## *1:W35: Regular expressions and spreadsheets*

1. What regular expressions do you use to extract all the dates in this blurb: <http://bit.ly/regexexercise2> and to put them into the following format YYYY-MM-DD?

***1.1 ANSWER***

(\d+).(\d+).\s?(\d+)

Substitution:

$3-$1-$2

Result:

Et billede, der indeholder tekst

Automatisk genereret beskrivelse

1. Write a regular expression to convert the stopwordlist (list of most frequent Danish words) from Voyant in <http://bit.ly/regexexercise3> into a neat stopword list for R (which comprises "words" separated by commas, such as <http://bit.ly/regexexercise4>). Then take the stopwordlist from R <http://bit.ly/regexexercise4> and convert it into a Voyant list (words on separate line without interpunction)

***2.1 ANSWER – from Voyant list to R list***

(\S+)\n

Substitution:

“$1”,

Result:

Et billede, der indeholder tekst

Automatisk genereret beskrivelse

***2.2 ANSWER – from R list to Voyant list***

"(\S+)",

Substitution:

$1\n

Result:

Et billede, der indeholder tekst

Automatisk genereret beskrivelse

1. In 250 words, answer the following question: "What are the basic principles for using spreadsheets for good data organisation?"

***3.1 ANSWER***

*To ensure that one is using spreadsheets properly for good data organisation several things must be kept in mind. First, one should consider the software that will subsequently be fed the data: programs like R or Python cannot tell different tables in one sheet apart neither combine them to one, which makes it crucial to make your data well ‘placed’ and structured before entering another software in which the analysis will be done. This means that one sheet should consist of no more than one big data frame, in which all observations should be noted.*

*Consistency is key when notating each observation: all values in each column should be on the same format and no cell should be left empty. Additionally, only one value should appear in each cell and common notation should be used, for example dates should be noted like YYYY-MM-DD. One should never notate the measure of the observation inside the cell (like 0.8 sec) but instead use intuitive naming of the columns (like time\_sec) and then leave the cell with the value only (like 0.8).*

*One should keep the data raw, and thus not use calculations or font or colour highlighting, since each step of the analysis should be transparent to others: each calculation step will not be saved in a typical sheet and the raw data will worst case be lost. The highlights will not be transferred into programs like R or Python anyways, and thus doesn’t make sense to introduce anyway. Lastly, one should remember to make both backups and data validation and hereby ensure that data entry errors are avoided.*

1. Challenge (OPTIONAL)! Can you find all the instances of 'Dis Manibus' invocation in the EDH inscriptions in <https://bit.ly/regexexercise5>? Beware of the six possible canonical versions of the Dis Manibus formula (see day 1 slides)!