

Data importation

Five examples of data importation for DtmVic

This tutorial contains a series of examples of data importation that aim at capturing or transforming data to comply with the DtmVic format files. Each example corresponds to a directory included in the sub-directory “DtmVic_Examples_D_Import” included in the directory “DtmVic_Examples” that has been downloaded with DtmVic.

Importation examples D.1—D.5

[To select the examples, press the right button of the mouse](#)

- Introduction Page D.2
- Example D.0. *Capture of dictionary and data.* Page D.5
- Example D.1. [EX_D01.Importation.XL.](#)
(Importation of numerical and textual data from an Excel ® file) Page D.7
- Example D.2. [EX_D02.Importation.Free](#)
(Importation of numerical data from a free format file) Page D.11
- Example D.3 [EX_D03.Importation.Fix.](#)
(Importation of numerical data from a fixed format file) Page D.15
- Example D.4. [EX_D04.Importation.Text.Free.](#)
(Importation of Textual Data from a free format file) Page D.18
- Example D.5. [EX_D05.Importation.Text.Num.XML.](#)
(Importation of both numerical and Textual Data from a XML format file). Page D.21

- Introduction -

Internal DtmVic format for input data and texts

The aim of the importation procedures is to transform a pre-existing text file into the “Internal DtmVic format”. The knowledge of the internal DtmVic format could be useful to some advanced users; it is not indispensable for the beginners.

Let us remind that DtmVic is a software devoted to exploratory analysis of multivariate numerical and textual data. The leading case that exemplifies all the possibilities of the software is a sample survey data set, comprising both responses to closed questions and responses to open-ended questions (the closed questions may lead to numerical [quantitative] or categorical [qualitative] data).

In the most general configuration, three files constitute the internal DtmVic input data set:

- 1) The **dictionary file**, that provides the names (or identifiers) of the numerical and categorical variables. It includes the names of the categories corresponding to each categorical variable. That latter feature is rather uncommon in statistical software, but seems indispensable to explore high dimensional categorical data sets.
- 2) The **data file**, that contains the values of these variables for a set of individuals (or: observations), together with the identifiers of the individuals.
- 3) The **text file** made of the responses to open ended questions. The text file (known as text file type 2) concerns the same respondents as the data file, in the same order. A simplified “text file format” (text file type 1) can be used when dealing only with a series of texts, without associated data file and dictionary file.

Some applications may involve only the text file (see for instance the example A4 of Tutorial A), whereas others may need only the dictionary and the data files (application examples A1, A2, A3, C1, C2, of Tutorials A and C).

Internal “DtmVic format”

These three types of files are in simple text format (extension “.txt”, readable through a “notepad” or a text editor, or also with a word processor, provided that they are saved as simple text files). As an introductory exercise, they can be recorded directly from the keyboard, or with the help of the menu “DataCapture”.

In most cases however, they have to be imported from (often large) pre-existing files. The transformation into DtmVic format is then transparent to the user.

Table 1 shows an example of a small DtmVic dictionary, involving four variables. Table 2 displays an example of a DtmVic data file (same four variables, six individuals or respondents). Table 3 presents a text file relating to three open-ended questions and three respondents.

Table 1: Example of an internal DtmVic dictionary for 4 variables.

Gender (2 categories); Age (0 categories = numerical variable); Age broken down into 4 categories; Educational level (3 categories). [fixed format, comments in *italic*].

2	GENDER	<i>(number of categories [2] in columns 1-4; blank; title of the variable)</i>
MALE	MALE	<i>(short identifier [column 1-4]; blank; identifier [< 20 characters])</i>
FEMA	FEMALE	<i>(short identifier [column 1-4]; blank; identifier [< 20 characters])</i>
0	AGE	<i>(number of categories [0] in columns 1-4; blank; numerical variable)</i>
4	AGE_CODE	<i>(number of categories [4] in columns 1-4; blank; title of the variable)</i>
AGE1	18_24	<i>(short identifier [column 1-4]; blank; identifier [< 20 characters])</i>
AGE2	25_39	<i>(short identifier [column 1-4]; blank; identifier [< 20 characters])</i>
AGE3	40_59	<i>(short identifier [column 1-4]; blank; identifier [< 20 characters])</i>
AGE4	>60	<i>(short identifier [column 1-4]; blank; identifier [< 20 characters])</i>
3	EDUCATION	<i>(number of categories [3] in columns 1-4; blank; title of the variable)</i>
EDUL	LOW	<i>(short identifier [column 1-4]; blank; identifier [< 20 characters])</i>
EDUM	MEDIUM	<i>(short identifier [column 1-4]; blank; identifier [< 20 characters])</i>
EDUH	HIGH	<i>(short identifier [column 1-4]; blank; identifier [< 20 characters])</i>

Table 2: Example of an internal DtmVic data file for the previous 4 variables:

Gender, Age broken down into 4 categories, Educational level. 3 respondents (individuals, observations)

'1006'	1	76	12	1	<i>(Identifiers of the respondents : between quotes, without blank, less than 20 characters. Separators between values: at least one blank space)</i>
'1007'	2	20	2	2	
'1008'	2	29	3	2	

Table 3: Example of an internal DtmVic text file (type 1) for three texts (see: application example EX_A04.Text.Poems of Tutorial A for texts in English).

