1 Introduction

The integration of Tanagra into a spreadsheet, such as Excel[1] or Open Office Calc[2] (OOCalc), is undoubtedly an advantage. Without special knowledge about the database format, the user can handle the dataset into a familiar environment, the spreadsheet, and send it to specialized tools for Data Mining when he want to lead more sophisticated analysis.

The add-on for OOCalc is initially created for Windows OS. Recently, I have described the installation and the utilization of Tanagra under Linux[3]. The next step is of course the integration of Tanagra into OOCalc under Linux.

Mr. Thierry Leiber has realized this work for the 1.4.31 version of Tanagra. He has extended the existing add-on. **We can launch Tanagra from OOCalc now, either under Windows and Linux**. The add-on was tested under the following configurations: Windows XP + OOCalc 3.0.0; Windows Vista + OOCalc 3.0.1; Ubuntu 8.10 + OOCalc 2.4; Ubuntu 8.1 + OOCalc 3.0.1.

This document extends a previous tutorial, but we work now under the Linux environment (**Ubuntu 8.10**). All the screen shots are in French because my OS is in French, but I think the process is the same for Linux with other language configuration.

2 Dataset

We use the CEREALS.XLS dataset (http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/cereals.xls -- from the StatLib server http://lib.stat.cmu.edu/datasets/1993.expo/). It describes the characteristics of 76 breakfast cereals. The data file is in the Excel format, OOCalc can handle it. We lead a PCA (Principal Component Analysis) in this tutorial in order to illustrate the data manipulation, but the main subject is the connection between OOCalc and Tanagra under Linux environment.

3 Installing the Add-On into Open Office Calc

We consider that Tanagra is correctly installed under Linux and works using Wine (see http://tutoriels-data-mining.blogspot.com/2009/01/tanagra-sous-linux.html)[4]. We launch OOCalc.

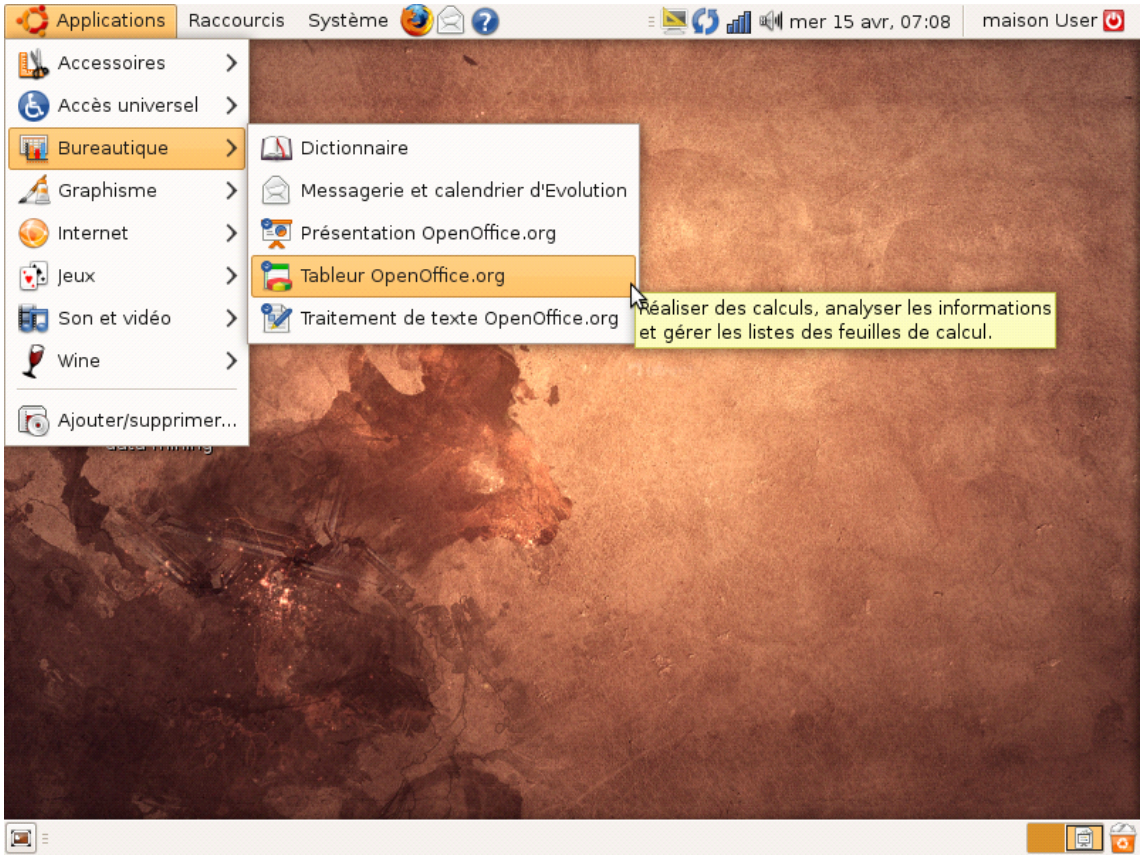
---

1 http://data-mining-tutorials.blogspot.com/2008/10/excel-file-handling-using-add-in.html

