PROCESSING ORBIS HISTORICAL DISK

Sebnem Kalemli-Ozcan, Jingting Fan, and Veronika Penciakova

in cooperation with Bureau van Dijk

I. PRELIMINARIES

This manual relates to the processing of the historical disk delivered by BvD. The user receives an external hard drive, which contains the following folders:

- **Financials Histo Dec SQL**
- **Financials Histo Dec text**: the contents of this folder, and the processing of files contained within are described in section III.
- **Orbis Dec text**: the contents of this folder, and the processing of files contained within are described in section II
- **Ownership histo Dec SQL**
- **Ownership histo Dec text**: the contents of this folder, and the processing of files contained within are described in IV

This manual details the processing of the files in the "text" folders since these contain files that can be read into Stata. The "SQL" folders contain the same information as those in the "text" folders, but in a different format (ie one that cannot be read into Stata).

This section describes system requirements and pre-processing steps that facilitate the processing of data contained in the historical disk. The hardware used for this task was the Windows 10 Enterprise workstation with Intel Xeon CPU E5-2680 @ @.75GhZ and 128 Gb of RAM.

1. SYSTEM REQUIREMENTS

The historical disk is around 280 GB. We recommend using an 8 TB drive to process and store the data. The guide describes how to process and store the BvD data using Stata MP, version 12 or higher.

2. FOLDER STRUCTURE

In the 8 TB drive, create the following folder structure that logically corresponds to the internal organization of the Orbis database.

1. **Firm_description**: will contain the processed descriptive firm-level data including name, address, legal form, industry classification, and national identifiers. The steps for processing these data are described in section II. Within this folder, create the following sub-folders:
    - 1.1. **Codes**: stores do-files for processing data.
    - 1.2. **Txt:** contains unzipped (unprocessed) data text files. Within this folder, create a subfolder called **chunky**, which will be referenced during the processing procedures described below.

1.3. **Dta:** contains processed data files. Within this folder, create two sub-folders, which will be needed for processing the data: **intermediate** and **final**. Within each of these folders, create a series of individual folders for each country covered in the BvD data, using the two-letter country ISO code. This country ISO code is used by BvD to create the identification numbers for companies, their owners, and subsidiaries.

2. **Financials**: will contain the processed firm-level financial data. The steps for processing these data are described in section III. Within this folder, create the same sub-folders as in folder 1 (Firm_description).
3. **Ownership**: will contain the processed firm-level ownership data. The steps for processing these data are described in section IV.
    1.1. **Codes**: stores do-files for processing data.
    1.2. **Txt:** contains unzipped (unprocessed) data text files. Within this folder, create a subfolder called **chunky**, which will be referenced during the processing procedures described below.
    1.3. **Dta:** contains processed data files. Within this folder, create two sub-folders, which will be needed for processing the data: **entity_information** and **ownership_structure**. Within each of these folders, create two sub-folders: **intermediate** and **final**. Within each of these folders, create a series of individual folders for each country covered in the BvD data, using the two-letter ISO code.
4. **Temp**: this folder is needed for running Stata directly from the 8 TB drive, which is detailed in section I.3.

3. RUNNING STATA

By default, Stata runs from the C:\ drive, which often has less capacity than other installed drives. When processing the historical disk, Stata should be run from the 8 TB storage disk.

1. Using a text editor (such as Notepad), save the following two lines of code as a *run_stata.bat* within the 8 TB drive.

```
set STATATMP=DRIVE:\FOLDER\Temp

"C:\Program Files (x86)\Stata14\StataMP-64.exe"
```

- **Note** that the first line of code references the location of the 8TB drive, and the second line of code references the location of Stata. Both lines should therefore be changed accordingly.

2. In order to run the codes for processing the historical data open Stata by double clicking on *run_stata.bat* .

II. FIRM DESCRIPTION DATA

The folder *Orbis Dec text* on the historical drive contains 32 RAR files that need to be processed. This section describes the content and details the processing of the 12 files that contain firm-level descriptive data. The remaining files contain financial information and are therefore described in

section III. Each of the 12 descriptive files discussed in this section are static, contain the latest year for which information is available, and therefore do not have a time dimension. All of these files contain the firm ID (or BVDID) in the first column and can be linked by merging on this identifier.

