Contents

List of Figures

Basic Guide to working with S & P CapitalIQ

Stefan Miels

April 28, 2020

Abstract

This guide aims at providing student assistants and students writing their thesis at the Institute of Accounting and Auditing with an introduction to using S & P CapitalIQ. The guide is restricted to basic topics and makes no claims to completeness or perfect accuracy. Still, I am confident that it provides sufficient assistance to your thesis or research work. If you come across information that is either false or lacking, I would be happy to receive feedback from you (stefan.miels.1@hu-berlin.de).

1 Preliminary Questions

In this section, I will address some of the questions that you need to have answers to before beginning to work with CapitalIQ. By the time you read this guide, you are likely to have already answered some if not most of these questions yourself. For this reason, I encourage you to glance through the headers of the subsections and then assess whether reading a subsection will yield you any new information. From section 2 onwards I will cover information that should be of use to all readers of the guide.

1.1 How do I get access to the database?

In order to get access to S & P Capital IQ , you need an active account with our faculty's Lab for Empirical and Quantitative Research (LEQR). In case you do not already have an account, you are required to apply for one via the LEQR's homepage (leqr.wiwi.hu-berlin.de | Button 'Membership'). The application has to be signed by the person that you named as your supervisor and handed in by you at the LEQR's office (Room 131 at SPA1) afterwards. The LEQR staff will then provide you with a PDF file containing both your account data as well as instructions for accessing the various databases which the faculty has licenses for. The crucial part for you is that you will need to connect to remote desktop **capitaliq.wiwi.hu-berlin.de**.

1.2 At what times can I use CapitalIQ?

You are required to reserve a timeslot for yourself each time you want to connect to the remote desktop (Figure 1). This is due to the fact that only one faculty member is permitted to use

the connection at any given time. In principal, you can reserve timeslots at any time of the week. However, access is limited to four hours per user per day. Unfortunately, there are some irregularities in the availability of the LEQR's servers. When these happen while you want to access the database, your best option is to contact the LEQR staff and inquire about the problem (rdc-team@lists.hu-berlin.de).
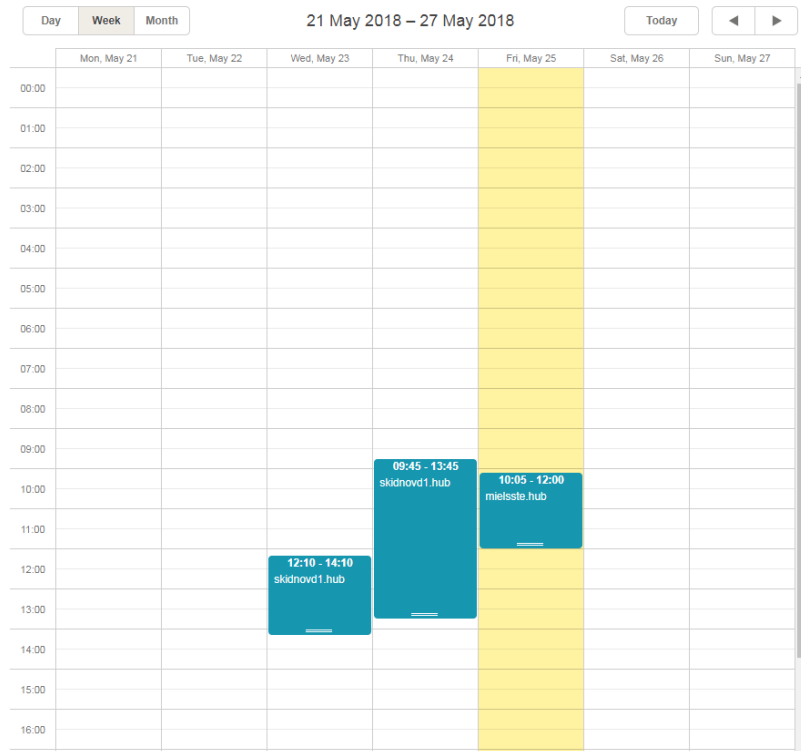

Figure 1: Scheduler

1.3 Do I need to be at the faculty in order to access the database?

No. Using the databases requires being connected to the HU network, but not physical presence at the university. You can obtain access to the database from anywhere in the world by setting up your LEQR homeshare and connecting to the HU network via VPN. For more information on this issue, please refer to the instructions that were provided to you together with your LEQR account data.