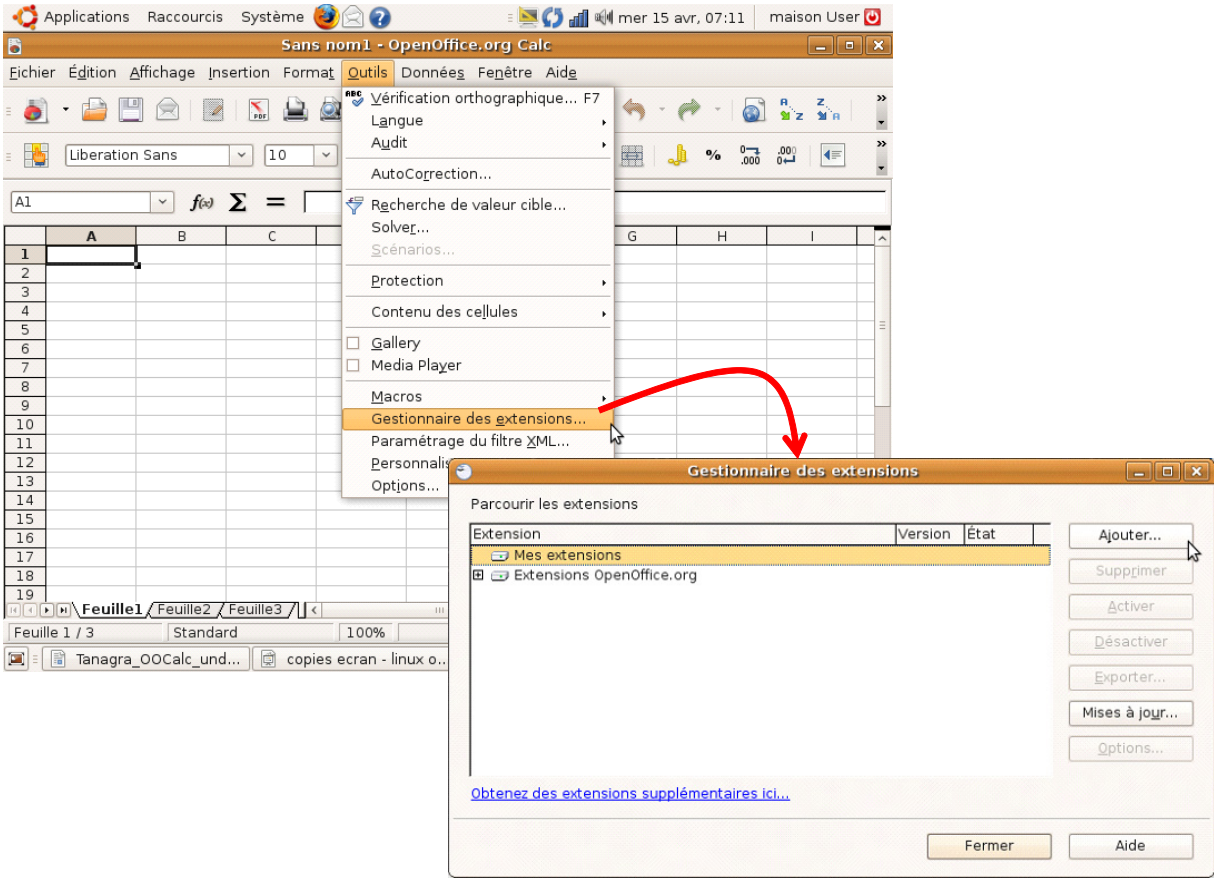
2 http://data-mining-tutorials.blogspot.com/2008/10/ooocalc-file-handling-using-add-in.html

3 http://data-mining-tutorials.blogspot.com/2009/01/tanagra-under-linux.html

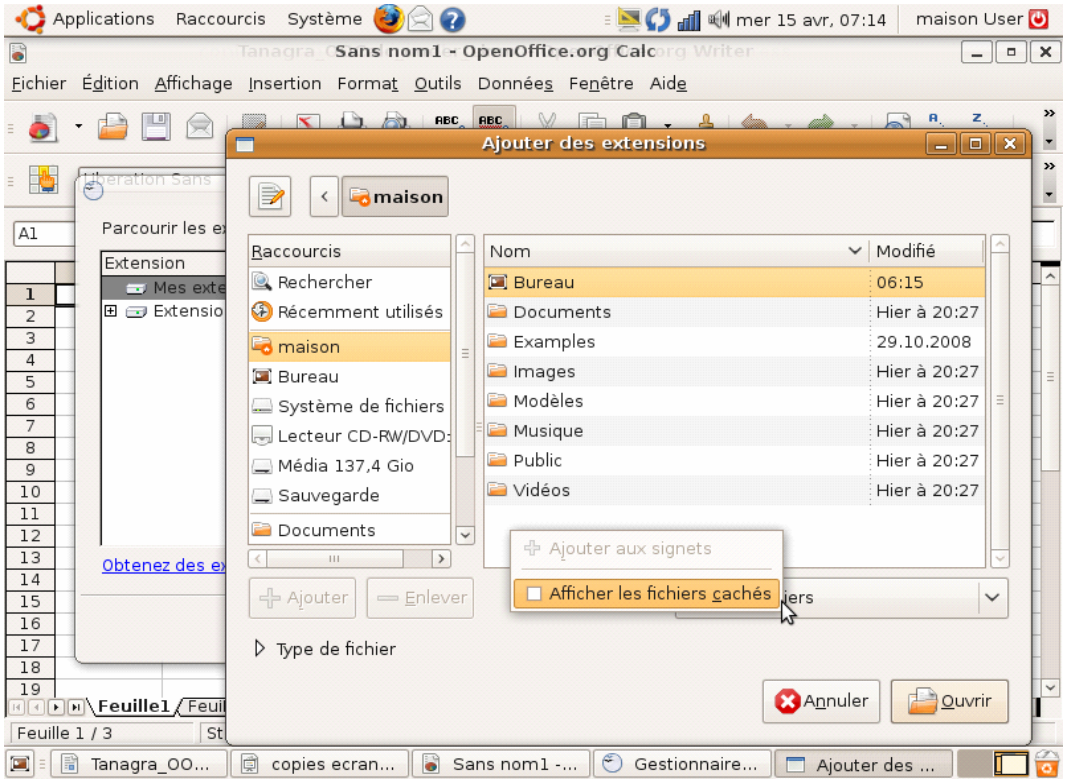
4 **WARNING. Under my UBUNTU 8.10 Intrepid Ibex, only the WINE version 1.1.18 (or previous) of WineHQ works fine with Tanagra**. It seems there is a problem with 1.1.19 (2009/04/14 version). You must downgrade your installation if you have any problem (see http://wine.budgetdedicated.com/archive/index.html).