1. DATA CONTENTS

The following describes the contents of the 12 firm description RAR files.

1. **All addresses.txt:** contains detailed address information. The variables include the main firm identifier BVDID, the first four lines of the street address (both in English and the native language), city (both in English and native language), postcode, country, country ISO code, region in country, type of region in country, telephone and fax numbers, and address type.
    a. There may be multiple entries per BVDID because one firm can have multiple address types. The most common address type is *incorporation address*, but others include *previous address, branch address,* and *postal address*.
    b. Depending on the purpose of the study, the user can create the dataset that only contains one observation per BVDID. For this, the user can implement the following steps. Identify cases where there is more than one entry per BVDID (using the *duplicates tag* command in Stata). If a BVDID does have multiple entries:
        i. First, keep the incorporation address
        ii. Some firms do not report an incorporation address, and therefore multiple entries per BVDID need to be dealt with using different criteria. Second, among remaining multiple entries, drop the *previous address*
        iii. The remaining cases of multiple entries usually have two types of addresses: *office* and *postal*. As a last step, keep the *office* address.
2. **Contact info.txt:** this file is similar to **All addresses.txt**, but has only one entry per BVDID. That is, while the **All addresses.txt** file contains information on a firm's *previous address, branch address,* etc., the **Contact info.txt** file only contains the latest address for each firm. The file contains the firm name, first four lines of the street address (both in English and the native language), postcode, city (in English and the native language), country, country ISO, metropolitan area (for the US), state/province (in US and Canada), county (US and Canada), fax and telephone number, website, email address, region in the country and region type.
3. **Identifiers.txt:** contains various firm identifiers for each BVDID, including a national ID number, the label of that national ID, the national VAT/tax identifier, trade register number, European VAT number, LEI (legal entity identifier), and ticker symbol.
    a. This file often has multiple entries for each BVDID. One common reason is that a country has more than one type of national identifier. For example, in many countries firms are assigned both a VAT/tax identifier and a LEI. Since both are types of *national IDs*, there will be two observations per firm. One observation where the *national ID* variable is populated with the VAT/tax identifier and the other where the *national ID* variable is populated with the LEI.

    b. Since countries differ in what constitutes the main national identifier, we recommend keeping the data with multiple entries per BVDID and de-duplicating the data (obtaining a data set with only one entry per BVDID) on a case-by-case basis.
4. **Industry classifications.txt:** contains the code and text description of various industry classifications for each BVDID. Data are available for three common industry classification systems: NACE Rev. 2, NAICS 2012 and USSIC. For each of these industry classifications there are data on the core, primary and secondary industry codes.
    a. Because one company can have multiple primary and secondary industry codes, there are often numerous entries per BVDID.
    b. The text description of each industry code takes up a lot of memory and we recommend dropping those variables in the cleaning process.
    c. In case the industry file is required with only one observation per BVDID we recommend keeping only the core industry code variables using the *duplicates drop* command in Stata. These core industry codes are 4-digit for NACE Rev. 2 and NAICS and 3-digit for USSIC.
5. **Legal info.txt:** contains one observation per BVDID. The variables included are previous company name, the date of the name change, status (active, dissolved, inactive, etc.), date of the status, standardized legal form (sole proprietorship, partnership, private limited company, public limited company, branch, etc.), national legal form (varies with the country), date of incorporation, state of incorporation (for the US), type of entity (bank, financial company, industrial company, insurance company, private equity firm, etc.), BvD firm size classification, listed status, and identity of the information provider.
6. **Auditors – current.txt:** contains information on the most recent auditors of the firms identified by BVDIDs. The variables include the auditing company name, legal function, appointment and resignation dates, address, country location, and the latest audit date. In some cases, there is also information on the individual auditors, including name, gender, age, nationality, and place of birth (though most of this information is blank). There may be multiple entries per BVDID because one firm has multiple auditors, and/or the auditors serve different functions (including alternate auditor and statutory auditor).
7. **Bankers – current.txt:** contains information on firm-bank relationships. Variables include the bank name, original advisor function (usually banker), appointment date, bank address and country. There may be multiple entries per BVDID because one firm has multiple banking relationships.
8. **DMC – current only.txt:** contains detailed information on the executives and Board of Directors of firms. Information includes individuals' names, position, job title, type of position, level of responsibility, appointment date, resignation date, whether the individual is also a shareholder, gender, date of birth, age, age bracket, nationality, college, college degree, major, graduation date, and details of compensation (salary, total, date). There are often multiple entries per BVDID because the data includes information on numerous executives and board members per company.
9. **DMC – previous.txt**: contains the same variables as *DMC – current only* and covers information on the previous executives and earlier board composition. The appointment and resignation