Free text format on less than 80 columns. Separator of texts : "***" followed, after four blank spaces, by the identifier (<= 20 characters); End of file: "====". All separators are in columns 1, 2, 3, 4.**

****	LAMARTINE
Voilà les feuilles sans sève,	
Qui tombent sur le gazon	
Voilà le vent qui s'élève,	
Et gémit dans le vallon	
Voilà l'errante hirondelle,	
Qui rase du bout de l'aile,	
L'eau dormante des marais...	
****	GAUTIER
L'automne va finir, au milieu du ciel terne,	
Dans un cercle blafard et livide que cerne	
Un nuage plombe, le soleil dort. Du fond	
Des étangs remplis d'eau monte un brouillard qui fond	
Collines, champs, hameaux dans une même teinte.	
****	VERLAINE
Les sanglots longs	
Des violons	
De l'automne	
Blessent mon coeur	
D'une langueur	
Monotone.	
=====	

Table 4: Example of an internal DtmVic text file (type 2) for three responses to three open-ended questions (see: applications examples A5 from Tutorial A and B1, B2, B3 of Tutorial B) and for three respondents.

Free text format on 80 columns. Separator of respondents: “----“ followed by the identifier (<= 20 characters); Separator of question: “++++”; End of file: “====”. All separators are in columns 1, 2, 3, 4. Note the blank lines for empty responses (last respondent, second and third questions).

```
---- 1006
 my sons, my kids are very important to me,
 being on my own I am responsible for their education
 and moral standard
++++
 education and moral standard of the youngsters, law and order
++++
 basically, British culture is traditional,
 people tend to keep themselves to themselves
---- 1007
 job, being a teacher I love my job, for the well being
 of the children
++++
 law and order, drug abuse, child abuse
++++
 accommodating, of course people from different races
 and culture have settled in here, (i.e., Irish, Jewish,
 Asians) and the British culture is working alright
---- 1008
 job, sometimes it is very hard to find a job
++++

++++

====
```

Preliminary Example D.0

Capture of numerical data and dictionary

Recording the dictionary presented above in the introduction and recording some data.

- 1) Click on the button **“Data Importation , Preprocessing, Data Capture, Exportation”**. (Basic Step from the main menu of DtmVic).
- 2) A new window appears.
- 3) Choose the item : **“Building the Dictionary (manually)”**.
- 4) In the new green windows, three yellow cases are ready to receive the information relating to the first variable.
 - “Variable number”**, : **“1”** (default value).
 - “Variable identifier”**, type: **Gender**;
 - “Variable type”**: type **“2”** [the type of a variable is the number of its categories] (special value: 0 for a numerical variable).
- 5) Since the number of categories is greater than 1, a second green window appears, inviting you to record the names of the categories: **male, female**.
- 6) A second variable is then proposed. It will be the age, the type of which is **“0”**, it is a numerical variable. No window appears since no categories are involved.
- 7) A third variable is proposed, you may record a categorical variable “age” in 4 categories... etc.
- 8) A report of the recorded data is printed in the lower window, while the right hand side window displays the dictionary in DtmVic internal format. It is that dictionary that will be recorded at the end of the capture process.
- 9) When all variables are recorded, one must click on the button **“SAVE DICTIONARY”**. We suggest to build a new directory (or: folder) in a workspace that is convenient to you, to open that directory, and to save the dictionary as “dic.txt”.
- 10) Then: **“RETURN”**.

We are back in the Data Capture window.

Choose now the item **“Creating the data file”**.

The window “Creating data source file” appears.

Click on **“LOAD DICTIONARY”**

- Ignore, at this stage, the button **“Update an existing data file”**.

The previous chosen directory appears. Select the dictionary “dic.txt”.

That dictionary is displayed in the upper right window.

Simultaneously, the yellow upper left window is ready to receive the data relating to the first individual :

- its identifier, (type for example **“Rita”**),
- then the value of the first variable for that individual: GENDER. A click on the right border of the caption window displays the two possible values. Let us choose **“female”**, to be consistent with the

identifier.

The second variable, AGE, is then proposed to the user. A numerical value must be inserted in the window, etc.

At the end of the record, the second individual or observation is proposed...

We suggest to record 3 or 4 individuals in this exercise, more if you wish...

Then press the button **“SAVE”** .

The same directory is again proposed. A name should be proposed for the data file. The extension “.txt” is recommended, to facilitate a quick access to the content of the file. Let us select for example the name “dat.txt”.

Press then the button: **“Create a first parameter file”**.

The window “Creating a starting parameter file” appears.

Click on **“Create a parameter file”**.

A DtmVic parameter file is displayed in the lower window.

That parameter file is automatically saved under the name: **“param_start.txt”**.

The parameter file does not include any statistical analysis command, except basic counts of categories, together with a computation of extreme and average values for the purely numerical variables.

It is only meant here as a check of the capture of the data.

Optional comments about the “first parameter file”

After an identification of the two input files, three “steps” of DtmVic are involved:

The step “AR DAT” that archives data and dictionary. The step “SELEC” that selects the variables for the subsequent processing. In this case, all the available variables are selected. The step “STATS” that computes the basic statistics mentioned above.

Click on **“Execute”**.

Read the results by clicking on the **“Basic numerical results”** item of the menu. These results are saved under the names: “imp.html” an “imp.txt” in the same directory.

End of example D.0

Example D.1: **EX_D01.Importation.XL**

Importation of numerical and textual data in “Excel ® format”.

(updated: June 18th, 2010)

Transforming a specific XL (csv) format file into DtmVic dictionary file, text file and data file.

This importation procedure can be applied to any text file (.txt) having the following features, for n individuals and p variables:

The first row ($p + 1$ elements) contains the name of the identifier and the p names of the variables (no blank space allowed within the name, less than 20 characters, preferably less than 10) separated with a semicolon (or a comma, or a tab). Blank spaces are allowed between names (free format).