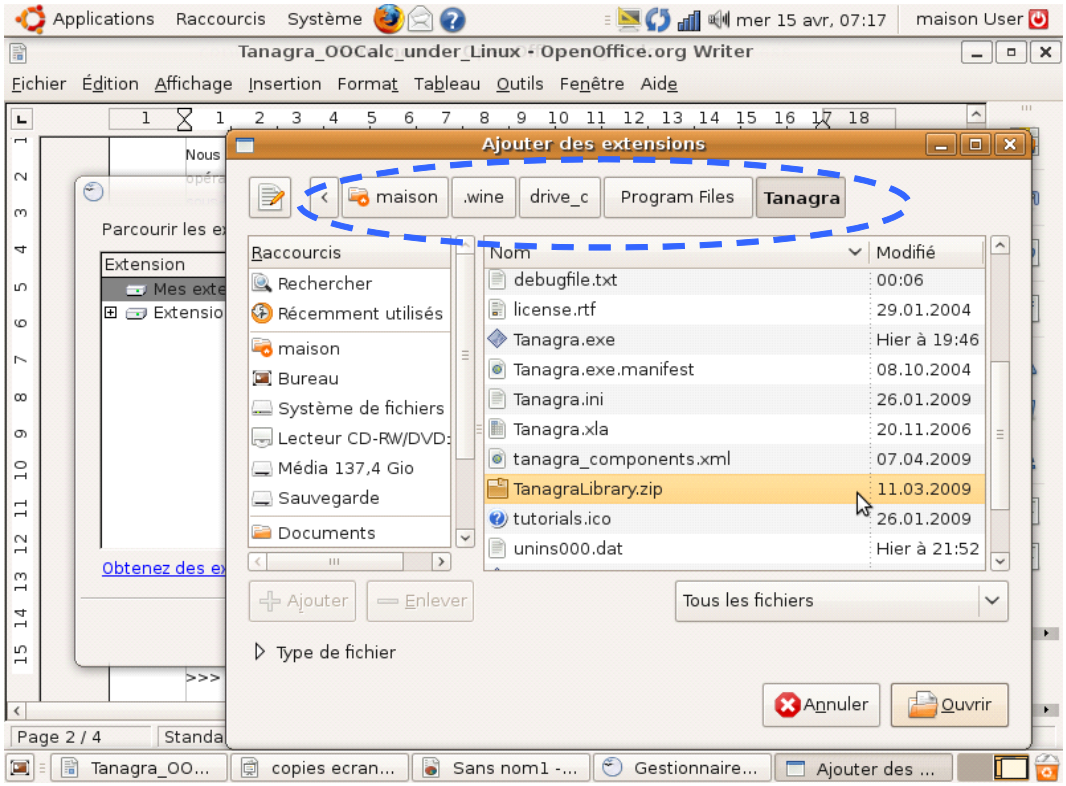
We use the 2.4 version of Open Office in this tutorial. It is the default version for Ubuntu 8.10. In order to install the Add-on, we click on the OUTILS / GESTIONNAIRE DES EXTENSIONS menu.



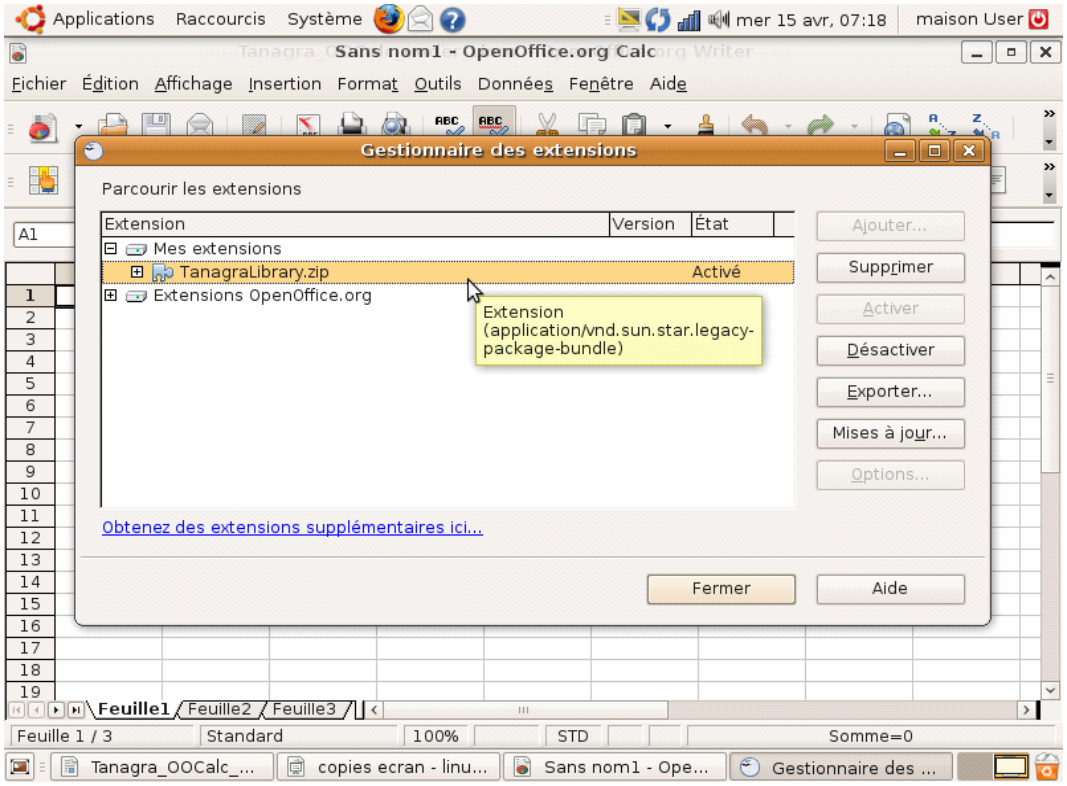
We click on the ADD button. The ".WINE" subdirectory is masked. We must first modify the settings.