1.4 Do I need to have any special programs installed to make use of CapitalIQ?

Data is retrieved from the database through an Excel Add-In, making Microsoft Excel the only program required for using CapitaIQ. Excel is installed on the remote desktop, so you do not need to have it on your own system. Since you may want to copy the data from the remote

desktop to your local hard drive in order to work with it, you might however need to figure out a way to store the data with whatever application you have available.

1.5 What level of Excel expertise is required for working with the Add-In?

Working with the Add-In only requires very basic knowledge of Excel's functionalities. In fact, the only two functionalities that you will probably not get around using are *Absolute and relative cell references* and *Autocomplete*. Even if you happen to be a novice at Microsoft Excel, getting a grasp of these concepts is unlikely to take you more than an hour. Additionally, I will provide some examples of their application in the context of the Excel Add-In in a later chapter. In short, a lack of Excel knowledge should not deter you from working with commercial databases.

2 Screening for companies

In most cases your data collection task will begin after you have already figured out a set of criteria by which you want to search for companies. In order to use these criteria for formulating requests to the database, you need to access CapitalIQ's web client. The client can be found on the website *capitaliq.com*. The website needs to be accessed(Figure 2), since this is the only way by which you can gain access to our faculty's account.

Figure 2: Web client login

Inside the web client you can use the *Screening* ribbon to access an interface which allows you to conduct company searches by means of a plethora of different criteria (Figure 3) . Alternatively, it is possible to search for equities or securities. However, given that company searches represent the most prominent use case of databases in empirical accounting research,

this guide is restricted to describing these particular requests [1]. The web client displays the accessible criteria in a list that is categorized by topic (see Figure 4).
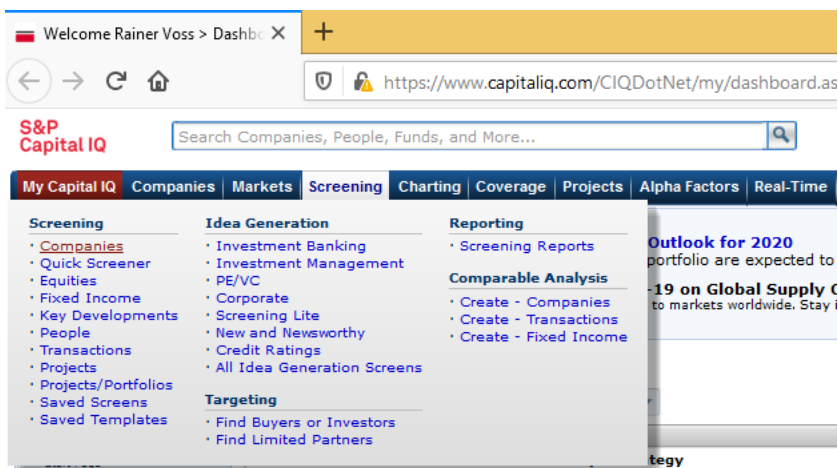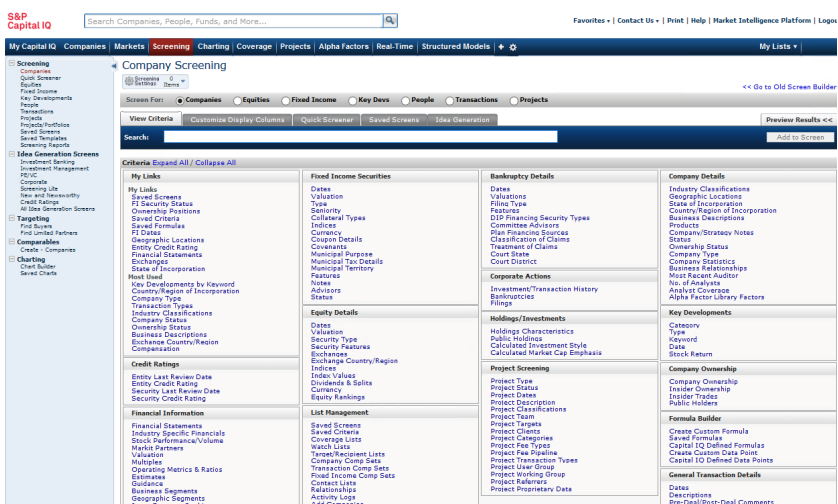

