Making a scientometric analysis using SAINT and visualizing it by using Gephi[1]

Contents

1. Introduction

This manual will guide you step by step through the use of the scientometric tools developed by the Rathenau Instiute and other parties. These tools allow you to make a visualization and network analysis of data from Thomson Reuters' Web of Science. To use the tools you need a computer running on windows (depending on your data size it needs to be a fast computer), Microsoft Access and firefox[2] (downloadable from: http://www.mozilla.org/en-US/firefox/new/). In addition, you need to have access to Thomson Reuters' Web of Science (http://apps.webofknowledge.com/WOS_GeneralSearch_input.do?highlighted_tab=WOS&product=WOS&last_prod=WOS&SID=W1Pj1CelLpGHLhk63KB&search_mode=GeneralSearch)

2. A short description of the software

The following software will be used in this manual:

1. SAINT, which stands for Science Assessment Integrated Network Toolkit
   Downloadable from: http://www.rathenau.nl/themas/thema/project/bibliometrische-softwaretools/download-page.html
2. Microsoft Access
3. Standard queries
4. Gephi (Bastian et al. 2009)
   Downloadable from: https://gephi.org/

[1] https://gephi.org/
Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy, *Gephi: An Open Source Software for Exploring and Manipulating Networks* (2009).

[2] Imacros can only be used in firefox. If you are downloading from the web of science by hand, you can also us other browsers.

SAINT

The Science Assessment Integrated Network Toolkit consists of the following parts

- ISI data importer (convert .txt files from Web of Science to MS Access database)
- Word splitter (strips off the morphological marking so that the word stem remains and removes stop words)
- Network tools (extracts the community structure of a network following Blondel et al. (2008))

Microsoft Access

- Database coupling and analysis

Gephi (Bastian et al. 2009)

- Visualizes networks and provides network analytics

3. A step by step manual

Step 1: Downloading data in the right format from Thomson Reuters' web of science (WoS)
We start with a description of how to download data in the right format from the Web of science (WoS). At this point, we will do this manually. For bigger datasets, one can automatize this process by using imacros, a description of which can be found at the end of this manual.

a) Go to the *Advanced Search* tab in the web of science
b) Now we will need to decide what search words to use. In this manual we will take as our dataset all articles with topic word earthquake* (i.e. the word *earthquake* and all letters that are added to it, such as earthquakes, for more information on operators and regular expressions see the help function of WoS) from 1 november 2010 to 30 november 2010. This means we type in the *search* box *ts=earthquake**.
c) Go to the first *Adjust your search settings* tab and switch off the lemmatization
d) Press the *search* button

Now, you should find a summary of your search results at the bottom of the screen.

e) Click on the number of results (depending on your subscription to the WoS you found 294 or less articles).

You see a list of the articles that contain your search term.

