Datafile structure - EpiData version 2.2

EpiData

EpiData uses four types of files:
1. Dataform definition file. E.g. first.qes
2. Actual datafile containing the data. E.g. first.rec.
3. A file with the defined checks. E.g. first.chk
4. Supplementary files, e.g. first.not with notes taken during dataentry or first.log with documentation.

This document contains a detailed description of the format for datafiles with the fileextension .REC. The datafile structure is the same as Epi Info version 6, except for changes introduced by new field types.

EpiData will read native Epi Info datafiles, but the field types phonenumber or local extension phonenumber will be treated as text fields and colour information will be ignored.

Epi Info will read native EpiData datafiles if they do not contain the field type soundex, EuroToday, reverse date (e.g. 1999/12/23), reverse today's date or encryption fields.

EpiData does not add a End-Of-File marker to the datafile, but will know how to handle the EOF marked added by Epi Info.

The datafile consists of a header describing the fields (variables) contained in the datafile plus the data.

The header consists of
- one line describing the number of fields (variables) in the datafile plus a code signifying the background colour of the entry form and one line for every field (variable).
- N lines each describing a field (variable) in the datafile

First line in header

First line in header contains information on number of fields in the datafile, background colour for data entry form, password for datafiles with encryption fields, a marker regarding how fields were named in EpiData and the datafile label.

Only the number of fields and the colour code is required.

The line begins with an integer value signifying the number of fields in the datafile. Please note that a field can be either a variable or a heading, so this number cannot be used to count the number of entry fields in the datafile. The number is of variable width, but it is separated from the rest of the the first line by a space (chr(32)).

Next in the first line is an integer value signifying the background colour of the dataform. This number is not used by EpiData, but is kept for compatibility with Epi Info. Width is variable, but it is separated from the rest of the the first line by a space (chr(32)).

If the datafile contains encryption field(s), the next item in the first header line will be the encryption password. The password is encrypted using the password it-self as key. It is then base64-coded to ensure readible characters. Is is saved in the header line prefixed by the marker "~kq:" and suffixed by the marker ":kq~".

EpiData uses two different systems for naming fields/variables. If the system "Use first word" is used the first line in the header will contain the marker VLAB in uppercase letters following the two integer numbers (and – if present – the encryption password). If VLAB is not present, then "Automatic fieldnaming" was

used when creating the datafile. VLAB marker is used internally by EpiData and is not important when converting to other datafile formats.

The datafile label is stored in the first line of the datafile following the identifier “Filelabel: “, i.e. the word ”Filelabel” plus a colon plus a space (#32). If ”Filelabel: ” is found in the first line then the rest of the first line contains the datafile label (max. 50 characters in length).

Field definitions

Following the first line the header contains one line per field.

Each line holds information on the entry field and of the text preceding the entry field (the ”question”).

Please note that a field with the length 0 (null) is a label (e.g. a heading) without an entry field and should be ignored if converting the EpiData datafile to other programs.

Data concerning positions are messured in character and lines. This position is converted to a positions messured in pixels based on the current font when the dataentry form is shown in EpiData.

The field description lines are formatted like this:

#	Name	Position in line	Width	Values	Description
1	Display character	1	1	# or _	The character to be displayed in the entry field. Not used in EpiData but kept for compatibility with Epi Info.
2	Fieldname	2-11	10	a-z, A-Z, 0-9, space	The field name. 1-10 characters (only 1-8 characters in EpiData versions 1.0 – 2.1b). Unused positions are filled with spaces. The name must begin with an alpha character. No spaces are allowed in the field name.
3	Unused	12	1	space (#32)	
4	Question column	13-16	4	0-9 An integer value	The column in which the text preceding the entry field begins. If the line is only text, it is displayed in this column. The first column on the left of the screen is column 1.
5	Question line	17-20	4	0-9 An integer value	The line in which the text preceding the entry field begins. Top line of the questionnaire is line 1.
6	Question color	21-24	4	0-9 An integer value	The colour code for the text preceding the entry field. Colour codes are shown in EpiData (from version 2.1b) in tools, colorcodes. Value not used by EpiData.
7	Field column	25-28	4	0-9 An integer value	The column where the entry field begins. The first column on the left of the screen is column 1. Label or heading fields will have a value of 0 (null).
8	Field line	29-32	4	0-9 An integer value	The line in the questionnaire where the entry field is shown. Top line of the questionnaire is line 1. Label or heading fields will have a value of 0 (null).
9	Field type	33-36	4	0-9 An integer value	A integer value signifying the field type. See tabel below.

