Merging Data with the Epi Info™ 7 Data Packager

If you have distributed your project for data entry and would like to combine the data from all of your colleagues and merge it into a centralized "master" data set so it can be analyzed, you could use the Epi Info 7 Data Packager to complete this process.

Figure 1

Data Packager Overview

For those familiar with prior versions of Epi Info™, the traditional data merging capabilities are still present in the Classic Analysis module. However, a simpler way of merging Epi Info™ 7 data can be accessed through the Enter module. This feature is known as the Data Packager.

The Data Packager provides a simple interface for packaging the data you've collected into a single compressed and encrypted file. The packager can also optionally de-identify data and filter certain records based on conditions you specify (e.g. specifying that only records entered in the last 7 days should be packaged). Once you've sent the package to another user, they would use the Enter module's **Import from data package** tool to merge the package's data into their form.

For this to be successful, both sender and recipient *must* have a copy of Epi Info™ 7 installed and *must* have identical (or near-identical) forms. If the form used to create the package differs significantly from the form that will receive the package's data, the process will not be able to proceed.

Matched and Unmatched Records

The data packager uses each record's unique ID value to determine if two records match. (Note that the unique ID is assigned to a record when the record is saved and is not normally displayed during the data entry process.) If the import process detects that a record in the data package already exists in the destination form, the record in the destination form is *updated* with the data present in the data packager. If the import process does not detect that a given record already exists, then the record from the data packager is *appended* to the form.

Imagine a scenario where you're receiving data packages from the same individuals on a weekly basis. Some new records are included in each week's package, but most of the records are the same ones you've been importing each week. Would you want those existing records to be appended each week? Probably not; if they were appended, you would end up with a large amount of duplicate records. A lot of data cleaning might be required to carry out useful analysis. You would likely prefer that those existing records in your data set be *updated* with any new data. For example, a case entered six weeks ago may have died or been re-interviewed, and the record in the investigator's computer was updated. When you import the data package, it will update the corresponding record in your data set with the new information.

To illustrate what would happen for both matched and unmatched records during a data packager import, assume the following scenario exists: Data from Table 1 resides in your data entry form. You have just been given a data package consisting of data from Table 2. You want to merge the data from the data package into your form.

Notice that Case ID #3 exists in both tables. It has a record ID value of 7e88e70c-faa9-4b87-8a96-df403df92fde. The age of the individual is different, however; the data package contains an age of 33 for case #3 while your data set contains an age of 32. After the import process has completed, your database will look like that shown in Table 3.

Table 1

Record Globally Unique ID	Case ID	Name	Age
01b91724-564e-4473-bfab-6d74a1377e4a	1	John	12
457452f2-ce73-45b7-a50f-024c80184cad	2	Mary	43
7e88e70c-faa9-4b87-8a96-df403df92fde	3	Susan	32

Table 2

Record Globally Unique ID	Case ID	Name	Age
7e88e70c-faa9-4b87-8a96-df403df92fde	3	Susan	33
a5a0e36d-a4df-41cb-a7c1-0ba44a895efe	4	Peter	31
dfc3547c-797a-4516-8f41-f08e6f6e4e0f	5	Andy	76

Table 3

Record Globally Unique ID	Case ID	Name	Age
01b91724-564e-4473-bfab-6d74a1377e4a	1	John	12
457452f2-ce73-45b7-a50f-024c80184cad	2	Mary	43
7e88e70c-faa9-4b87-8a96-df403df92fde	3	Susan	33
a5a0e36d-a4df-41cb-a7c1-0ba44a895efe	4	Peter	31
dfc3547c-797a-4516-8f41-f08e6f6e4e0f	5	Andy	76

The importer determined that of the three records in Table 2, two of those records did not exist in the destination table and therefore appended those records (cases #4 and #5). However, it determined that case #3 already existed in the destination table and so instead of appending it, it updated the destination table's information for case #3. The update is why the resulting data contains an age value of 33 for case #3 instead of 32.