f) Scroll down. There you will find the options that are shown in Figure 1

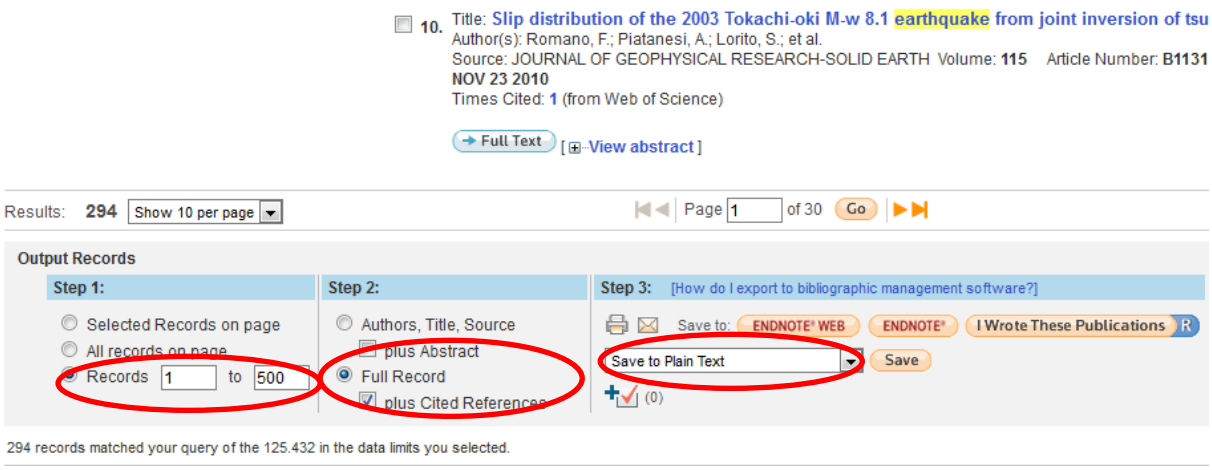

Figure 1: The options for saving records in the Web of Science

g) WoS allows you to download 500 records at a time
h) You will want to check the *Full Record* box and the *plus Cited References* box
i) The results need to be in plain text format, so in the drop down menu choose Save to *Plain Text*
j) Click the *save* button and follow the instructions for saving the text file at a convenient folder and name it.

Step 2: Converting text data to a MS Access database using the SAINT ISI data importer

To be able to analyze the data we saved we need to import our text file in MS Access. This is done by the *ISI data importer* tool.

a) On the tab *ISI input files* select the text file we just saved by using the '…' button as is shown in Figure 2 .
b) Go to the *Output* tab and provide a location and a name for the new file using the '…' button. First time users will need to click the second '…' button . (just adding a name in the file name

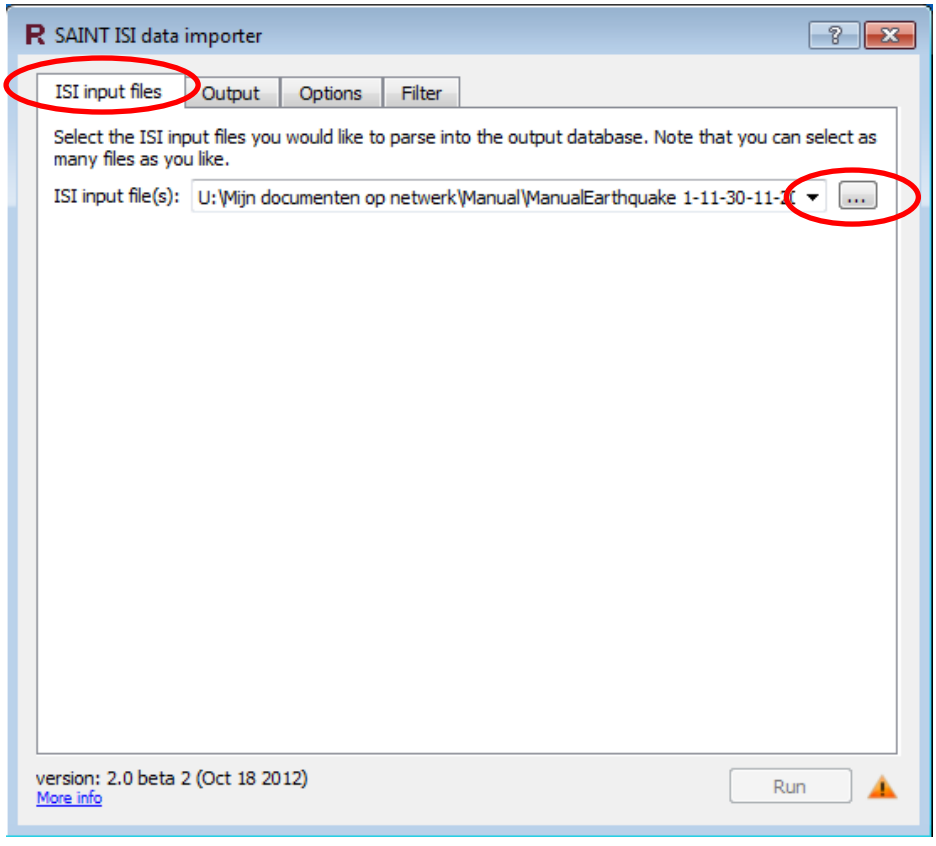

Figure 2:ISI data importer window

box which appears after clicking the '…' button will do to create a new file).