Figure 3: Screening ribbon

Figure 4: Screening Interface

Since you are likely to use more than one criterion at a time, it will be necessary to employ a methodology that allows you to connect multiple criteria in one request. The connections are made through the use of the Boolean search operators. Roughly speaking, these operators determine whether the criteria have to be fulfilled cumulatively. If you have no experience in working with logical operators, you can refer to the definitions given in the web client, which pop up when you hover over one of the operators (fields "NEW", "OR" and "AND" in Figure 5). CapitalIQ displays an overview of the criteria that you used as well as the applied search operators (Figure 6). Once you have specified your full requests (which are called *Screens*), you

---

[1]From what I have seen, the interfaces for equities and securities are not all that different from the one for companies. So most of what is described here should also prove useful in context of these applications.

may save them so that you can quickly access them later. When doing so, please keep in mind that the CapIQ-user is a shared account which can be accessed by all members of the LEQR. Consequently, the list of saved screens contains not only your own past requests, but also the ones of other faculty members. For this reason, it is vital to give your screen a unique name that you will be able to recognize later on.
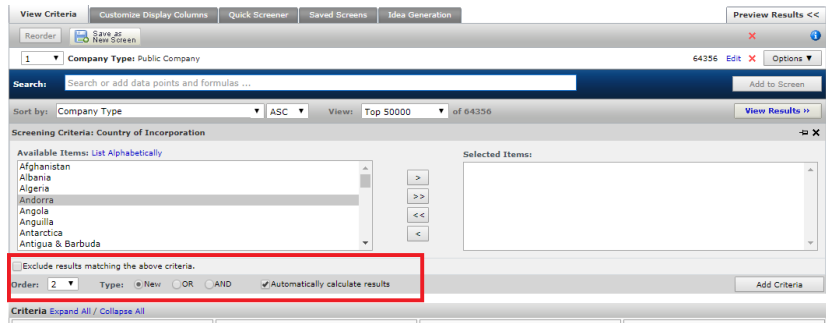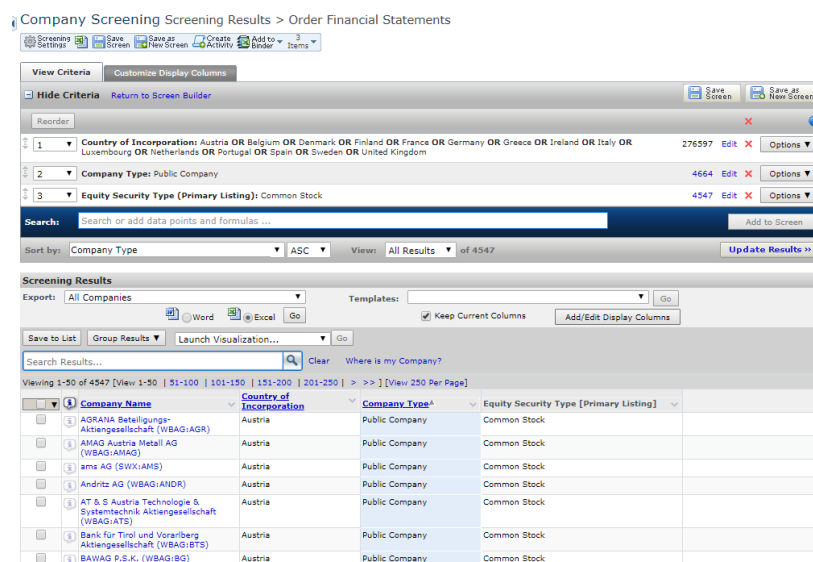