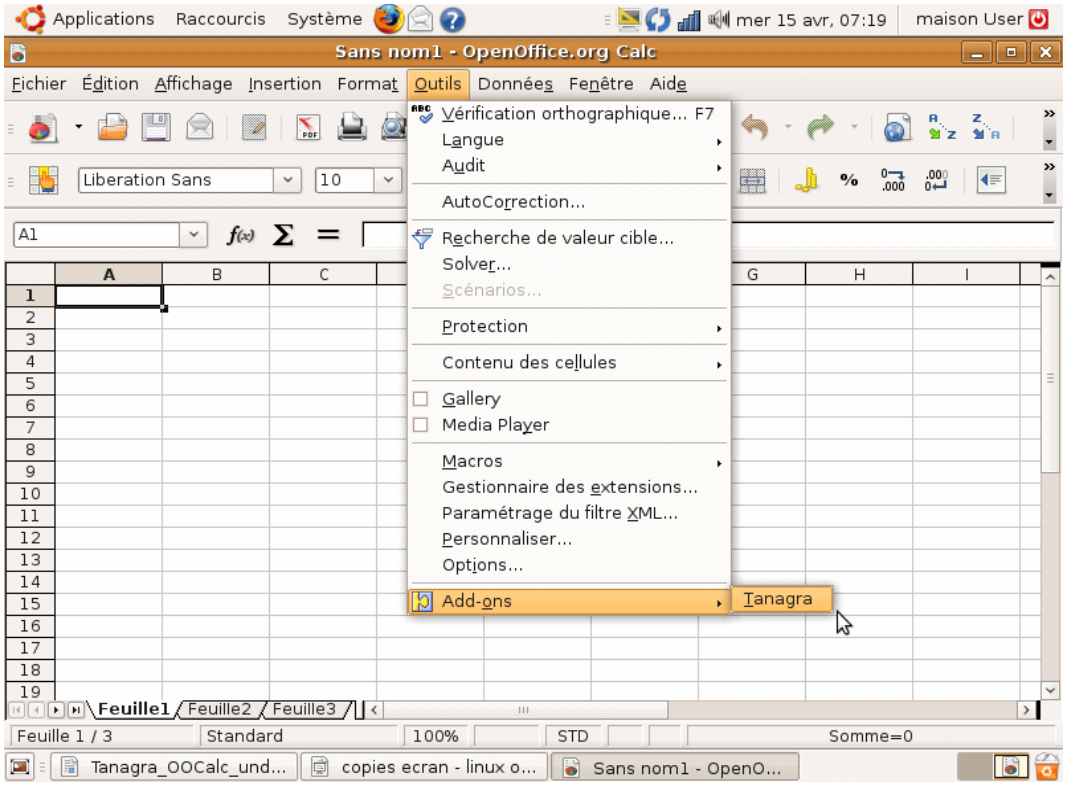
We can now reach the installation directory of Tanagra ".WINE/DRIVE_C/PROGRAM FILES/TANAGRA". We select the « TanagraLibrary.zip » file.



The add-on is now installed. We can close the dialog settings.



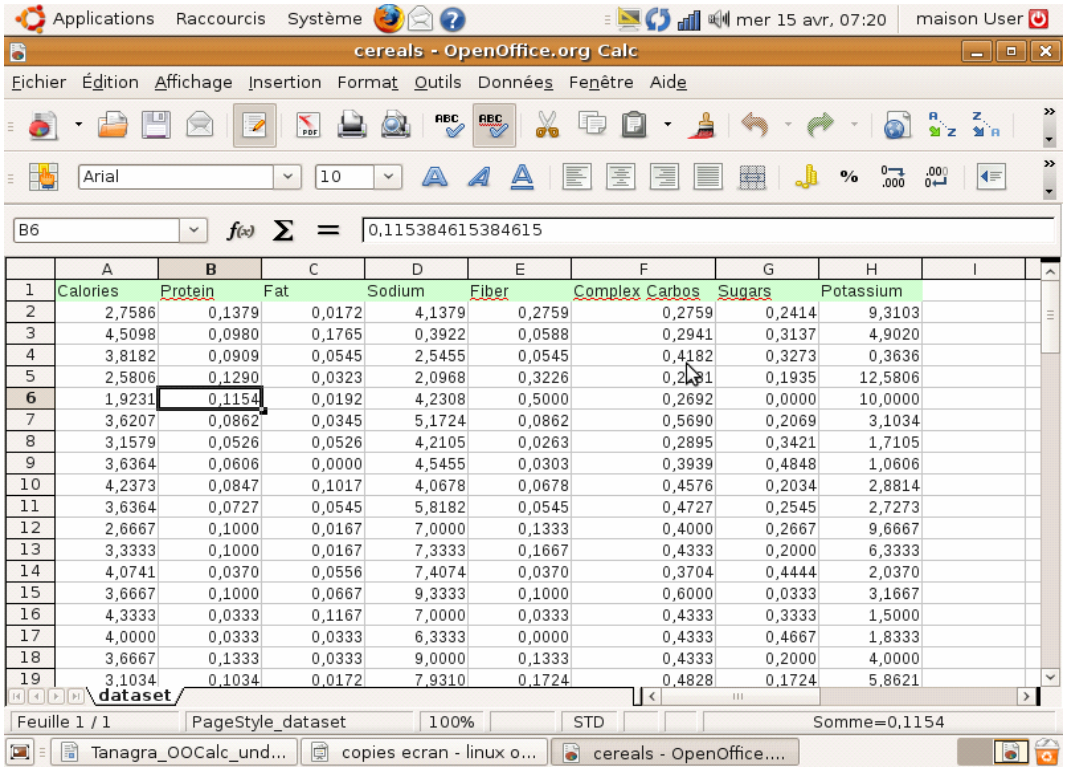
At this time, even if the add-on is inserted into OOCalc, it is not visible. We **must restart** the application. We have now a new item (Tanagra) in the TOOLS menu.