(If you wish you can stop the importer from merging authors with the same name in the *Options* tab or request only some publication types in the *Filter* tab)

If you have specified both files, the *Run* button receives a green tick and you can click it.

c) Click the *Run* button

The parser now starts analyzing the data showing its progress. When it is finished, it allows you to open the new Access file.

Step 3: Create the analyses in MS Access that you want to visualize 1: the word splitter and the word-reference co-occurrence queries

In the newly created Access file you find all kinds of tables and queries that can be used for scientometric analyses. You can of course also add your own queries. In this manual, we will discuss the queries that are used to make a make a title word-reference co-occurrence analysis (van den Besselaar and Heimeriks 2006).

If we make a list of title words, we may want to exclude stop words and we do not want plurals of the same form or adjectival forms from the same stem to be considered as different words. To this end, the SAINT word splitter was developed.

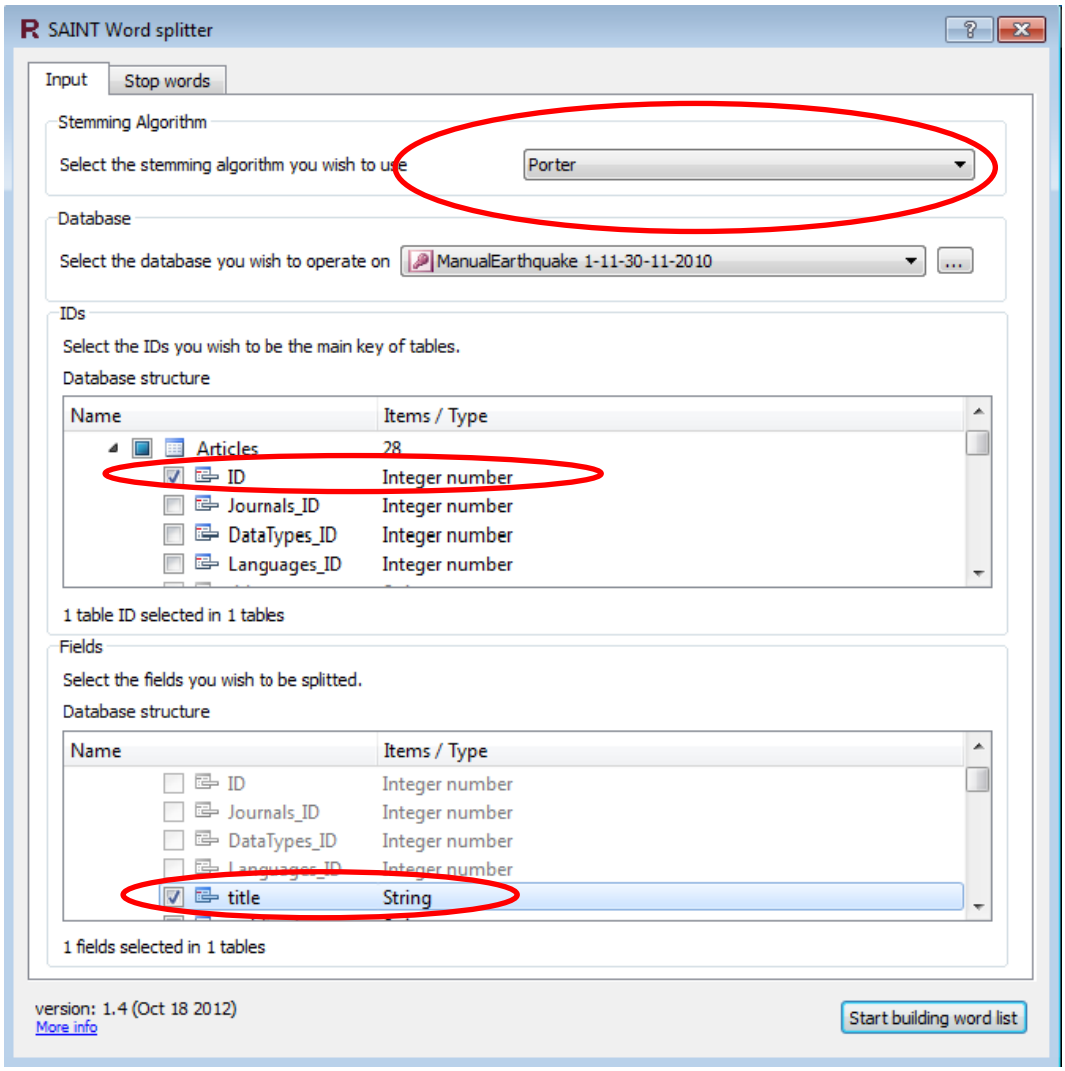

Figure 3 The SAINT word splitter selection window

a) Open the word splitter
b) Select one of the word splitters that are offered. For an overview of the differences see Jivani (2011)
c) Select the Access table with your data
d) Select in the IDs box only the ID of the column you want to split (see Figure 3)
e) Select in the *Fields* box only the column with the words/sentences you want to split (see Figure 3)
f) Go to the *Stop words* tab and select in the *text based* tab a stop words file (e.g. stopw-full.txt) or make a database file with your stop words in the *database based* tab or write some regular expressions and add them in the *regular expressions based* tab
g) Click the *Add to list button*
h) Click the *Start building word list* button

The word splitter will add a table to the Access database called *wordstems*. It may take a while before this table is visible. What may help is refreshing the database or closing Access and reopening it.

Step 4: Create the analyses in MS Access that you want to visualize 2: word-reference co-occurrence queries

We now want to know how similar articles are with respect to the combinations of references and title words. This can be done by using a set of standard queries which calculate the Jaccard index showing the similarity of articles with respect to their word-reference combinations.

We will now import (not copy!) the queries one by from the template database *Query repository*. Yet, some of them need some manual adjustment, so we will do it step by step.

