
  Immunization reporting system: HMIS                       1                   2
  Immunization reporting system: Immunization/EPI prog      1                   2
  Immunization reporting system: If other, specify          Inp:EIDSR
                                                            Out:DHMT Western Area

  Source document Penta3 available - November               2                   1
```

SESSION 13

Data Compilation, Cleaning, and Quality Assurance

Compare data files

SESSION 13

Data Compilation, Cleaning, and Quality Assurance

- The input file and reference file are listed at the top. Each case in either file appears listed on the left, identified by the facility code.
- For each case, any difference between the input file and the reference file will be listed, with values for the input file under the column "Input File" and for the reference file under the column "Reference File" (far right).
- If the case exists in one file but not in the other, CSDiff will output "Case missing" in the relevant column.
- In the screenshot on the previous slide, the input and the reference files contain data for the facility with facility code 01141234. The results will show only differences between the two cases with the same ID.

Compare data files

- If the differences are only in spelling of facility name, facility location, other or text only fields, no edits need to be made.
- If differences arise in other questions, make a list by facility of questions that have a mismatch. Send this list to the supervisor for resolution of discrepancies. When the discrepancies have been resolved, return to the original data set and edit the record to reflect the changes.

SESSION 13

Data Compilation,
Cleaning, and
Quality Assurance

Questions

- What should you do if you find that teams are not filling a certain survey question appropriately?
- Describe the supervisor's role in quality assurance of the DV/SA.
- Describe methods for checking data completeness during survey implementation
- Describe methods for assessing quality of survey implementation during implementation.

SESSION 13

Data Compilation,
Cleaning, and
Quality Assurance

Compare data files

SESSION 13

Data Compilation,
Cleaning, and
Quality Assurance

Practice

- Your instructor will provide an example data file or files for use in practicing the techniques for data compilation, cleaning and quality assurance.
- Practice concatenating several data files.
- Practice indexing and exporting the data to Excel. Review the data in both the CSPro data entry module and in Excel.
- Practice splitting a data set into interviewer and supervisor datasets. Practice using the Compare function to compare values between interviewers and by supervisors.
- You have 60 minutes.