The n remaining rows ($p + 1$ elements) contain the identifier of the individual (less than 20 characters) and the values of the p variables (for categorical variables, no blank space are allowed within the alphanumeric values; preferably less than 10 characters) separated with a semicolon (or a tab). Blank spaces are allowed between values (free format) and within textual variables.

Only one type of separator (semicolon or tab) can be used in a file.

1- Looking at the data, preliminary steps

The folder “**EX_D01.Importation.XL**” contains the file “**database_global.xls**”.

The file `database_global.xls` corresponds to a frequent situation: the first row of the table contains the variable identifiers, the first column comprises the observations identifiers.

To begin with, we will have a look (**outside DtmVic**) at the original file to be imported.

This file is under Microsoft Excel ® format. The reader who is not provided with that software should skip the next instructions... or use the free software “Open Office” instead.

1.1) Search for the examples directory **DtmVic_Examples**

1.2) In that directory, open the directory of example D.01, named **EX_D01.Importation.XL**

1.3) Click on the file: “**database_classical.xls**” (basic dictionary **and** data **and** texts) to obtain a view of the data through an Excel spreadsheet.

- The first row contains the names of the 17 variables (there are 18 columns, but the first one relates to the identifier of individuals).

Note two important constraints:

- a) the names of variables must have less than 20 characters,
- b) these names should not contain blank spaces (replace them by underscores, if any).

Note that these names will be truncated down to 10 characters to build the identifiers of the categories. It is then important that these first 10 characters allow for identifying the variable.

The remaining rows consist of 1043 lines (it is the same sample of individuals from a socio-economic sample surveys serving as example in the applications A.5, B.2, B.3).

The sequence of characters in the first cell of each line is the identifier of individual, the following sequences being the values of the 17 variables. Blank cells means “no-answer” or “missing value”.

1.4) We must save this file as a text file in “.csv” format. (command: File, then “Save as”)We obtain a free format file with semicolons as separators. The file in “csv” format is provided in the example directory.

Important:

If there are some semicolons in the data file, they should be replaced by another symbol before saving the “Excel file” as a “csv file”.

Note also that before saving the file, the format of the cells must be “standard”, to avoid some additional small blank spaces in numbers of more than 3 digits that are misinterpreted by the csv file. In the French version and in some European versions of Excel, the “decimal commas” should be replaced by the usual decimal dots.

If your version of Excel does not allow for “saving as a csv file”, you can save the file using “tabs” as separators, and then, change the “tabs” into “semicolons” (basic step: “Data capture, data importation”, then: button “specific preprocessing”, then: button “replacing tabs with semicolons”). This suppose that the initial data set does not already contain semicolons : in such a case, you should replace these semicolons with another symbol before the importation process).

In many versions of Excel, the csv format uses commas as separators, instead of semicolons. You can then transform these commas into semicolons (provided that the initial data set does not already contain semicolons: you should replace then these semicolons with another symbol before the importation process).

2) Sequence of operations

2.1) Click on the button: “Data Importation , Preprocessing, Data Capture, Exportation”, (Basic Steps from the main menu of DtmVic). A new window appears.

2.2) Choose the item: “Importing Dictionary ,Data and Texts”. The new window “Data Importation” is displayed.

2.3) Press the button entitled: “Excel® type files (saved as csv files)”.

A new window entitled “Data Importation from an Excel (r) file” appears.

If the Excel file has been saved using “tabs” or “commas” as separators, click on one of the optional buttons:

“0. Change tabs into semi-colons”.

“0. Change commas into semi-colons”.

Select the file saved with tabs or commas, and convert it. Note that a new name is given to the created file. The importation process will continue using this new file.

2.4) Then, click on the button: **“Start the Importation Process”**

In the new window, click on: **“1. Select input data file.”** (widen the window if necessary).

Select the previously saved file: **“datbase_global.csv”** (or the file produced by one of the previous buttons “0”)

The left hand side memo contains, for each variable, all its observed values. In the case of continuous numerical variables, the number of values could be the same as the number of observations. In the case of textual data, the number of values is the number of “words” (separators : blank, periods, commas)

- The central memo is a summary of the previous one. For each variable, we can read within the brackets the number of distinct values observed in the file.
- The letter (A) in parenthesis means that some letters or non-numerical values have been observed.
- The letter (N) indicates that only numerical values have been obtained.

It is then easier to choose the status of the variables: categorical (**CHAR**), numerical (**NUM**) , textual (**TEXT**), variables to be abandoned (**DISCARD**).

You have then to select one or several consecutive items in the list, and choose, for each item (i.e.: each variable), one keyword among the four following keywords {char, text, num}.

- **“CHAR”** means that we are dealing with a category of a nominal variable. Such variable could be coded with at most 6 characters. For instance, ‘male’ and ‘female’ for coding the gender (or “0” and “1”, or “10” and “20” ...). Conventionally, the first item (identifier) should be a **“CHAR”**.
- **“NUM”** means that we are dealing with a numerical value.
- **“TEXT”** means that the records (up to 8000 characters, another constraint!) will feed the textual data file.
- **“DISCARD”** means that the records (whatever the prior status) will be suppressed in the imported file.

Clearly, a variable with a few distinct values containing letters (A) should be a categorical variable “char”.

Similarly, a variable with hundreds of purely numerical values (N) will probably deserve the type: “num”.

If expected numerical values contain letters (A), it could be than in the original Excel file, the missing values or ”Do not apply (DNA)” are represented by alphanumeric symbols. These symbols should be replaced with blank spaces in the original file, or directly in the “csv file” before the importation. If you give the status “num” to a variable whose values contain letters, the importation process will be stopped before being completed, entailing a waste of time.