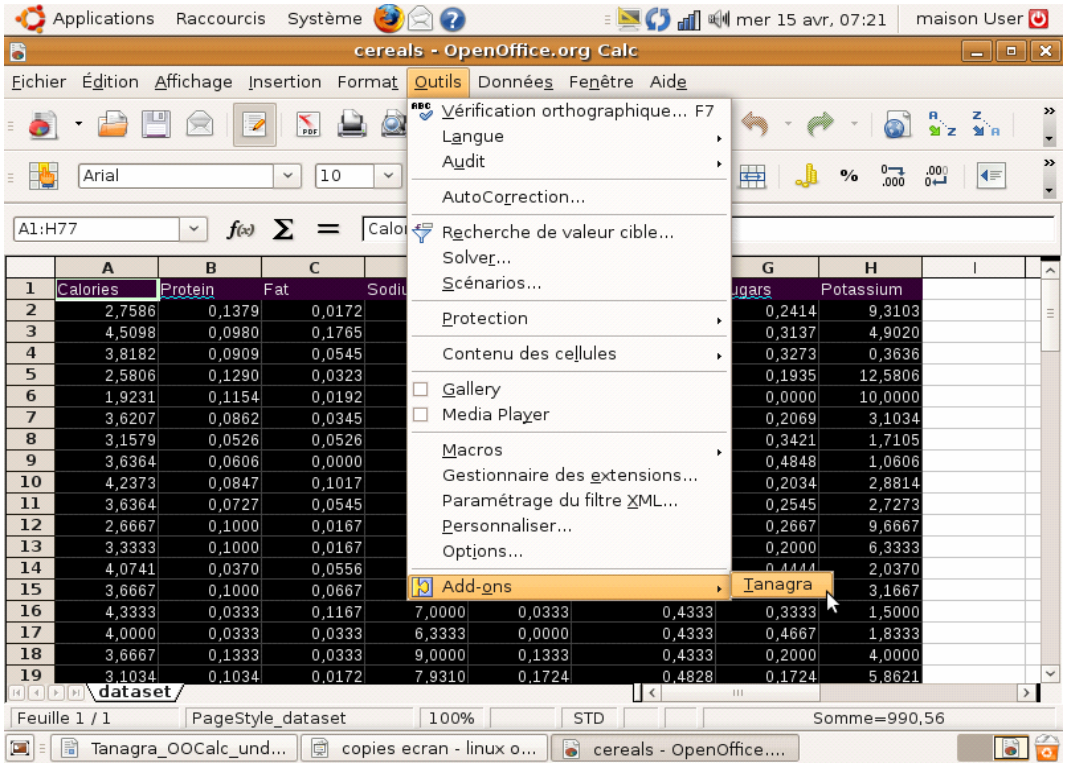
4 Principal component analysis

4.1 Launching Tanagra from OOCalc

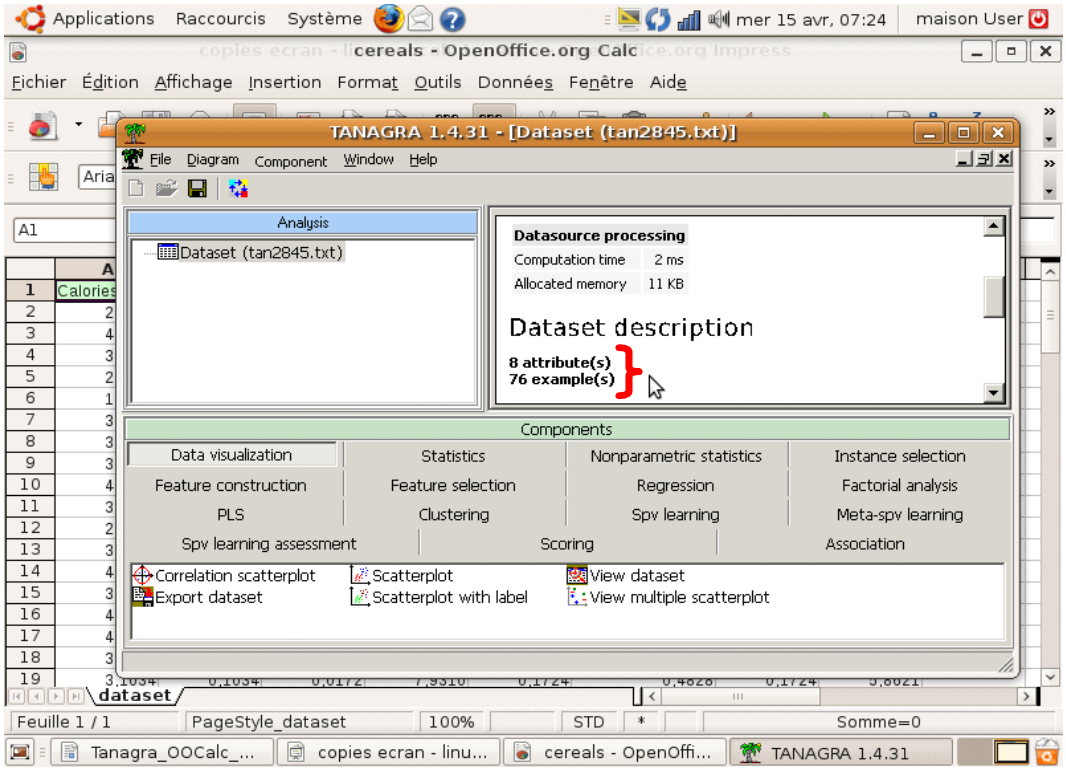
We launch OOCalc and we load the data file CREALS.XLS (FILE / OPEN menu).



We select the range of data cells, including the first row corresponding to the variable name. Then we click on the TOOLS / ADD-ONS / TANAGRA menu.