date variables can be used to compare the board composition of firms in the **DMC-previous.txt** file to the **DMC-current only.txt** file.

10. **Other advisors – current.txt:** contains the name, function, appointment/resignation dates, address, country, and contact information of advisors employed by the firm. The kinds of advisors in the data include financial advisor, bankruptcy trustee, insurance advisor, investment advisor, law firm, public relations consultant, etc. Since one BVDID can hire various kinds of advisors, the data often contain multiple rows per BVDID.
11. **Overviews.txt:** This file contains information found in the annual reports or websites of companies. The variables covered include history, primary business line, secondary business line, main activity, secondary activity, strategy, organization and policy, and related topics.
12. **Trade description.txt:** for each BVDID, the file contains information on the type of filing (mainly annual report and local registry filing), company description (called *trade description*), a list of products/services, company class (for insurance only), specialization (for banks only), peer group name (industry), and peer group size.

2. PROCESSING DATA FILES

The non-financial data described in the previous section are primarily in string form. String variables take up a lot of memory in Stata. As a result, it is not possible to unzip and upload any of these files into a single Stata data file in the conventional workstations. Since many users will use the data to conduct research on particular countries, or subsets of countries, we break down each large data file into more manageable individual country data files. We recommend the following procedure, which proved feasible with the hardware specified at the beginning of this manual:

1. Unzip each individual file and save the text file in the *Firm_description/Txt* folder. Let's use the address.txt file as an example.
2. Process the text file using the large text file chunking utility *chunky* in Stata[1], which breaks apart a text file into smaller chunks, the maximum size of which can be specified. Save these files in the *Firm_description/Txt/chunky* folder. We recommend a maximum size of 2GB for the files above because even a 2GB text file can turn into over a 20GB Stata file if all the variables are in string format. For example, when processing the address file, the user can end up with around 20 smaller text files, each with a maximum size of 2GB. These files will contain information of a given type (firm address coming from file address.txt, for example) across a number of countries. Moreover, two or more chunks can contain parts of the same country.
3. Using *import delimited*, import these small text files into Stata format and save them in the *Firm_description/Dta/intermediate* folder. For example, after this step, the user will have 20 individual address files. At the end of this step, the user can erase the intermediate text files in *Firm_description/Txt/chunky.*

[1] This user-written command is written by David C. Elliott, Nova Scotia Department of Health, Halifax DCElliott@gmail.com

4. Clean each resulting Stata files by dropping unwanted variables. This step helps reduce the overall size (sometimes cutting the size by half) of each Stata file. See below for recommendations regarding the address file and others.
5. Break apart these Stata files into even smaller pieces by country into separate country folders within *Firm_description/Dta/intermediate*. To do so, use the first two letters of each BVDID. The user will end up with a series of country folders and each will contain a number of address files. The total number is less than 20 files, unless that country's BVDID appears in every single address file, of which there are 20. Once this step is complete, the user can erase the original data files located in *Firm_description/Dta/intermediate.*
6. Create a final appended file by stacking together (with Stata command *append*) all of the individual files of a given type for a given country and save the file into the appropriate country folder in *Firm_description/Dta/final*. Once the appended file is created, the user can erase all of the intermediate data files in the country folder of *Firm_description/Dta/intermediate*.

Below we provide recommendations on dropping unwanted variables from some of the Stata files. These files were chosen because they contain particularly useful information for researchers.