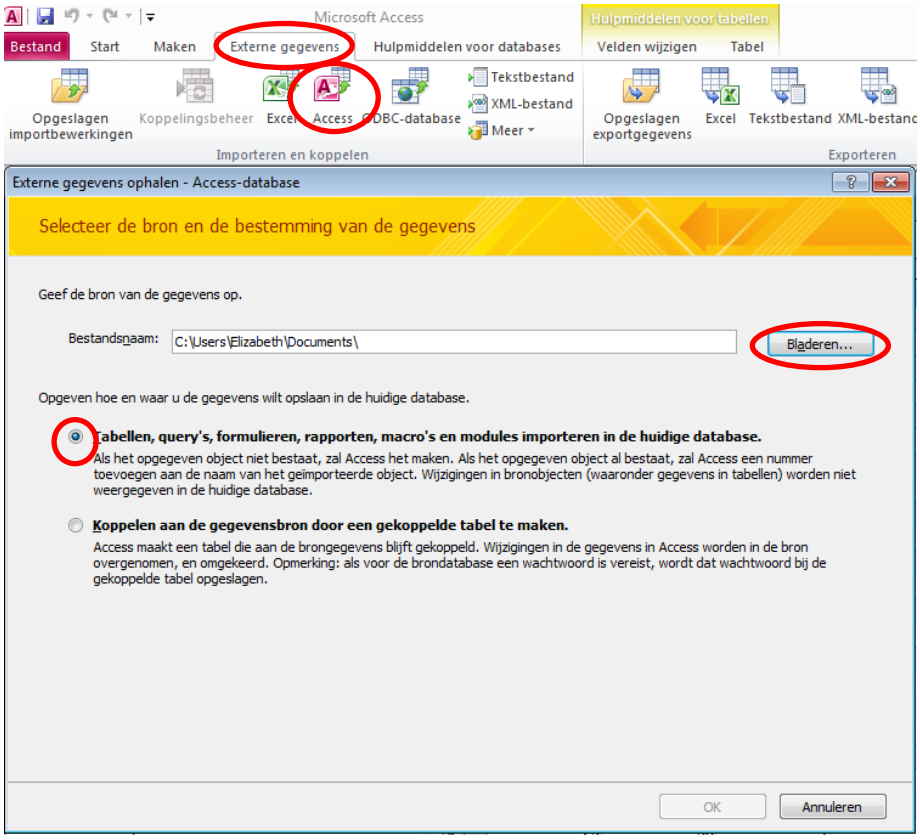

Figure 4: The Access import window

a) Import query A0 Article by means of the import wizard shown in Figure 4, from the *queries* tab, do not run it yet.
b) Open the DataTypes table and write down the numbers of the types of data you want to use (if you have already selected them from WoS, you can skip this step and the next one.)
c) Open the A0 Article query and put an expression like the following in the "criteria" row (depending on your choice of types of data): 1 Or 2 Or 5 Or 6 Or 8 Or 10 Or 13
d) Run the A0 query
e) You will be prompted with the question as to whether you really want to paste an x number of records into the new table. Press "yes".
f) Now Access has made a table *Article selection*, which contains only a column of IDs.
g) Import query A1 and run it
h) You will be prompted with the warning that you are about to change a table. Press "yes".
i) Access now has made a table *0 Article ID wordref comb* in which for each article all title word-reference combinations have received an unique number.
j) Import A2 and run it
k) You will be prompted with the question as to whether you really want to paste an x number of records into the new table. Press "yes".
l) Access now has made a table *4 selected article wordref count*, containing a column with IDs and a column with counts.
m) Import query A3 and run it
n) You will be prompted with the warning that you are about to change a table. Press "yes".

The 32 query provides us with a table called *3correlation table*. This table shows us all word - reference combinations. This means that if all articles share all words and references the maximum number of rows in this table is the number of references * the number of title words * the number of articles . So if the dataset is very large and homogeneous, the total number of rows can become very large, bringing the limits of MS Access into sight (2 GB).

o) Import queries A4-6 (there are 2 A4 queries, import them both) all together, but run only A5 and A 6 in their respective order (A4 queries will automatically run).
p) Now close Access