10	Field width	37-40	4	0-9 An integer value	Width of entry field in characters. Label/headings fields has a width of 0 (null), which identifies the field as a label or heading without an entry field. See note below regarding encryption fields.
11	Entry field colour	41-44	4	0-9 An integer value	Colour code for the entry field. Colour codes are shown in EpiData (from version 2.1b) in tools, colorcodes. Value not used by EpiData. See note below regarding encryption fields.
12	Unused	45	1	Space (#32)	
13	Question	46 - to end of line	Variable	Any displayable character	The text preceding the entry field (i.e. the "question"). May contain prefixed spaces used to position the question correctly.

Note regarding encryption fields:
EpiData ver. 2.2 introduces encryption fields, i.e. fields where data are encrypted in the datafile. To ensure that the datafile is readable by all programs the data are first encrypted and afterwards Base64-coded. Base64 coding makes the encrypted data longer, e.g. 10 bytes of data will be 16 bytes in length when base64-coded. In EpiData the field width (item 10 in the table above) signifies the width of the data in the datafile, i.e. the base64 coded data. The number of characters allowed in the encryption entry field are saved in the entry field colour (item 11 above). An entry width of 1-14 characters are saved as 111+length (121 signifies that 10 characters can be entered in the field). A width of 15-80 characters are saved as the number 15-80.

Field type codes (see item 9 i table above)

Value	Field type	Field length	Comments
0	Integer	1-14 chars.	Integer number fields. Contains only numbers 0-9 or spaces. The field can be up to 14 characters in length, but EpiData saves integer fields with a length of 5 or more to Double Real Fieldtype with the field type code 100 (see Double Real field type)
1	Alpha	1-80 chars.	Text fields. Can contain all ANSI characters.
2	Date	5, 8 or 10 chars.	US style data fields, i.e. dates in the form mm/dd, mm/dd/yy or mm/dd/yyyy. What form is used is read from the length of the field. Datafiles created with EpiData will always create fields with 4-digit years, i.e. a length of 10, but to ensure compatibility with Epi Info short datetypes are allowed.
3	UpperAlpha	1-80 chars.	Upper-case text fields. Can contain all uppercase ANSI characters.
4	-	-	Not used in EpiData. The number is reserved to ensure compatibility with Epi Info. However, as far as I know the number is unused in Epi Info, too. The source code of Epi Info labels field type code 4 as "CheckBox".
5	Boolean	1 character	Boolean field or yes/no field. Can contain space (ANSI #32), the letter "Y" or the letter "N".
6	Double Real	1-14 chars.	Double Real number field. If the field type code is 6 or 100 then the number of decimals is null. A double real number field with one or more digits after the decimal separator will have the code 100+the number of decimals. Decimal separator will always be a dot (".") even if EpiData allows users to enter real numbers using a comma a decimal separator. An example: A field entered in EpiData as ###.## will signify