Figure 5: Search operators

Figure 6: Search criteria

As soon as you have finished formulating your request, you may view its results and export them to Excel. Be that as it may, you should refrain from immediately exporting and instead take a moment to examine your results list. The list will feature the names of the companies you have found as well as the firm characteristics that are related to your search criteria. You have the option to add further characteristics before exporting. For most of the firm characteristics this is a superfluous exercise, since obtaining them via Excel will be far more convenient. Nevertheless, there is one characteristic, the so-called *Excel Company ID*, for which it is crucial to attain it at this early stage. When we work with Excel functions later on, this variable is used to identify the companies that you are referencing and is therefore indispensable for obtaining

any information [2]. Subsequent to this, you can export your data in either Word or Excel format (see Figure 7). As Microsoft Excel is used for data retrieval later on, the Excel-format should be the strictly better choice in almost all scenarios.
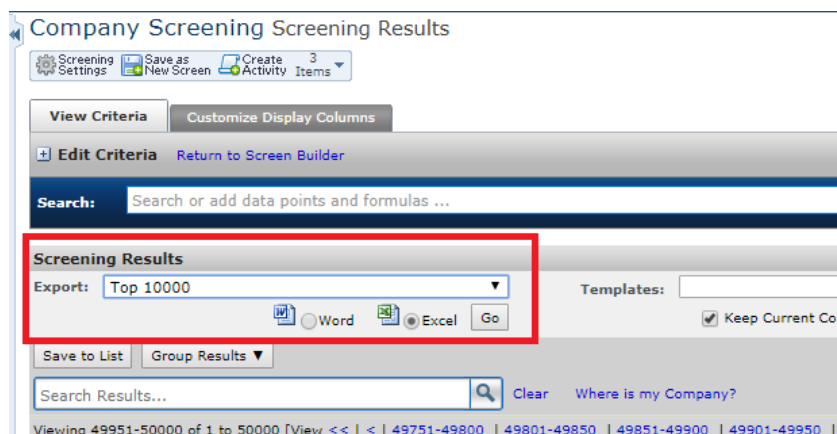

Figure 7: export options

Once you have exported your results, you will obtain an Excel file consisting of two worksheets. The first displays your results list with all the firm characteristics that you have decided to include. Worksheet number two contains the *Screening Report*, which is an overview of the criteria that you used in your search and the Boolean operators with which you concatenated them. This worksheet provides a convenient tool to reproduce your searches later on.

However, you do not need to make use of this worksheet if you have saved your screen in the web client earlier. In any case, you can now start working with the Excel Add-In.

3 Using the Excel Add-In

3.1 Fist-time Use

After starting your session on the remote desktop, you open Excel 2016. Excel will then automatically load the Add-In for CapitalIQ functions, which allows you to retrieve data from the database.

The Add-In is displayed to you in the form of an extra header in Excel's ribbon which is named after the database. The related menu offers you a range of functions that let you interact with CapitalIQ (Figure 8).

---

[2]It is also possible to use the ISIN Code, rather than the aforementioned ID, as an identifier. This alternative has its merits since it allows for merging the obtained data with data from other sources in later steps of the analysis. However, ISIN codes are not necessarily time-invariant nor are they guaranteed to uniquely identify a company. For these reasons, you should exercise caution when choosing the ISIN approach.
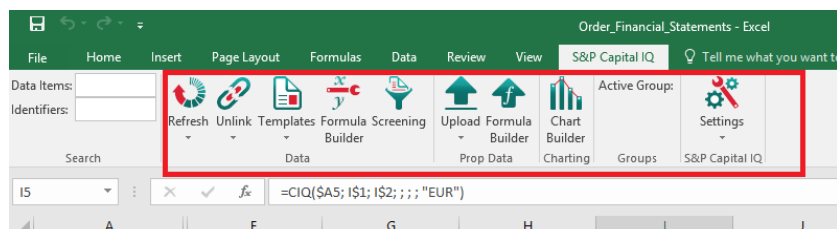
Figure 8: Active Excel functions