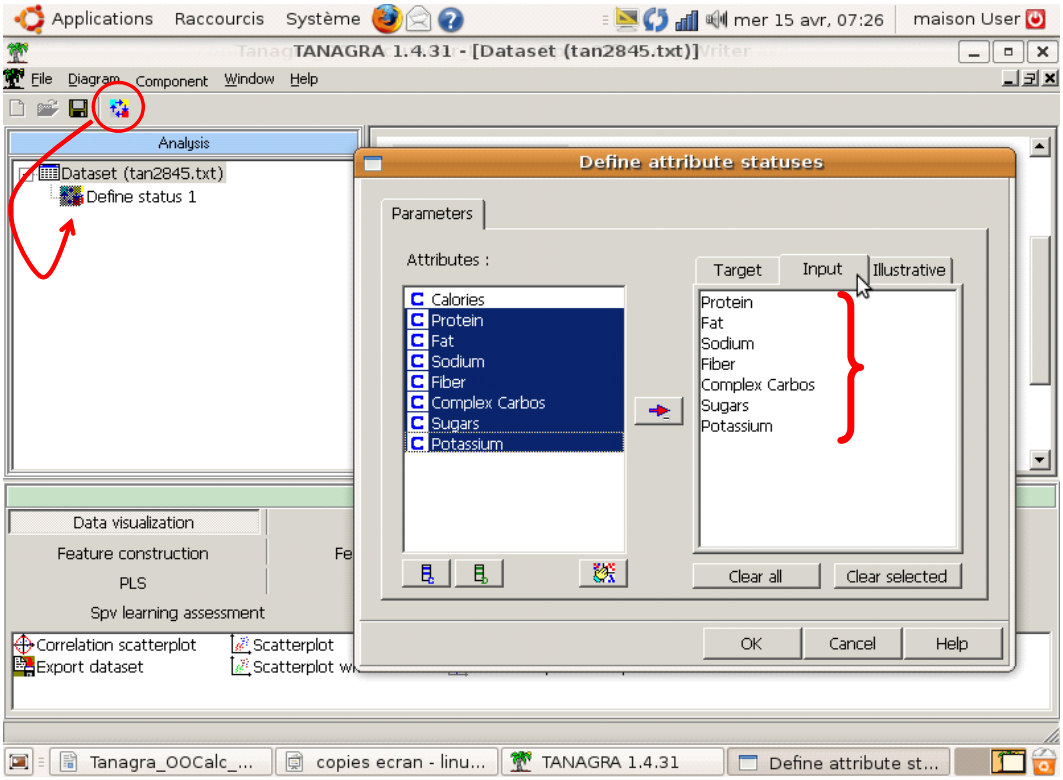


Tanagra is automatically launched. A new diagram is created and the dataset is loaded. There are 8 variables and 76 examples in the dataset.