			a double real number field with the length of 6 and the field type code will be 100+2=102. A field entered as ###### will have the code 6, signifying 0 decimals.
7	-	-	Not supported by EpiData, which converts the data to text. Used for phone-number fields in Epi Info.
8	-	-	Not supported by EpiData, which converts the data to text. Was ment to be used for time fields in Epi Info.
9	-	-	Not supported by EpiData, which converts the data to text. Used for local extension phonenumber in Epi Info.
10	Today	5,8 or 10	Today's day field in US date style. Data cannot be entered in this field by the user but will contain the current date of the time the record containing the field was saved. Use same format as Date fields (see field code 2)
11	EuroDate	5,8 or 10	European style date field, i.e. dd/mm, dd/mm/yy or dd/mm/yyyy. Which date type is used is read from the length of the field. EpiData only creates date fields with 4-digit years (i.e. a field length of 10), but shorted date types are kept for compatibility with Epi Info.
12	IDNUM	5-14 chars.	Automatic ID-number field. Field will be filled out by EpiData with an incrementing integer number. Contains only numbers 0-9.
13	-	-	Not supported by EpiData. Field type code is reserved by Epi Info.
14	-	-	Not supported by EpiData. Field type code is reserved by Epi Info.
15	-	-	Not supported by EpiData. Field type code is reserved by EpiData.
16	EuroToday	5,8 or 10	European style today's date field. Data cannot be entered in this field by the user but will contain the current date of the time the record containing the field was saved. Use same format as EuroDate fields (see field code 11). Field type is not supported by Epi Info.
17	Soundex	5-80 characters	Soundex code field in the format A-000 where A can be any uppercase letter and 000 can be any three numbers. When converting EpiData datafiles to other programs this field type can be converted to a text field with a length of 5. Field type is not supported by Epi Info.
18	Entryption field	3-80 characters.	Encrypted and base64-coded text. Uses Rijndael encryption. Password in saved in the first line in the header (see description above).
19	Reverse date (yyyy/mm/dd)	10 characters	Dates in the format YYYY/MM/DD. Not supported by Epi Info.
20	Reverse today's date	10 characters	Today's date in the format YYYY/MM/DD. Data cannot be entered in this field by the user but will contain the current date of the time the record containing the field was saved. Not supported by Epi Info.

A few more remarks:

- A missing value is saved as a number of spaces (ANSI #32) corresponding to the length of the field.
- Epi Info creates field names using only upper-case letters. When datafiles are created in EpiData the user can optionally create lower-case, upper-case or mixed-case field names. The lettercase of fieldnames created in EpiData should therefore be kept if possible when converting to other datafile-formats.

- The length of the header of an Epi Info v6 file should not exceed 500. However it is not consistent across the Epi Info v6 modules whether the header length can indicate a position up to 999. Since 4 places are left for indication of which line in the qes file that ”line” in the rec file refers to: EpiData allows up to 999 lines in the QES file.

Data

The rest of the datafile consist of the data saved in fixed format based on the width of each field.

Datalines are a maximum of 79 characters. Therefore a record can be one or several lines in the datafile dependent on the total of the widths of the fields.

The last line of the record can be less than 79 characters. The last line is terminated by ”!” (normal record), ”?” (deleted record) or ”^” (verified record – used only in Epi Info). If the record consist of more than one line then the first lines will be terminated by ”!”.

Missing values are saved as spaces. Numeric values are prefixed by spaces if not all positions are used. Other fields are suffixed by spaces if not all positions are used.

Epi Info terminates the datafile with an End-of-file marker (hex 1A). EpiData does not add this marker.

Examples of datafiles created by EpiData.

First example

The file consists of 22 fields (variables), none of which are labels-only, and 3 records. The 2nd record is marked "Deleted" (it is terminated by a question-mark).