3.2 Retrieving Data

You begin the data retrieval process by either copying your list of identifiers from the *Company Screening Report* you exported earlier, or by downloading the saved screen from the web client. If you want to download a previously saved screen, you do so by clicking on the *Screening* button in the *Data* section of the Excel menu. This will open up a window which lets you choose from the screens saved on your account (Figure 9). By default, these screens are ordered from the most recently generated to the oldest.
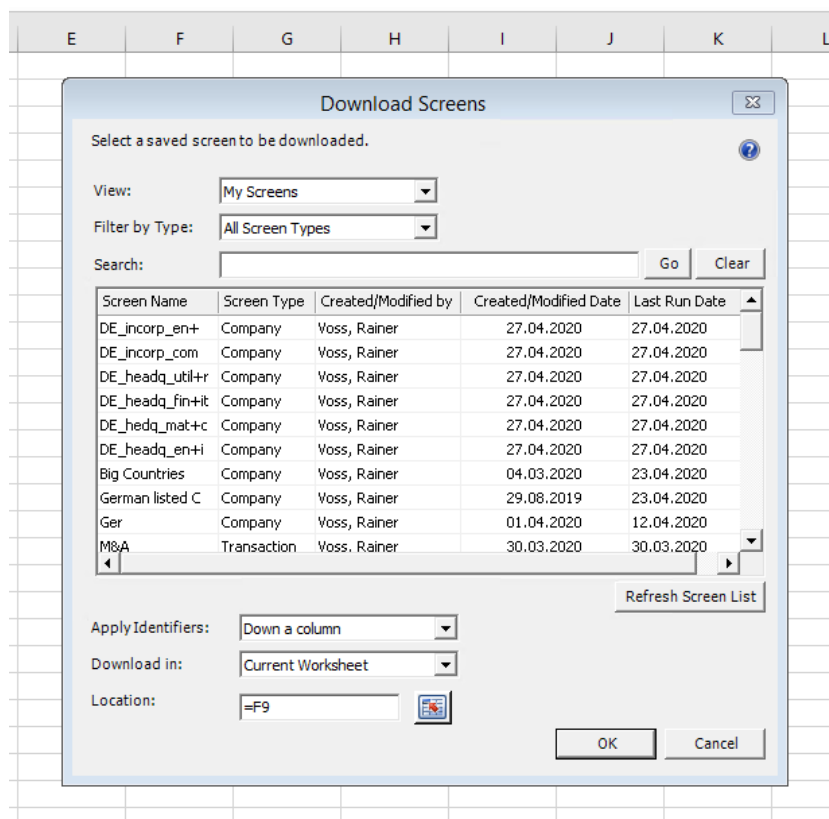

Figure 9: Saved screens

The second button that we will focus on is the *Formula Builder* (Figure 10),which can also be found in the *Data* section. With this tool, you can construct requests to the database by

generating an Excel formula. Once again, the button opens a window which has an API to the database. Here, you can search for data items and specify the properties of the requested item (No. 4 [3]). The search for data items can either be conducted through a key word search in the search bar (No. 2) or by skimming the item categories given on the left (No. 3). In the bottom half of the window, you are given a preview of the formula that you constructed (No. 5). However, you may choose not to insert this formula, but instead insert only the name of the data item you chose (No. 6). This allows for referencing the data item's name when building formulas later on. Lastly, you are asked to choose the worksheet and cell in which you would like the formula to be inserted (No. 7). By default, this will be the cell that was selected when you accessed the *Formula Builder*.

Instead of specifying the components of the request explicitly within the *Formula Builder*, you may choose to reference cells which contain the values of interest. Setting up cells to reference in the *Formula Builder* might appear to be a redundant exercise, but it actually enables you to make better use of Excel's *Autocomplete* function later on. One component for which you have to set up a cell reference is the *Identifier* (No. 1), in which you need to reference the *Excel Company ID* (or alternatively the ISIN).

At this point, absolute and relative cell references come in to play. Excel's *Autocomplete* function allows you to copy your formula to cells that are located either within the same column or within the same row as the formula. This is necessary since otherwise you would have to build a new formula for each firm and item that you want to collect data for. Setting up absolute and relative cell references ensures that your formula still refers to the correct identifiers when being dragged to other cells. You may specify the references to be absolute or relative either directly within the *Formula Builder* or you can choose to adjust the formulas in Excel afterwards.