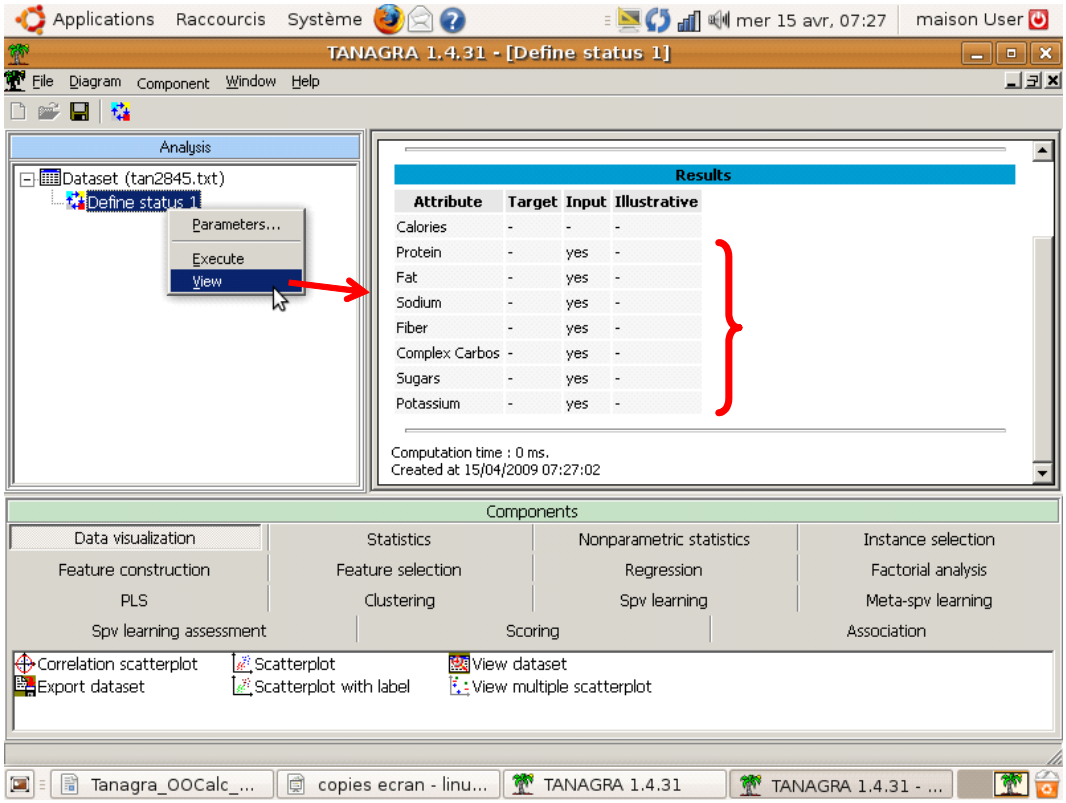


4.2 PCA with Tanagra

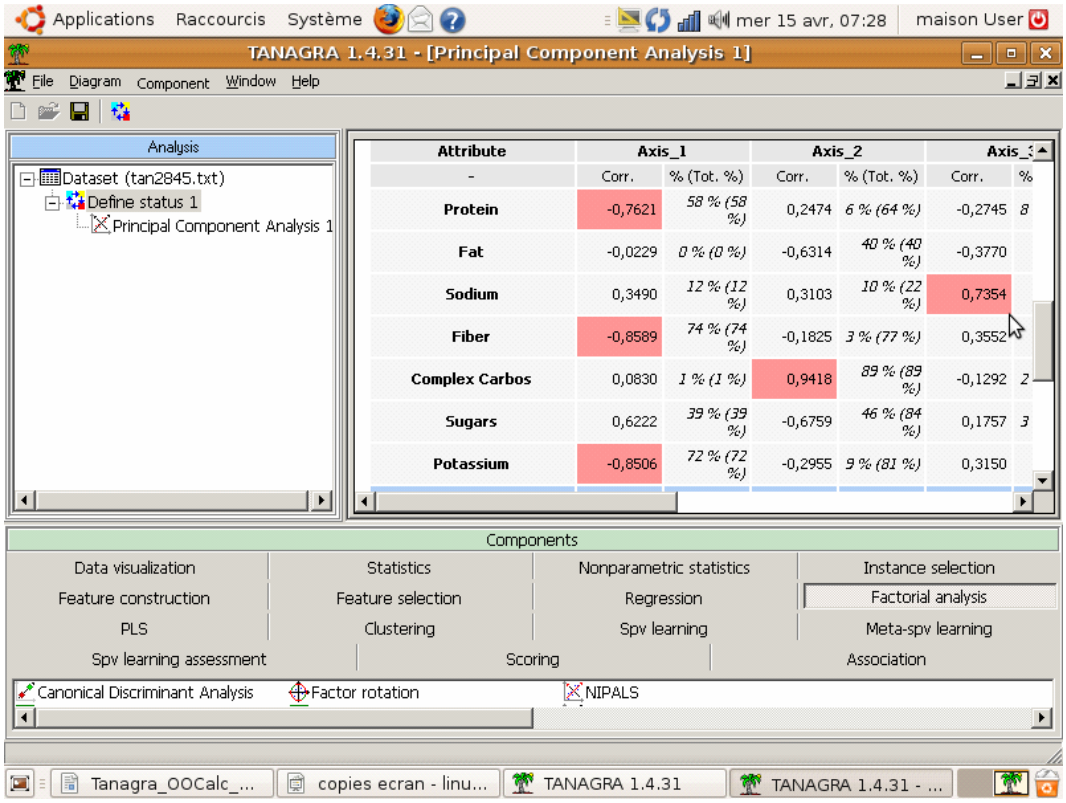
We use the DEFINE STATUS (from the shortcut into the tool bar) component in order to specify the INPUT variables (all the variables except CALORIES).



We click on the VIEW menu, we obtain the following result.



We can insert now the Principal Component Analysis component (FACTORIAL ANALYSIS tab). We click also on the VIEW contextual menu.



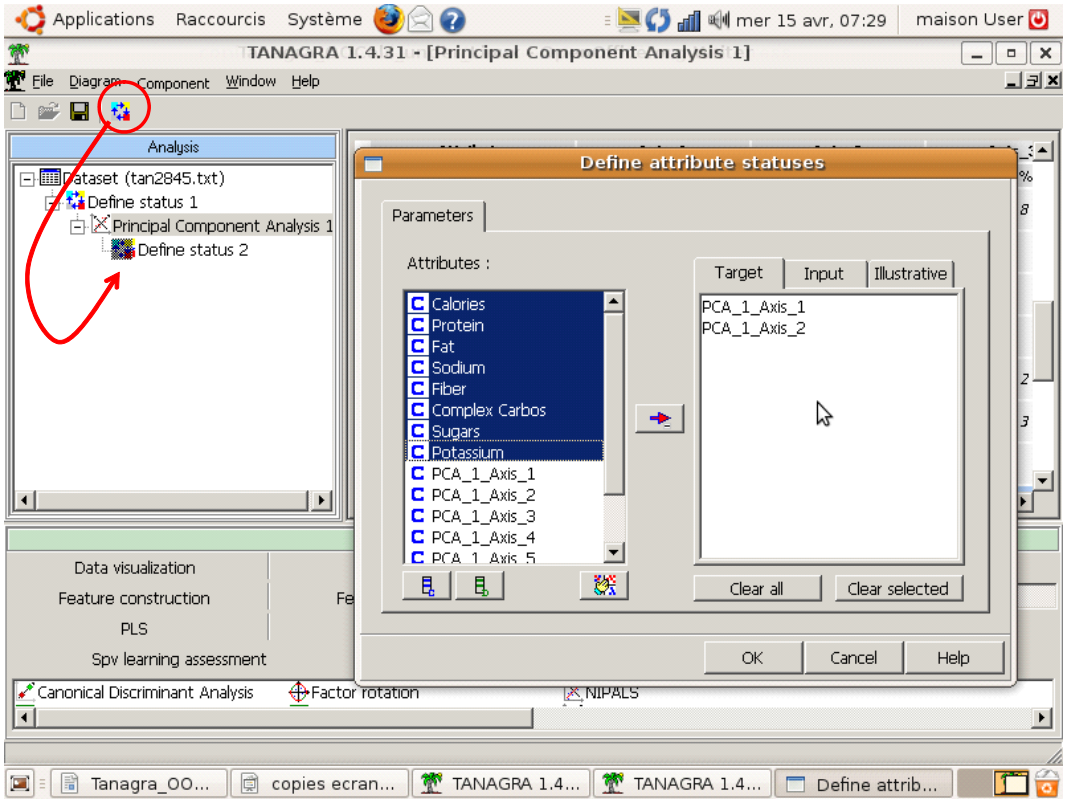
The first 3 axes concentrate 80% of the available information. We see that: (1) the products with

high sugar have low potassium, fiber and protein; (2) the products with high sugar and fat have low complex carbos; (3) the sodium seems not related to other characteristics.