2.5) Once the attribution of types is completed, click on the button **“3. Updating and continue”**.

2.6) In the new window, Click on **“Values and counts”**.

A further check of the consistency of the selected types or the variables. A list of all the categories found in the data file, with the corresponding frequencies is displayed. Basic parameters are also provided for numerical variables. We will not dwell on this output serving mainly as a technical check.

2.7) Click then on **“Create dictionary and data”**.

A new window entitled “Creating a dictionary and a data file“ appears on the screen.

2.8) Click on **“Name for the new dictionary”**.

You have to choose a name for the forthcoming DtmVic dictionary, always in the same directory (the extension “.txt” is recommended) select for example: “dtm_dic.txt.

2.9) Click on “Name for the new data file”

You have to choose a name for the forthcoming DtmVic data file, always in the same directory (the extension “.txt” is recommended). Select for example: “dtm_dat.txt”

2.10) [if textual data have been selected] Click on “Name for the new text file”

You have to choose a name for the forthcoming DtmVic text file, always in the same directory (the extension “.txt” is recommended). Select for example: “dtm_text.txt”

2.11) Click on “Create new dictionary”

A DtmVic dictionary is created (number of lines = total number of variables + number of found categories). The DtmVic dictionary is displayed in the right hand side memo.

2.12) Click on “Create new data file”.

2.13) [if textual data have been selected] Click on “Create new text file”.

A message box producing the numbers of different types of variables is displayed.

2.14) Click on “Create a first parameter file”.

The window “Creating a starting parameter file” appears.

[Reminder: In DtmVic, the phrases “Parameter file” and “Command file” are equivalent].

A DtmVic parameter file (or: command file) is displayed in the lower window.

The command file is automatically saved under the name: “**param_start.txt**”.

The command file does not include any statistical analysis command, except basic counts of categories, together with a computation of extreme and average values for the purely numerical variables. It is only meant here as a check of the importation of the data.

Optional comments about the “first command file”

After an identification of the two input files, three “steps” of DtmVic are involved:

The step “**ARDAT**” that archives data and dictionary. The step “**SELEC**” that selects the variables for the subsequent processing. In this case, all the available variables are selected. The step “**STATS**” that computes the basic statistics mentioned above.

2.15) Click on “Execute”. Back in the main menu window, the sequence of steps is displayed.

2.16) Click on the button: “Basic numerical results”.

The button opens a created (and saved) html file named “**imp.html**” which contains the main results of the previous basic computation steps. After perusing these numerical results, **return** to the main menu. Note that this file is also saved under another name: The name “**imp.html**” is concatenated with the date and time of the analysis (continental notation). That file keeps as an archive the main numerical results whereas the file “**imp.html**” is replaced for each new analysis performed in the same directory. Likewise, a simple text format file “**imp.txt**” is created and saved.

End of example D.1

Example D.2: **EX_D02.Importation.Free**

Data Importation: Numerical data in “free format”

Transforming a specific free format file into DtmVic dictionary and data files.

The format file to be imported can be easily derived, for example, from a SAS ® data file, after performing the “PROC CONTENT “ to save the data file under a simple text file.

1- Looking at the data, preliminary steps

To begin with, we will have a look at the original dictionary and data files to be imported.

We will use the editor of the button **“Open an existing Command file”** of DtmVic as a simple text editor.

- 1) **In the main basic window, click the button “Open an existing Command file”**
- 2) **Search for the examples directory DtmVic_Examples**
- 3) **In that directory, open the directory of example D.02, named EX_D02.Importation.free**
- 4) **Select the file: “dicbase.txt” (basic dictionary)**

It contains two columns (the format is free, any number of blank spaces may separate the elements of a same line). The first column contains the names of the 14 variables (there are 15 rows, but the row “0” relates to the identifier of individuals). Note two important constraints: a) the names of variables must have less than 20 characters, b) these names should not contain blank spaces (replace them by underscores, if any). The second column contains one keyword among the three following keywords {char, text, num}.

→ **“char”** means that we are dealing with a category of a nominal variable. Such variable could be coded with at most 6 characters. For instance, ‘male’ and ‘fema’ for coding the gender (or “0” and “1”, or “10” and “20” ...).

→ **“text”** means that the records (up to 30 characters) will not be taken into account in a numerical data file. However, to keep the order of the variables, a dummy variable taking the constant value 1 will be recorded instead.

(Warning: during the importation of an Excel file (EX_D01), the status “Text” had quite a different meaning: it meant “Textual data”, up to 8000 characters, leading to the creation of a dtm textual file)

→ **“num”** means that we are dealing with a numerical value.

- 5) **Return, and select now the data file: “database.txt” (basic data file).**

The file is made of 1043 lines (sample of individuals from a socio-economic sample surveys serving as example in the applications A.5, B.2).

The first sequence of characters is the identifier of individual, the following sequences, separated by a semicolon

“;” are the 14 variables, described, in the same order, by the file “dicbase.txt”. Blank space(s) within commas means either “no-answer” or “missing value”.

6) Nature of the importation process

The importation process consists in building a DtmVic dictionary and a DtmVic data file from the original data file.

→ The names of the variables are extracted from the first column of the “dicbase.txt” file.

→ The number of categories for each variable and the names of these categories are built from an analysis of the data file “database.txt”. For each variable, all the different sequences of characters observed in the data file are detected, and counted. The categories are sorted according to the alphabetical order of their identifiers.

