MSCI ESG Reports Scraping

In this document, we showcase the methodology used to scrape the MSCI reports. Only ESG Reports are downloaded.

Location and Access

1. The documents reside in the location - MSCI ESG Direct
2. Access to this page is required to scrape. Please contact `aj.egb@cbs.dk` for more details

Code

1. The scripts to download the reports can be found here
2. The main files are:
    a. **scrape_reports.py** - Used to download the reports using Selenium. This script should handle 95% - 98% of the cases, but some might error out due to connection errors, etc
    b. **concat_entries.py** - To record the names and details of downloaded reports into a **logger.csv** file
    c. initialize.py - Sets up required files used in **a and b**
3. The code is well commented to assist anyone taking up the task

Code execution Order

1. **initialize.py**
2. **scrape_reports.py**
3. **concat_entries.py**

Bypassing Errors

1. While looking for a certain report during execution, we may either not find that element because:
    a. **ESG Report doesn't exist**: Up until the first 1800 files (sorted alphabetically), if the ESG report doesn't exist, its Industry report has been downloaded. However, as this was more error prone and time consuming, after that, if the ESG report doesn't exist, it has been marked as "**found but report does not exist**" under the "name" column after the company's name
    b. **Download link is broken (403 Forbidden error):** Many times, the report was found, but the download link was broken. This has been marked as "**found but other error**" after the company's name in the CSV

c. **Unknown Error**: Very few times, the download was unsuccessful even though the ESG report exists. In these corner cases also, it has been marked as "**found but other error**" after the company's name and must be downloaded manually

Reporting

1. The downloaded files are stored as a zipped files in batches here
2. A **logger_final.csv** also exists that contains what Reports were downloaded and what were unavailable as well as what files were missed because of error
3. Example snippet of the document:

	report	tearsheet	industry-report	data-record-id	name
0	Yes	Yes	No	0	1&1 AG
1	No	No	No	1	1-800-FLOWERS.COM, INC. -> found but other error
2	No	No	No	2	10X Genomics Inc -> found but report does not exist

- If a report exists and has been downloaded successfully, it will have only the company name
- If a report doesn't exists, it will have suffix - "found but report does not exist"
- If a report status is unknown, it will have a suffix - "found but other error" – Only this particular status has to be downloaded manually, but there are very few of them.