Selecting the Import Type

When importing records from an Epi Info™ data package, you can specify if you want to update and append, update only, or append only. Selecting "Update and append records" will cause the importer to behave as described in the Matched and Unmatched Records section and is the default option.

Selecting "Update records only" will update matching records but skip appending any unmatched records.

Selecting "Append records only" will add any unmatched records and skip updating any records that match.

Encryption and Security

Three security mechanisms are built in to the data packager:

- File encryption
- Column data removal
- Row filtering

File Encryption

The data packager compresses and encrypts the package file that is created. The encryption uses a standard 256-bit symmetric key encryption algorithm provided by the Microsoft .NET Framework v3.5. Specifying a password during the data packaging process is recommended but not required.

Because the encryption keys are stored in the software itself, a determined and/or experienced attacker would be able to determine what the keys are and potentially crack the file very quickly. The encryption on the file is therefore not intended to provide world-class security, but rather, to prevent unskilled eavesdroppers from viewing the file's contents. (Think of it like the deadbolt on your front door; it will keep out the nosy neighbors, but a skilled thief will be able to bypass it without much effort.) For enhanced security, consider using additional encryption software.

Column Data Removal

Removing data from selected columns gives you a way to exclude personally identifiable information (PII) or other confidential information from the package.

Row Filtering

Row filtering can be used to package only certain records. For example, you can package only records where the case status is "Confirmed" and that were interviewed in the last three days.

Creating a Data Package

1. Open the project and form using the Enter Data module.
2. Select **File** > **Package for Transport…** from the Enter Data menu. The Package for transport dialog opens.
3. Click the **…** (ellipsis) button next to the **Package path** box. A Windows Browse for Folder dialog appears.
4. Select the *Desktop* from the list of available locations to be the path. (NOTE: You can specify any location within your computer except the same location as the project file).
5. Click **OK**. The Package path box in the Package for transport dialog is populated with the folder path to the desktop.
6. Type "DemoPackage" into the **Package name** box. (You can call it whatever you want).
7. Check the **Append timestamp to file name** box.
8. Type "1234" into the **Password** box. (user determines whatever password to use..)
9. Type "1234" into the **Verify password** box.
10. Click the **Save Script** button. A Windows Save dialog opens.
11. Type "PackageScript" into the **FileName** box.
12. Click **Save**. An informational box appears stating that the script has been saved.
13. Click the **Package** button once. The data packaging process starts.
14. When the packaging process has completed, verify that the last item in the status box contains the words "Package creation complete."
15. Click **Close** to exit the data packager.

You have successfully created an Epi Info™ 7 data package file on the desktop. Since the script was saved, users can quickly re-package the data later using the same settings without needing to go through all of those steps again.

To finish the process, you would attach the DemoPackage.edp7 file (now residing on the desktop) to an e-mail message.

Importing a Data Package

You have just been sent an Epi Info™ 7 data package containing the records in a project. To start analysis, you need to merge these records into your master project.

1. Open the project and form using the Enter Data module.
2. Select **File** > **Import Data** > **From Data Package** from the Enter Data menu. The *Import From Encrypted Data Package* dialog appears.
3. Click the **…** (ellipsis) button next to the **Encrypted data packages to import** box. A Windows Open File dialog appears.
4. Navigate to the location where you saved the data package(s) that were sent to you.
5. Select the package(s) that was sent to you by your colleague. (NOTE: If you are importing more than one data package file at once, click on the IS BATCH IMPORT checkbox. This will import all the files within the specified directory at once.)
6. Click **Open**. The file path appears in the **Encrypted data packages to import** box.

7. Type “1234” for the password. The dialog should appear as shown in Figure 1.
8. Click **Import**. The import process begins.
9. When the import process has completed, click the **Close** button. The Enter Data window refreshes and all records imported as well as the ones contained in the database appear in the record counter.

Figure 1

Resource: The following links provide additional information about packaging and importing data for sharing using the Data Packager:
http://wwwn.cdc.gov/epiinfo/script/ViewTutorial.aspx?id=10

http://wwwn.cdc.gov/epiinfo/script/ViewTutorial.aspx?id=11