→ The DtmVic data file starts with the same identifier inserted within quotes, the categories of the categorical variables will be consecutive integers starting with the value “1”, instead of an alphanumeric symbol. The numerical values will be identical to those of the original data file, except the missing values replaced, in this version of DtmVic, by the conventional value “999.”.

2) Sequence of operations:

2.1) Click on the button: “Data Importation , Preprocessing, Data Capture, Exportation”, (Basic Step from the main menu of DtmVic).

A new window appears.

2.2) Choose the item : “Importing Dictionary, Data and Texts”.

The new window “Data Importation” is displayed.

2.3) Press the button entitled: “Free Format Files”.

The window “ Free format. Finding the states of each categorical variables, frequencies“ is displayed.

2.4) Click on : “Select basic dictionary”

You have to select the file “dicbase.txt” in the previous directory: **EXD02.Importation.free.**

2.5) Click on: “Reading the basic dictionary”.

The content of the file “dicbase.txt” is displayed in the left hand side memo.

2.6) Click on: “Select basic data file”

You have to select the file “database.txt” in the same directory: **EXD02.Importation.free.**

2.7) Click on “Values and counts”.

All the nominal variables (symbol : “char” in the basic dictionary) are built from the two basic files.

A new button “Show results (counts)” is displayed. Clicking on that button displays the list of all the categories found in the data file, with the corresponding frequencies. We will not dwell here on this output serving mainly as a technical check.

2.8) Click then on “Create dict. and data”.

A new window entitled “Creating a dictionary and a data file “ appears on the screen.

2.9) Click on “Name for the new dictionary”. You have to choose a name for the forthcoming DtmVic dictionary, always in the same directory. (The extension “.txt” is recommended).

2.10) Click on “Name for the data file”. You have to choose a name for the forthcoming DtmVic data file,

always in the same directory. (The extension “.txt” is still recommended).

2.11) Click on “Create new dictionary”. A DtmVic internal dictionary is created. The DtmVic dictionary is displayed in the right hand side memo.

2.12) Click on “Create new data file”.

After a while, a message box displays the number of individuals.

A new button is produced: **“Create a DtmVic parameter file”**. Please click on that button.

Note: In our terminology, the phrases “Parameter file” and “Command file” are equivalent.

The window “Creating a starting parameter file” appears.

2.13) Click on “Create a first parameter file”.

A DtmVic parameter file is displayed in the lower window. The command file is automatically saved under the name: “param_start.txt”. The command file does not include any statistical analysis command, except basic counts of categories, together with a computation of extreme and average values for the purely numerical variables.

It is only meant here as a check of the importation of the data.

Optional comments about the “first command file”

After an identification of the two input files, three “steps” of DtmVic are involved:

The step **“ARDAT”** that archives data and dictionary. The step **“SELEC”** that selects the variables for the subsequent processing. In this case, all the available variables are selected. The step **“STATS”** that computes the basic statistics mentioned above.

The right hand side memo indicates how to run that parameter file.

2.14) Click on “Execute”.

Back to the main window, a message-box displays the sequence of computation steps.

2.15) Click on: “Basic numerical results” button

The button opens a created (and saved) html file named **“imp.html”** which contains the main results of the previous basic computation steps. After perusing these numerical results, **return** to the main menu. Note that this file is also saved under another name: The name **“imp.html”** is concatenated with the date and time of the analysis (continental notation). That file keeps as an archive the main numerical results whereas the file **“imp.html”** is replaced for each new analysis performed in the same directory. Likewise, a simple text format file **“imp.txt”** is created and saved.

→ *N.B. You can do again the same exercise in the sub-directory named “other example”.*

The files to be imported are “dictibaz.txt” and databaz.txt”. The dictionary “dictibaz.txt” corresponds to a full-size questionnaire (446 variables) but the data file is restricted to 346 individuals (10 % of the real sample).

End of Example D.2

Example D.3: **EX_D03.Importation.Fix**

Importing numerical data in “fixed format”.

Transforming a specific fixed format file into DtmVic dictionary and data files.

The format file to be imported, very dense, is still used by some surveys institute, although it refers implicitly to an obsolete conception of the data format: The fixed format is probably inherited from the FORTRAN format, together with the tradition of saving spaces on punched cards!

1- Looking at the data, preliminary steps

To begin with, we will have a look at the original dictionary and data files to be imported.

We will use the editor of the button **“Open an existing Command file”** of DtmVic as a simple text editor.

- 1) In the main basic window, click the button **“Open an existing Command file”**
- 2) Search for the examples directory **DtmVic_Examples**
- 3) In that directory, open the directory of example D.03, named **EX_D03.Importation.Fix**
- 4) Select the basic data file: **“databrut_small.txt”** (basic data)

The file, extremely dense, is made of 335 lines (sample of individuals from a semiometric survey [for more details, see examples of application C.1 and C.2]). The first sequence of characters is the identifier of individual, the following sequences being the variables. It is clear that we need a careful description of the position and the length for each variable, including the identifier...

- 5) Select now the basic dictionary file **“dictibaz_small.txt”**

It contains **five columns** (the format is free, any number of blank spaces may separate the elements of a same line).

The **first column** contains the number of the variable (if this information is not readily available, replace each number by, e.g., 1, but do not suppress the column).

The **second column** contains the names of the 76 variables (there are 77 rows, but the first one relates to the identifier of individuals). Note again two important constraints: a) the names of variables must have less than 20 characters, b) these names should not contain blank spaces (replace them by underscores, if any).

The **third column** contains one keyword among the two following keywords {**char**, **num**}.