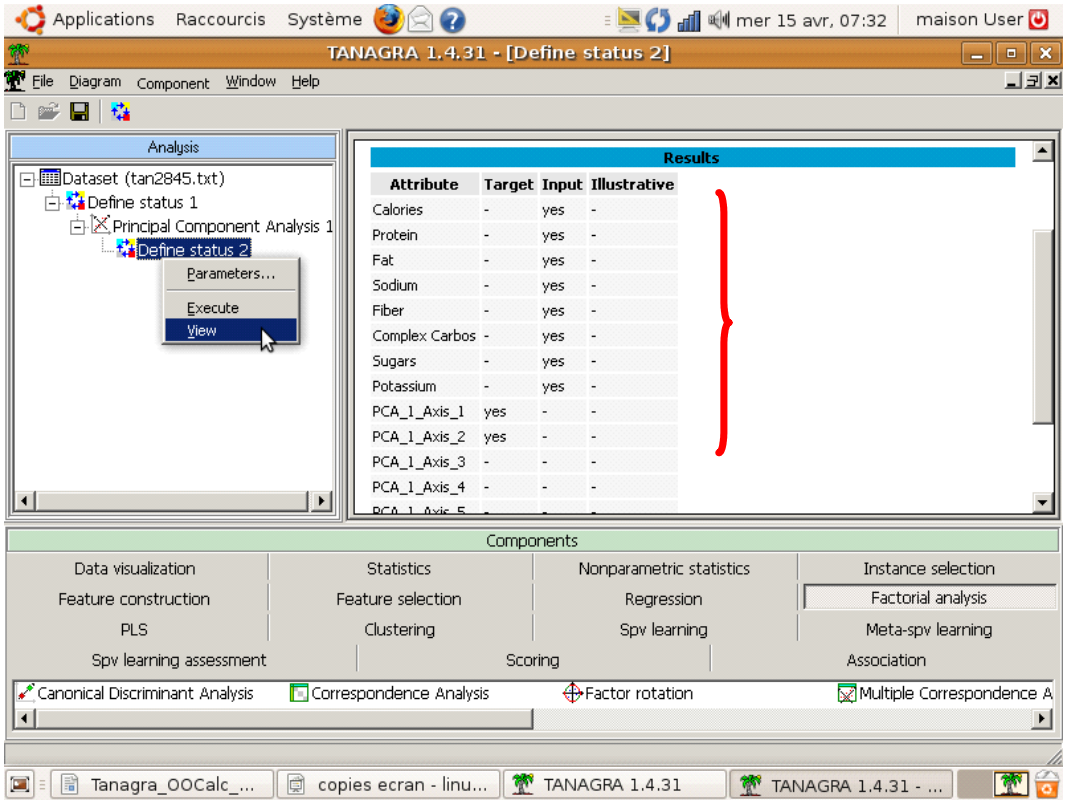
4.3 Correlation circle

The correlation circle is mainly used for the interpretation of the axes. We locate the variables in relation to axes; we can also see the situation of an illustrative variable which is not used during the learning phase. This is the case of the CALORIES variable.

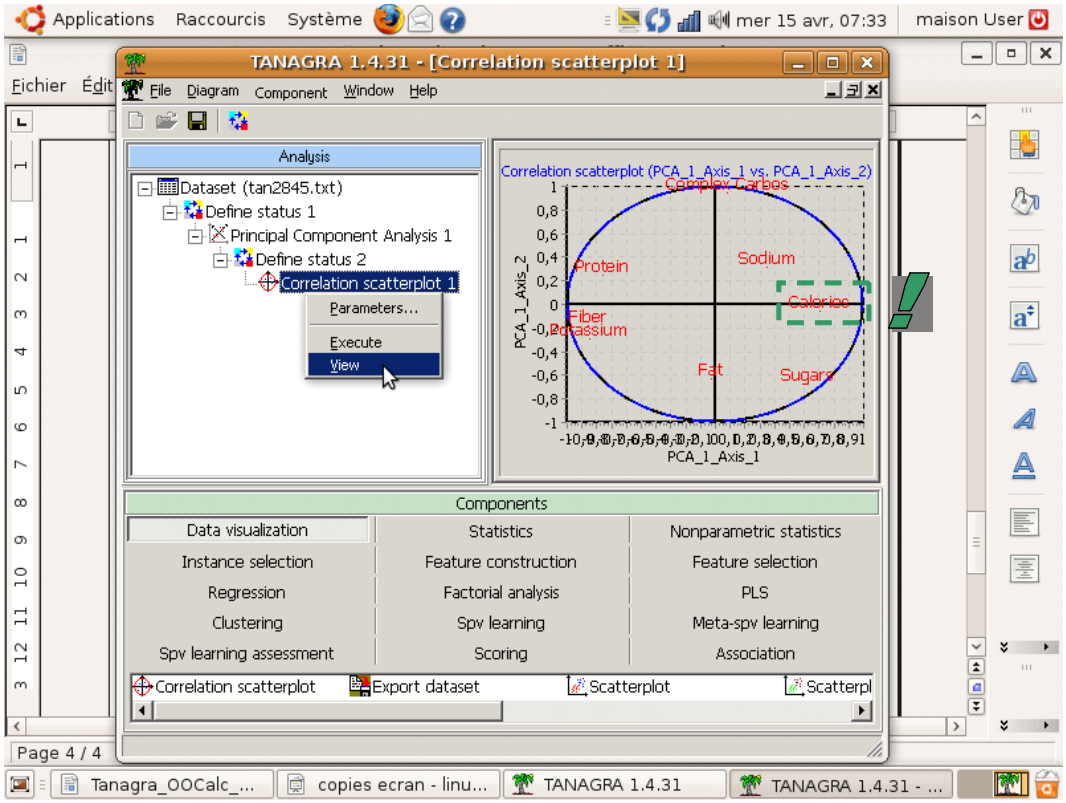
We insert the DEFINE STATUS component. We set as TARGET the 2 first axes (PCA_1_AXIS_1 and PCA_1_AXIS_2); as INPUT all the variables, including CALORIES.



We click on the VIEW menu. The types of variables are outlined.



We can insert now the CORRELATION SCATTERPLOT component (VISUALIZATION tab).



CALORIES variable seems mainly related to the first axis.

5 Conclusion

OOCalc has several advantages: it is as powerful as the state of the art commercial spreadsheet tool (Excel); it is free; it operates under various operating systems.

Making easier the data transfer between OOCalc and specialized tools for Data Mining using add-ons is a good strategy. It seems that the great majority of searchers use a spreadsheet for their data manipulation and preparation (http://www.kdnuggets.com/polls/2008/tools-languages-used-data-cleaning.htm).