The queries we just ran calculate the Jaccard similiarity coefficient (also called Tanimoto similarity coefficient)[3] for the word-reference combinations of the articles. This coefficient allows us to think of the strength of connections between articles as edges in a network. Within a network one can have clusters (or communities) of more closely connected nodes. The next step will therefore be to calculate which communities can be found in the articles on four levels of aggregation. To do this, we will use the network community detection tool based on a paper by Blondel et al. (2008), which is part of the SAINT network tools.

q) Open the SAINT Network tools and choose the community detection tool
r) A short explanation of the theory behind the tool appears. To continue click the *next* button
s) Select your database and click *next*

[3] N.B. The formulas of Jaccard and Tanimoto for the distance coefficient are not the same.

t) You are now requested to choose a table containing your nodes. The queries we just ran made a *Labels* table containing the Article ID, year of publication, journal and the times it is sited. Select this table.
u) Now the wizard wants to know in which column the IDs are, so select the ID column from the list
v) The next step is to select the edges, which can be found in the *TWCR Matrix* table (TWCR stands for *title word cited reference*)
w) The program now needs to know which nodes from the *TWCR Matrix* table are to be connected. Normally, the tool will automatically select ID1 and ID2 from the table, if this is not the case take the numeric variables.
x) The last step is to allow the tool to weigh the strength of the connection between the nodes. Select *the Entries have a defined strength* option and select in the field below the Jaccard column
y) Now you need to decide where you want your communities to be saved. In order to use them later in Gephi, it is practical to save them as fields in the nodes table.
z) The last step is to give a name to the new columns. It is useful for the other queries we will run to call them com%1. Below in the same window, you see an overview of the new *Labels* table.
aa) Press *commit*

In the results box you will now find information about the communities that were detected on maximally four levels of aggregation. In our small dataset, there are only three levels of aggregation present, which we will need to remember for the following steps.