→ **“char”** means that we are dealing with a category of a nominal variable. Such variable could be coded with at most 4 characters. For instance, ‘male’ and ‘fema’ for coding the gender (or “0” and “1”, or “10” and “20” ...).

→ **“num”** means that we are dealing with a numerical value.

The **fourth column** contains the number of columns occupied by each variable.

The **fifth and last column** contains an indicator of the position of the variable. That information is somewhat

redundant with that of the fourth column in this case. As for the first column, each number can be replaced by the value 1, but the column must not be omitted.

6) Nature of the importation process (identical to that of Example D.1 and D.2)

The importation process consists in building a DtmVic dictionary and a DtmVic data file from the original data file.

→ The names of the variables are extracted from the second column of the “**dictibaz_small.txt**” file.

→ The number of categories for each variable and the names of these categories are built from an analysis of the data file “**databrut_small.txt**”. For each variable, all the different sequences of characters observed in the data file are detected, and counted. The categories are sorted according to the alphabetical order of their identifiers.

→ The created internal DtmVic data file starts with the same identifier put within quotes, the categories of the categorical variables will be consecutive integers starting with the value “1”, instead of an alphanumeric symbol. The numerical values will be identical to those of the original data file, except the missing values replaced, in this version of DtmVic, by the conventional value “999.”.

2 - Sequence of operations:

2.1) Click on the button “Data Importation , Preprocessing, Data Capture, Exportation”, (Basic Step from the main menu of DtmVic). A new window appears.

2.2) Choose the item : “Importing Dictionary, Data and Texts”.

The window “Data Importation” is displayed.

2.3) Press the button entitled: “Fixed Format Files”.

The window “ Fixed format. Finding the states of each categorical variables, frequencies... “ is displayed.

2.4) Click on : “Select basic dictionary”

You have to select the file “**dictibaz_small.txt** ” in the previous directory: **EX_D03.Importation.Fix.**

2.5) Click on “Reading the basic dictionary”.

The content of the file “**dictibase_small.txt**” is displayed in the left hand side memo.

2.6) Click on “Select basic data file”

You have to select the file “**database.txt**” in the same directory: **EXD03.Importation.Fix.**

2.7) Click on “Values and counts”.

All the nominal variables (symbol : “char” in the basic dictionary) are built from the two basic files.

[In fact, at this stage, the data file is converted into a free format file].

The sequel is identical to that of the two first examples.

A new button “**Show results (counts)**” is displayed. Clicking on that button displays the list of all the categories found in the data file, with the corresponding frequencies. We will not dwell here on this output serving mainly as a technical check.

2.8) Click then on “Create dict. and data”.

A new window entitled “Creating a dictionary and a data file “ appears on the screen.

2.9) Click on “Name for the new dictionary” You have to choose a name for the forthcoming DtmVic

dictionary, always in the same directory. (The extension “.txt” is recommended).

2.10) Click on “Name for the data file” . You have to choose a name for the forthcoming DtmVic data file, always in the same directory. (The extension “.txt” is still recommended).

2.11) Click on “Create new dictionary” . A DtmVic internal dictionary is created. The DtmVic dictionary is displayed in the right hand side memo.

2.12) Click on “Create new data file”.

After a while, a message box displays the number of individuals.

A new button is produced: **“Create a DtmVic parameter file”**. Please click on that button.

Note: In our terminology, the phrases “Parameter file” and “Command file” are equivalent.

The window “Creating a starting parameter file” appears.

2.13) Click on “Create a first parameter file”.

A DtmVic command file is displayed in the lower window. The remaining operations and comments are identical to those of the introduction.

The command file is automatically saved under the name: “param_start.txt”. The command file does not include any statistical analysis command, except basic counts of categories, together with a computation of extreme and average values for the purely numerical variables.

It is only meant here as a check of the importation of the data.

Optional comments about the “first command file”

After an identification of the two input files, three “steps” of DtmVic are involved:

The step **“ARDAT”** that archives data and dictionary. The step **“SELEC”** that selects the variables for the subsequent processing. In this case, all the available variables are selected. The step **“STATS”** that computes the basic statistics mentioned above.

2.14) Click on “Execute”.

Back to the main window, a message-box displays the sequence of computation steps.

2.15) Click on: “Basic numerical results” button

The button opens a created (and saved) html file named **“imp.html”** which contains the main results of the previous basic computation steps. After perusing these numerical results, **return** to the main menu. Note that this file is also saved under another name: The name **“imp.html”** is concatenated with the date and time of the analysis (continental notation). That file keeps as an archive the main numerical results whereas the file **“imp.html”** is replaced for each new analysis performed in the same directory. Likewise, a simple text format file **“imp.txt”** is created, renamed and saved.

End of Example D.3

Example D.4: **EX_D04.Importation.Text.Free**

Importation of textual data in “free format”.

Transforming a specific free format text file into DtmVic text files (type2) .

The DtmVic format for textual data is described in Table 4 of the introduction of this Tutorial. It contains two types of separators: separators of individuals: “----“ and separators of questions “++++”, located in columns [1,2,3,4]. There is one constraint for the length of a line (80 characters) but, in principle, no constraint about the number of lines for one question or for one individual. However the number of open questions should not exceed 12, the number of closed questions should not exceed 1000. The number of individuals is limited to 22 500 in the present version of DtmVic.

Remark about DtmVic text file type 1:

Another separator (separator of texts : “**”) could be used in the case of DtmVic text file type 1, exemplified by Table 3 of the introduction.**

This kind of internal format can be easily built directly from the original corpus of texts without using the importation procedure (see the example: EX_A04.Text-Poems of Tutorial A).