---

[3]All numbers referenced in this subsection refer to those given in Figure 10.
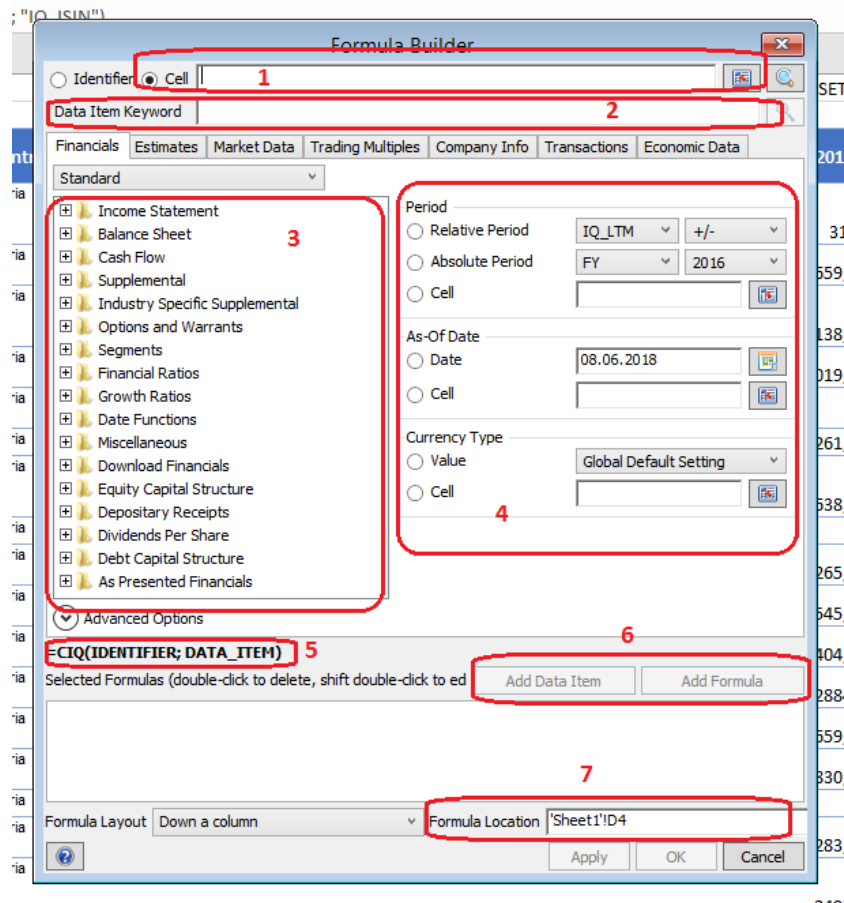
Figure 10: Formula Builder

For certain types of data, e.g., stock prices or other financial market data, it might be necessary to refresh the data at later points in time. For these cases, the Add-In offers a *Refresh* option, which spares you of having to manually retrieve the data multiple times. When using this option, you should be careful to only refresh the data for which this is necessary. Refreshing the entire workbook will take a very long time and might in some cases even cause Excel to crash. Selecting the range of cells that you want to refresh is done via the dropdown menu below the *Refresh* button.

Since you will often want to use retrieved data with stata or other programs that are not installed on the remote desktop, it is advisable to save a copy of the Excel file to your local desktop. However, since Excel will not be able to compute the functions on desktops that do not have the Add-In installed, you should only copy the **values** from the original file.

Lastly, it should be mentioned that the Add-In lets you set some global options for your data retrieval, e.g., the currency in which monetary values should be printed out (Figure 11). This can be useful for increasing the speed with which you retrieve data.
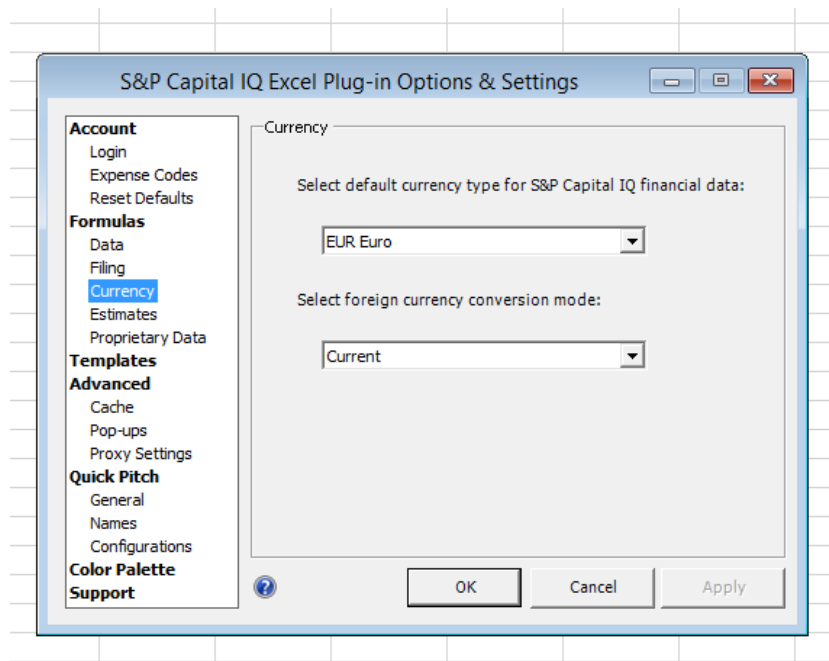
Figure 11: Add-In settings