Now we return to Access. To get a good overview of the available information in the database, we will import the following standard queries: A7 *Make labels final*, the B tables *Origins of trains* and the *Summary table*, the C tables which all have *count* in their name and the Z queries as was shown in Figure 4.

bb) Before running query A7, open the query and check whether the names of columns are correct and whether the number of communities found in the *results* box of the community detection tool matches the number of communities in the query. If there are less communities in your dataset, delete the communities that do not exist.
cc) Run query A7 , this will create a table called *Labels redux*, which contains an overview of the properties of the articles in the dataset.
dd) Open query *B0 summary table* and make sure the highest community level available is found in the position in which you find com4 (i.e. if there are less than 4 levels, take the highest level and change all occurrences of com4 into the highest available level. In our example database we need to change com4 >> com3).
ee) Run the B0,1 and C1-5) in their respective order.
ff) Run the C6 query, you will be prompted with an “Enter parameter value” dialogue box click OK
gg) Click OK again when a second “Enter parameter value” dialogue box appears
hh) Open query *Z1* and make sure the level of the label *com4* is the highest community level available (i.e. if there are less than 4 levels, take the highest level and change all occurrences of com4 into that level. In our example database we need to change com4 >> com3)
[Some of the community level names still need to be adjusted to com1, com2 etc.]

The Z queries will put the data together that are needed for exportation to Gephi. However, Gephi requires semicolon separated csv files. Therefore, we will need to save the nodes and edges tables that were created by the Z queries (respectively *Nodes to csv* and *TWCR-Edges to csv*) as textfiles. This can be done in the same tab as we used to import queries.

ii) Open the external data tab and click on export to text file button as shown in Figure 5.

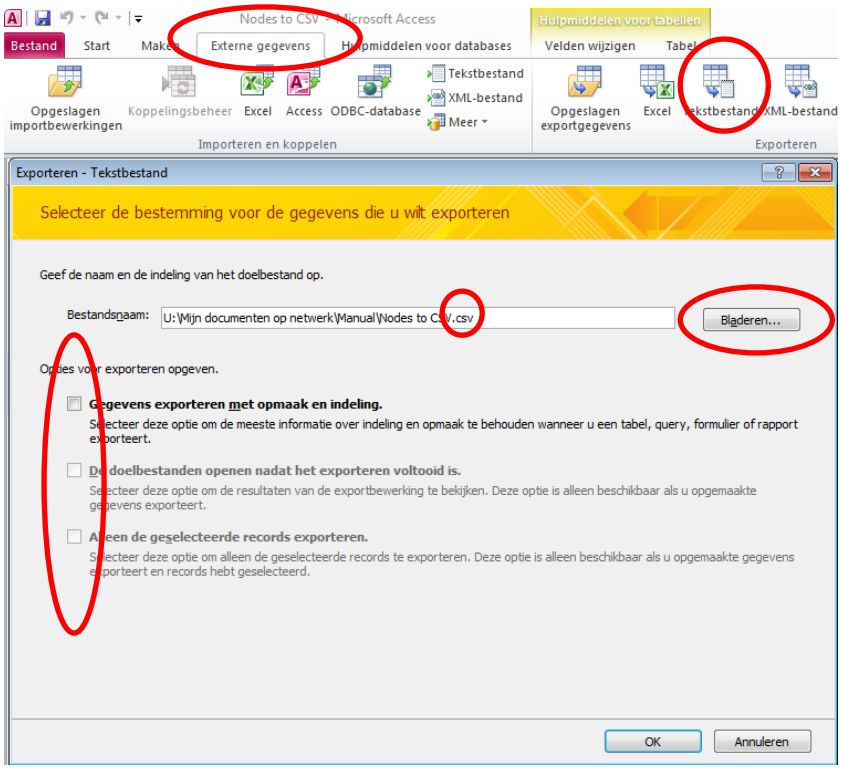

Figure 5: The save to text window

jj) Make sure the file will be saved in the right place and replace the ending .txt by .csv.
kk) Do not check any boxes and press OK
ll) In the wizard that follows make sure you choose in the first step *delimited*, in the second step *semicolon* as delimiter, [none] as text qualifier and field names in the first row is switched on and save the file.
mm) Do the same for the edges file.