No importation procedure is needed in that case.

To begin with, we will have a look at the original textual data to be imported.

1- Looking at the data, preliminary steps

We can use the editor of the button “**Open**” (line: “Command file”) of DtmVic as a simple text editor.

- 1) **Click on the button “Open”**
- 2) **Search for the examples directory `DtmVic_Examples`**
- 3) In that directory, **open the directory of example D.04, named `EX_D04.Importation.Text.Free`**
- 4) **Select the basic text file: “`TDA1_text_free.txt`” .**
- 5) (the responses are those involved in application examples A.5, B.1, B.2, B.3).

The free format of that file is the following:

- Each line corresponds to an individual (a respondent).
- The separators are the character #, which serves to separate the identifier of a respondent from its first response, and also to separate two consecutive responses.

We deal here with three open ended questions, since we have three # per line (a character # at the end of a line means an empty response to the last open question).

6) Nature of the importation process

The importation process consists in building a DtmVic text file from the original text file.

It consists in inserting the different separators.

The DtmVic format is closer to the usual format of texts in everyday life, easier to consult and peruse. However, the matching of both textual and numerical files is more easily carried out with the basic textual format (one individual = one row).

2 - Sequence of operations:

2.1) Click on the button “Data Importation , Preprocessing, Data Capture, Exportation”, (Basic Step from the main menu of DtmVic). A new window appears.

2.2) Choose the item : “Importing Dictionary ,Data and Texts”.

The window “Data Importation” is displayed.

2.3) Press the button: “Textual data (free format)”.

The window “ Importation of a text file“ is displayed.

2.4) Click on : “Open text file”

You have to select the file “TDA1_text_free.txt” in the directory: **EXD04.Importation.Text.Free.**

2.5) Click on “Convert into DtmVic file”.

The 100 first line of the new DtmVic text file are displayed in the right hand side memo.

A prudent message is given “Conversion apparently completed”.

Two message boxes give successively the number of individuals (1043) and the maximum length of the identifiers (10).

→ The DtmVic text file is automatically saved under the name:”**DtmTextFile.txt**”

2.6) Click then on the button : “Create a first parameter file”.

The window “Creating a starting parameter file” appears.

[Reminder: In DtmVic, the phrases “Parameter file” and “Command file” are equivalent].

2.7) Click on “Create a first parameter file”.

A DtmVic parameter file (or: command file) is displayed in the lower window.

The command file is automatically saved under the name: “**param_tex_start.txt**”.

The command file does not include any statistical analysis command, except basic counts of words for the first open question (parameter NUMQ = 1 in the step SELOX).

It is only meant here as a check of the conversion of the data.

Optional comments about the “first parameter file”

After an identification of the two input files, three “steps” of DtmVic are involved:

The step “ARTEX” that archives the three sets of responses to the three open questions. The step “SELOX” that selects the open questions for the subsequent processing. In this case, by default, the first question is selected. The step “NUMER” that performs the numerical coding of the selected text.

The right hand side memo indicates how to run that parameter file.

2.8) “Return to the Main menu” of DtmVic.

In this main basic window, click the button “**Open**” (line: “Command file”)

Select the parameter file **“param_tex_start.txt”**. (as a general rule, the three files [parameter, data and dictionary] must be in the same directory).

-- **“Return to execute”**.

2.9) Click on “Execute”.

Back to the main window, a message-box displays the sequence of computation steps.

2.10) Click on: “Basic numerical results” button

The button opens a created (and saved) html file named **“imp.html”** which contains the main results of the previous basic computation steps. After perusing these numerical results, **return** to the main menu. Note that this file is also saved under another name: The name **“imp.html”** is concatenated with the date and time of the analysis (continental notation). That file keeps as an archive the main numerical results whereas the file **“imp.html”** is replaced for each new analysis performed in the same directory. Likewise, a simple text format file **“imp.txt”** is created, renamed and saved.

End of Example D.4

Example D.5: **EX_D05.Importation.Text.num.XML**

Importation of numerical and Textual data in “XML format”.

(updated: April 29th, 2009)

A specific XML format allows for dealing with both numerical data and textual data in a unique file. Such format could be generated from some online questionnaires in the framework of a mySQL database.

The internal DtmVic format for textual data is described in Table 3 of the introduction of this Tutorial. It contains two types of separators: separators of individuals: “----“ and separators of questions “++++”, in columns [1,2,3,4]. There is one constraint for the length of a line (80 characters) but, in principle, no constraint about the number of lines for one question or for one individual. Another separator (separator of texts : “****”) could be used in another context when dealing with a series of texts (without theoretical limitations of size) without numerical data file matching these texts (Example: series of novels, poetries, discourses, chapters, etc.).

To begin with, we will have a look at the original XML data file to be imported.

1- Looking at the data, preliminary steps

We will use the editor of the button **“Open”** (line: “Command file”) of the main menu of DtmVic as a simple text editor.

- 1) Click on the button **“Open”**
- 2) Search for the examples directory **DtmVic_Examples**
- 3) In it, open the directory of Example D.05, named **EX_D05.Importation.Text.num.XML**
- 4) Select the unique file: **“TDA2_dtm.xml”**.

(The data are the same as those of example A.5 of Tutorial A : instead of three files – dictionary, numerical data, textual data- we have now only one (rather large) file).

The structure is schematised below.

All the tags can be chosen by the user, except the tag `<individual>` that indicates an end of record. However; that keyword **“individual”** is only a default value. It can be changed during the importation process.

For an individual, a missing tag means “no-response”. It is advisable, but not necessary, to put the tags in the same order.