The datafile uses variable-labels. The question for the first field is, according to the datafile, "Int1 The first integer". But because of the VLAB flag, the fieldname "Int1" is removed and the variable-label is therefore "The first integer". The datafile has the label “This is the filelabel”.

```
22 1 VLAB Filelabel: This is the filelabel
#INT1.........1...1..30...7...1...0...1.112.Int1.The first integer
#INT4.........1...2..30...7...2...0...4.112.Int4.A 4-digit integer
#INT6.........1...3..30...7...3.100...6.112.Int6.A 6-digit integer
#INT9.........1...4..30...7...4.100...9.112.Int9.A 9-digit integer
#INT13........1...5..30...7...5.100..13.112.Int13.A 13-digit integer
#FIX11........1...7..30..10...7.101...3.112.Fix1:1...
#FIX53........1...8..30..10...8.103...6.112.Fix5:3...
#FIX104.......1...9..30..10...9.104..10.112.Fix10:4..
_TXT1.........1..11..30...7..11...1...1.112.Txt1..
_TXT4.........1..12..30...7..12...1...4.112.Txt4..
_TXT9.........1..13..30...7..13...1...9.112.Txt9..
_TXT14........1..14..30...7..14...1..14.112.Txt14.
_UPP1.........1..16..30...7..16...3...1.112.Upp1..
_UPP4.........1..17..30...7..17...3...4.112.Upp4..
_UPP9.........1..18..30...7..18...3...9.112.Upp9..
_UPP14........1..19..30...7..19...3..14.112.Upp14.
_BOOL.........1..21..30...7..21...5...1.112.Bool..
_SOUNDEX2.....1..22..30..11..22..17..20.112.Soundex20.
_EUDATE.......1..24..30..10..24..11..10.112.EU.date..
_USDATE.......1..25..30..10..25...2..10.112.US.date..
_EUTODAY......1..26..30..10..26..16..10.112.EU.today.
_USTODAY......1..27..30..10..27..10..10.112.US.today.
11111111111111111111111111111111.111.11111111.1111aaaaaaaaaaaaaaaaaaaaaaaaa!
aaAAAAAAAAAAAAAAAAAAAAAAAAAAAAYF-623              01/05/200005/01/200025/05/2!
00005/25/2000!
22222222222222222222222222222222.222.22222222.2222bbbbbbbbbbbbbbbbbbbbbbbbb!
bbBBBBBBBBBBBBBBBBBBBBBBBBBBBYA-535               02/05/200005/02/200025/05/2!
00005/25/2000?
```

Second example

The file consists of 16 fields and 2 records. First variable (LABEL1) is a heading without entry field. The 2nd record is marked "Deleted" (it is terminated by a question-mark).

The datafile includes an encryption field and the password is shown in the header's first line inclosed in "~kq:" and ":kp~". The password is "pwtest" (can be tested by uncode using base64 and then decrypt using Rijndael with the key "pwtest").
The datafile uses variable-labels (VLAB marker is present). The datafile has the label "Example of EpiData datafile".

```
16 1 VLAB ~kq:9IrX0B+q:kq~ Filelabel: Example of EpiData datafile
_LABEL1        1   1  30   0   0   0   0 112 Heading: This is example of an EpiData datafile
#INTEGER3      1   2  30  18   2   0   3 112 INTEGER3
_ALFA10        1   3  30  18   3   1  10 112 ALFA10
_USDATE        1   4  30  18   4   2  10 112 USDATE
_UPPERALFA     1   5  30  18   5   3  10 112 UPPERALFA
_BOOL          1   6  30  18   6   5   1 112 BOOL
#FLOAT22       1   7  30  18   7 102   5 112 FLOAT22
#FLOAT6        1   8  30  18   8   6   6 112 FLOAT6
_USTODAY       1   9  30  18   9  10  10 112 USTODAY
_EUDATE        1  10  30  18  10  11  10 112 EUDATE
#IDNUM         1  11  30  18  11  12   5 112 IDNUM
_EUTODAY       1  12  30  18  12  16  10 112 EUTODAY
_SOUNDEX       1  13  30  18  13  17  10 112 SOUNDEX
_CRYPT         1  14  30  18  14  18  16 121 CRYPT
_REVDATE       1  15  30  18  15  19  10 112 REVDATE
_REVTODAY      1  16  30  18  16  20  10 112 REVTODAY
111First text12/24/2003FIRST TEXTY11.1133333304/23/200324/12/20031    23/04/20!
03T-230     4sYbOSRmeEYMTU==2003/12/242003/04/23!
222second t  12/25/2003SECOND T  N44.4455555504/23/200325/12/20032    23/04/20!
03S-253     9xZws8JEcX1=    2003/12/252003/04/23?
```