Step 5: Import your files in Gephi

a) Open Gephi and choose *new project*
b) Go to the *Data laboratory* tab
c) Make sure the nodes tab is on at the upper left side of the screen and click the *Import spreadsheet* button
d) Select the nodes file and make sure the separator is set to semicolon and the *as table* is set to nodes table
e) Select all columns
f) Check whether all your data have been imported via Window>context (Gephi is still beta)

g) Change the view to *edges* at the left side of the screen and import the edges (don't forget to change the *as table* option to *edges* and to check whether all your data have been imported.)
h) Save your file

For a description of how to use Gephi see https://gephi.org/users/

Step 6: Get a good overview of the data by means of a spreadsheet

By means of the C queries, we have retrieved interesting information on frequencies. To get an overview of the larger and most frequent patterns, we will put the results of these queries and any other ones you would like to make in an excel spreadsheet, which will show us an overview of the most frequently found patterns.

a) Open the spreadsheet template [TEMPLATE description of trails]

The template has several tabs, the first of which will provide you with a summary of the information in the other tabs. The first part of the summary is already part of one of the standard queries.

b) Copy the result of the standard query *B0 Summary table* in the upper left gray field containing the text *community level 2*. This text is just an example. We will be copying the results of our community level 1.
c) Copy the result of the *C2* query in the upper left gray field of the *keyword* tab
d) Copy the result of the *C6* query in the upper left gray field of the *title word combinations* tab. This may result in an overload for the clipboard and you may need to do the copy pasting in chunks.
e) Do the same for the respective query results of C3 in the tab *journals*, C1 in the tab *authors*, C4 in the tab *cited references* and C5 in the tab *top cited papers*.
f) Now make sure the formatting of the E to J columns is extended to all rows that are covered by the A-C columns for every tab by selecting the first rows of these columns that are filled in already and dragging them down from the lower right corner of your selected field (a black cross is visible when pointing your mouse to that corner).

Now the whole spreadsheet is filled in. In the yellow box you can change how large your selection is (i.e. the top 10, 20, 30).

However, you may want to make queries yourself and put them in the spreadsheet.

g) Make a query of the form: *com1, [information, e.g. country, institute etc.], article ID*. (For how to do this consult a MS Access manual)
h) Sort your query first by *community* and then by *Count of Articles_ID*
i) Go to (for example) the *keywords* tab in the spreadsheet and right click on the tab label
j) Select *copy or move* and in the following dialogue box *new worksheet*
k) Tick the *create a copy* checkbox and press ok
l) A new tab has appeared. Select columns A, B and C, right click and clear the contents
m) Copy the content of your MS Access query into the emptied columns, beginning at the fourth row of the first column
n) Give the page a new title and rename the tab

o) Go to the *summary* tab and add a new column after column L (not at the end of the row, but somewhere in between, later you can change its position if you wish).
p) Copy one of the other columns into the new column
q) Change the tab reference of the formula in the first cell to the name of your new tab
r) Drag the formatting of the first cell as far down as needed. The cells will fill with a summary of the information that was provided in the new tab.

References

Bastian, Mathieu, Heymann, Sebastien, and Jacomy, Mathieu (2009), *Gephi: An Open Source Software for Exploring and Manipulating Networks*.

Blondel, V. D., et al. (2008), 'Fast unfolding of communities in large networks', *Journal of Statistical Mechanics-Theory and Experiment*.

Jivani, Anjali (2011), 'A Comparative Study of Stemming Algorithms', *International journal of computer technology and applications*, 02 (06), 1930.

van den Besselaar, P. and Heimeriks, G. (2006), 'Mapping research topics using word-reference co-occurrences: A method and an exploratory case study', *Scientometrics*, 68 (3), 377-93.