How we worked on ‘The Blacklists’

88 billion Euros in 34,796 individual investments – out of which 5.63 billion Euro were placed in different controversial investments

In 2019 the Danish newspaper Dagbladet Information investigated the investments of the 17 largest Danish pension funds. The series of articles caused several of the funds to divest controversial investments.

Here you get a brief introduction to how we did it.

Journalists from Dagbladet Information collected data from the investment lists that the pension funds had published or lists that they handed over to us upon request. We compiled the lists and compared the investments with the exclusion lists that the same pension funds had published on companies that were blacklisted by the individual funds for different reasons, and which the funds publicly guaranteed that they would not invest in.

By comparing the data it was for the first time possible to map out an overall picture of the Danish pension funds equity investments and the individual funds investments in companies that one or more funds considered so controversial that they were blacklisted.

It was, and this is an important note, the 17 pension funds' own definition of controversial investments that we used.

The processing of the data was not without challenges. Firstly, the 17 pension funds do not have a shared plan for publishing this type of information. Some pension funds publish their share lists once a year. Others do it on an ongoing basis, and others still did not publish any lists at all but handed them over to us when we asked politely.

Because of these challenges there was an inherent risk that information in the compiled data – and thus in the entire dataset – would end up being outdated, since investments could have been divested since the publication of their lists.

Yet, the entire dataset was, at the time of publication of the data and the series of articles, the most accurate and up-to-date picture that it was possible to draw of the Danish pension investments.

We wanted to minimize errors as much as possible so we decided to present the combined data to all the 17 pension funds for comments before publication and adjusted the data that was pointed out as imprecise or incorrect. One of the pension funds found out that they had published an investment list that was outdated and full of factual errors. They produced an updated and correct list and we included this in our dataset.

In the following, we review the tools that Information used to collect, clean, process and analyze data from the pension companies.

The investment lists came in several different data formats. The easiest to data work with were those that were provided in spreadsheets, typically Excel files. However, it was by far the fewest we could get in that format. Most often, the investment lists were published online in pdf files.

We used the program Tabula, developed by Nerdpower, to extract the tables from the pdf files. Tabula is free and can be downloaded here: https://tabula.technology/

Tabula is used by major media organizations from all over the world in investigative journalism projects, including ProPublica and The New York Times.

With Tabula, you can extract tables from PDF files and convert them to comma-separated files that can be imported into Excel. There is an excellent introduction here:

http://schoolofdata.org/handbook/recipes/extracting-data-from-pdf-with-tabula/

The next challenge was to clean up the data. Even if the data had been extracted from the pdf files and imported into Excel and compiled into a single, unified spreadsheet, it was still completely useless because of the name standards and writing from the 17 different pension funds. The problem was that the different pension funds write names on their share investments differently. There were 17 different pension funds in the dataset thus potentially 17 different ways to state the name of the share. AP Møller-Mærsk alone is written in 10 different ways, including:

A. P. Møller - Maersk

A. P. Møller - Maersk A/S

A.P.M. - Maersk

A.P. Møller-Mæ

AP MOELLER-MAERSK A/S

AP Moller - Maersk A/S

The human eye can easily see that it is the same stock. The computer can't do that. We also had to check for human errors that might have occurred when typing in the single investments in the dataset. We wanted to pivot the data and link the individual investments to the blacklists to track the investments, but the results we would get would be useless, if we worked on uncleaned, dirty data.

Because of that it was necessary to gather the different designations for the same company under one common name. But when you have a data set with over 34,000 investment records, each potentially written in different ways, two things are needed to get a handle on the data and get it cleaned up into a usable data set: OpenRefine and knuckle-dusting. OpenRefine, like Tabula, is free and can be downloaded here: http://openrefine.org/

OpenRefine is designed exactly the task of cleaning up large datasets. It is an extremely powerful tool that can clean, transform, split and aggregate large datasets. One of

OpenRefine's functions is to find contents of cells in datasets that are similar to each other, group them into clusters and give them common names, which was exactly what we needed for our work with the investments dataset.

With OpenRefine, we could go through the entire sheet, find all places where it says A.P. Møller-Maersk in all sorts of strange ways and give them one common name and do the same with all the other different names in one and the same workflow.

There is a nice description of how to get started with OpenRefine here:

http://www.kwantu.net/blog/2016/12/28/how-to-clean-up-messy-data-using-open-refine

You can do a lot of the work automatically in OpenRefine, but you will have to check your data manually – and we strongly suggest that you present the data to sources for individual checks.

The first process was to collect and clean the entire set of data on the 17 pension funds investments. But as said before, they also publish exclusion lists – the blacklists – and we wanted to import these lists into the dataset and check for controversial investments. Some funds refuse to invest in coal and oil, others exclude tobacco, virtually all exclude companies that produce cluster weapons etc.

The process with the exclusion lists was entirely similar. The exclusion lists were retrieved from the web in the various data formats the lists were published in. Pdf files were extracted with Tabula. It was all cleaned up in OpenRefine. The lists were checked manually to weed out any remaining errors.

Again, we asked the individual funds for comments on the lists. We wanted to check their reasons for excluding investments in individual companies. As an example, the French oil company Total SA was excluded by a pension company on the grounds that the company was involved in the production of nuclear weapons. This was a mistake and the pension fund announced that it would remove Total SA from its next published exclusion list. For this reason, we decided to remove Total SA from the total exclusion list in the dataset.

It was a challenge that the individual pension funds had defined their own, and different categories in their exclusion lists. To solve this problem, we decided to define our own exclusion categories, based on the pension fund categories. We presented our categories to the pension funds to give them the opportunity to raise objections.

Finally the two lists - the entire investment list and the entire exclusion list – were combined in a single spreadsheet in Excel. Then we could apply formulas and calculate the total amount the Danish pension funds had invested in total and calculate the sum of investments in companies that were, by the pension funds own definitions – controversial.

Out of the 87 billion Euro invested by the 17 funds we found:

1 Billion Euro invested in coal activities and mining industries

0,9 billion Euro invested in oil, gas and energy companies

75 million Euro invested in tar sand extraction

And also investmens in companies blacklisted for human rights violations, weapons, tobacco, operating in occupied territories and other types of controversial investments.

Based on this final and combined spreadsheet, we could look for specific journalistic stories and begin researching and writing our stories. In our research we checked whether individual policies for blacklisting companies might have changed. A company could have been excluded in 2012 but might since have changed policy without being removed from the blacklist. We assessed that it was necessary for our articles that we checked up on the individual companies to see if we could find documentation that they were still controversial.

We then presented the research to the pension funds and asked for their comments. It could be, for example, that they had sold off the shares since they published their latest holding list. Or that they had a really good explanation for their investment, which we had to take into account.

For example, one of the pension funds investment list showed that they, as the only Danish pension company, had invested in the Indian arms company Larsen & Toubro. It initially looked like an obvious story, but the reality was that the pension company no longer owned the share. It had excluded the company since it last disclosed its shareholdings.

The data set could also be used to hunt for stories that was not based on the pension funds own exclusion lists. As an example we found the amounts the pension funds had invested in the arms giants that supplied equipment for Saudi Arabia's war in Yemen. This process was similar to the two described above. We compiled a list of relevant companies and inserted it into the spreadsheet to identify individual investments, and again – of course – we presented our findings to the pension funds for comments.