4 Common pitfalls

While working with commercial databases has a wide range of benefits, it also brings some aspects with it on which you will need to exercise caution. I have tried to compile the problems that most frequently occur down below.

4.1 Overestimating the qualities of commercial databases

Commercial databases typically have a wide coverage and are frequently updated. Yet, for almost all subsamples, their coverage is likely incomplete and information is not guaranteed to be up-to-date. Additionally, they are also error-prone. This is evidenced by the fact that collecting the same sample from two databases will almost always yield differing values on the same variables for a subset of companies. Unfortunately, you have very few ways to actually address these limitations. Two possible approaches are the collection of data from multiple sources to increase the coverage and addressing the limitations directly in your paper, i.e., not overselling the results that your data yielded.

4.2 Misunderstanding/not knowing data definitions

You can retrieve a multitude of data items from the database but might not be able to distinguish between all of them as they are often similarly named. The first step to making sure that your data item choice is in line with your objectives is that you check whether you know what the variable should be called in theory. But even then, some item names might seem ambiguous

to you. This could either be due to imprecise terminology on the part of Standard & Poor's or to the fact that they generate multiple variants of the same basic variable. In these cases, you should check whether CapitalIQ provides a crisp definition of the data item and additionally try to replicate the numbers for a few firms in your sample by hand. Replicating numbers unfortunately does not always work since database providers sometimes adjust values gathered from financial statements according to their internal methods. Typically, these methods are not disclosed to individuals outside of the provider's firm. In the case that you cannot obtain a reasonable level of confidence that the data item meets your informational needs, it is advisable to refrain from using said item and instead search for an alternative.

A similar issue arises from industry-specific or market-specific variables. These variables might have names that are similar or identical to their more general counterparts, but are only covered for firms operating within one industry or market. Fortunately, identifying these variables is typically relatively straightforward, as retrieving them will generate a lot of missing values.

4.3   Using variables with differing currencies, units or scales

This group of issues applies not only to the use of commercial databases, but to all empirical research. The easiest way of avoiding these issues is to evaluate the nature of the values at the time of retrieval and then write down the definition of the data item including the unit, currency and scale that were used. At least for the currency, it might also be advisable to explicitly state it when building the Excel formula, so that you can later check on what you actually used.

4.4   Messing up the data set after retrieving it

This is an issue that arises in later steps of the research process. As you have already obtained an Excel dataset, it may be tempting to use this very file for the transformation of data and for conducting your analyses. However, this is an error-prone approach that might render your entire data set useless. For example, removing duplicates from your data set in an incorrect way could shift the identifiers away from the related firm data. Also, since Excel does not record a history of your changes after closing a session, you might not remember which changes you made to the file. It is best to use an external program, i.e., a statistics package like stata or RStudio, for your analyses since the code you use to conduct the analyses provides a form of documenting the changes you made.

At the very least, you should always keep a copy of your raw, untouched request data so that you can start from square one in the worst case [4].

---

[4]Furthermore, you have to make your raw data available to the university when handing in your thesis, so keeping a copy of it is necessary anyway.

4.5 Structuring data in an impractical way

It is very unlikely that your research design remains unchanged from the time you make your first request to the point when you finish your analyses. From experience, the data you need will change multiple times during your work. Common adjustments to the data set include amending the length of the time series, dropping variables or firms from the analyses or including new ones. You will of course not be able to anticipate all of these alterations, so you should structure your data set in a way that allows you to make multiple adjustments at later stages without having to dedicate too much time to it. Additionally, it can be benefitial to already think about what data structure best fits your analysis. One example of this would be to set up panel data in a wide format so that you can transpose it to long format with just one stata command later on.