The first tag after `<individual>` must be the identifier of the individual (tag `<id>` in the example, but any other name is acceptable).

```

-----
<FileName.xml>

  <individual><id> identifier1 </id>
  <question1> response1</ question1>
  <question2> response2</ question2>
  .....
  <open>
  <Open_quest_1>   free response 1 </Open_quest_1>
  <Open_quest_2> free response 2 </Open_quest_2>
  .....
  </open>
</individual>

  <individual><id> identifier2 </id>
  <question1> response2.1</ question1>
  <question2> response2.2</ question2>
  .....
  <open>
  <Open_quest_1>   free response 2.1 </Open_quest_1>
  <Open_quest_2> free response 2.2 </Open_quest_2>
  .....
  </open>
</individual>

  .....
  .....
  <individual>
  .....etc.
</individual>

</FileName.xml>
-----

```

Comments complying with XML syntax are possible anywhere in the file.

Note that this simple format is directly provided by saving the MySQL data bases derived from “on line surveys” as a simple XML file (without attributes, the tags being nested as shown before).

The drawbacks are the following:

- An obvious drawback of the XML structure is the size of the file, owing to the presence of opening and closing tags for each variable and individual.
- Another problem is the presence of XML-dedicated symbols such as: &, <, >, ‘, “, . The dictionary must not contain such characters.

The advantages are the following:

- A unique file replaces three files (dictionary, data and text). (The same situation occurs in the case of an Excel file [see Example D_01] with a limitation of the size of the cells: less than 8000 characters).
- The order of variables can change from an individual to another.
- The order of individuals is no more important, since it is not necessary to match the data file and the text file.
- The length of individual records can vary, since the absence of a tag means “no response” to the corresponding variable.

5) Nature of the importation process

The importation process consists in building the three internal DtmVic files (dictionary file, data file and, possibly, text file) from the original XML file.

2 - Sequence of operations:

2.1) Choose the item : “ **Data Importation , Preprocessing, Data Capture, Exportation**”.

(Basic Step from the main menu of DtmVic). A new window appears.

2.2) Select the item : “**Importing Dictionary, Data and Texts**”.

The window “Data Importation” is displayed.

2.3) Press the button: “**XML specific file**”.

The window “Find and select the tags, Import XML data file“ is displayed.

2.4) If the tag separating the individuals in your XML file is not the keyword “**individual**”, type your own tag in the first small white window. Press “**enter**” to register the new tag.

2.5) As explained in the pop up purple window, two thresholds are necessary.

Threshold1 is the minimal number of respondent to an open question (default value: 40) (that default value means that if the sample size is 1000, we tolerate 960 non-responses for some open questions). The question is discarded if the number of response is less than Threshold1.

Threshold2 is the minimal length of the lengthiest response to an open question (default value : 60) (that default value means that if all the responses to a question have less than 60 **characters**, this question will not be selected as an open question, and discarded)

Remind that in this version of DtmVic, the number of open-questions should be less than 12, whereas the number of closed question should be less than 1000.

If you wish to change the previous default values Threshold1 and Threshold2, enter the new values in the two windows below (don’t forget to press the “**enter**” button afterwards).

2.6) Click on: “**List of tags and content**”

You have to select the file “**TDA2_dtm.xml**” in the directory: **EX_D05. Importation.Text.num.XML**. Some messages are produced, describing the different steps of the process.

2.7) If the XML file contains responses to open-ended questions:

Click on “**Create the textual data file to be imported**”.

2.8) Always in the case in which the XML file contains responses to open-ended questions:

Click on: “**import as an internal DTM file**”

The 100 first lines of the new DtmVic text file are displayed in the right hand side memo.

A message box gives the number of individuals.

→ The DtmVic text file is automatically saved under the name : “**Dtm_final_text_TDA2_dtm.xml.txt**”

→ The DtmVic data file and the DtmVic dictionary file **remain to be imported** from the created csv file: “**Dtm_import_num_TDA2_dtm.xml.txt**” (importation as an Excel file).

→ An intermediate file, “**Dtm_import_text_TDA2_dtm.xml.txt**” is also created, as a mere check. It is the text file in importation format. In fact, the importation of text has been completed and the final text has already been provided (“**Dtm_final_text_TDA2_dtm.xml.txt**”)

2.9) About the control files

Five control files are created to check the different steps of the process.

- The (huge) file: **“Check1_data_TDA2_dtm.xml.txt”** contains a list of all the tags encountered for all the individuals.
- The file: **“Check2_Tags_TDA2_dtm.xml.txt”** contains a list of all the encountered tags, with the parameters characterizing these tags (frequency, mean rank, average length of the content, minimum length maximum length). In this case, all the tags are present and have the same position.
- The file **“Check3_Dict_TDA2_dtm.xml.txt”** contains all the tags sorted according to their average rank (in this particular case the same rank for each individual), the tags corresponding to textual responses, the tags corresponding to numerical responses.
- The file **“Check4_Textual_TDA2_dtm.xml.txt”** contains all the encountered responses to open questions.
- The file **“Check5_final_text_TDA2_dtm.xml.txt”** complements the previous one.

These five files could be suppressed after checking the whole process.

As a conclusion:

The DtmVic text file is created: (**“Dtm_final_text_TDA2_dtm.xml.txt”**).

The dictionary file and the data file have been converted from the XML file to a standard csv file.

They remain to be imported through a standard Excel importation process, using the created file:

“Dtm_import_num_TDA2_dtm.xml.txt” as an input (see: example D.1).

End of Example D.5

End of tutorial D