1. **All addresses:** drop the variables in the native language (street address and city), telephone and fax number, and country (the ISO code should suffice). Format remaining variables as string. This step is needed because if one of the smaller data files (chunks) contains all missing values for a variable, it will be assigned a numeric format by Stata. If the user doesn't standardize formats in each file, appending individual files will generate an error because in some files the variable will be in string format, while in others the same variable will be numeric.
2. **Contact information:** drop the variables in the native language (street address and city), country, telephone and fax numbers, website, email address and nuts. Put all remaining variables into string format for the reason discussed above.
3. **Identifiers:** drop IP identification number, IP identification label (both used to identify the information provider), other company ID number, statistical number, v13 (always missing) and ISIN number (international securities number). Put all remaining variables into string format.
4. **Legal information:** drop also known as name, reason for filing exemption, delisted comment, no recent financials flag (this will be apparent in the financial data), historical record flag, historical record since, and information provider. We chose to keep only the variables that are useful in potentially linking the BVDIDs to other data sources in the country (various national identifiers).
5. **Industry:** drop the NACE Rev 2 main section (string variable), national industry classification, primary and secondary codes in this classification (we chose to focus on the standard classification systems, NACE, NAICS and USSIC), all of the variables with text descriptions of the industries, and the BVD major sector. All of the remaining variables can be stored in numeric format. The USSIC core, primary and secondary codes are often in string form. They can be put into numeric format by implementing the following steps.
    a. Replace primary codes with missing value if the current value is "NULL"
    b. Replace secondary codes with missing value if the current value is "NULL" or "JHB"
    c. Replace core codes with missing value if the current value is "NUL"

In the procedure described above, we suggest creating individual country folders for storing country files. BvD does not provide a comprehensive list of country codes covered in the data. We recommend generating a list using the available data. This can be done by following the steps described below.

1. **Process files containing country ISO:** There are three files in the BvD historical product with a variable called *country ISO*. In our experience, each individual file does not include all possible country abbreviations encountered in Orbis Historic product. We advise the user to use these three data sets to create a data set containing all possible two-letter ISO codes.
    a. The three files are **all addresses, contact info,** and **entities**. The contents of the first two files have been described in this section. The **entities** file is located in the *Ownership histo Dec text* folder.
        i. Unzip and save each of these files into *Firm_description/Txt.*
        ii. Using the *import delimited* command, read into Stata the first column (BVDID) and *countryisocode* and save the file in *Firm_description/Dta/intermediate.*
        iii. For the **all addresses** and **contact info** keep *countryisocode*. Use the *duplicates drop* command to get a list of country ISO codes. Save the resulting files in *Firm_description/Dta/intermediate*.
        iv. For **entities**, when the *countryisocode* variable is missing, replace it with the first two letter of the BVDID (this ensure the user doesn't miss the ISO codes assigned to individuals and unknown entities). Use the *duplicates drop* command and save the resulting file in *Firm_description/Dta/intermediate.*
2. **Generate full country ISO list**: append the three file created. Use *duplicates drop*. The data file now contains the full list of countries covered in the BvD historical product. Save the file in *Firm_description/Dta/final*.
3. **Create individual country folders:** the most efficient way to create individual country folders in *Dta/intermediate* and *Dta/final* is the implement the following steps:
    a. Open the data file containing the full list of country ISOs. Use the *levelsof* command to save the list into a local macro.
    b. Loop through each element of the local macro and use the *mkdir* command to create the country folder in *Dta/countries.*

III. Financial Data

Financial data is stored in two different folders in the disk. The first place is the folder *Financials Histo Dec text*, which contains three files: "Industry - Global financials and ratios", "Industry - Global financials and ratios - USD", and "Industry - Global financials and ratios - EUR". The second place is the folder *Orbis Dec text*, which contains 32 RAR files that need to be processed. Among the 32 files, 12 are non-financial files, discussed in the previous section, and the remaining 20 are financial files.

The two folders have some overlap.

- folder *Financials Histo Dec text* contains information on industrial firms, which, by BvD definition, exclude financial companies.
- folder *Orbis Dec text* contains several pieces of information covering:
  1) financial information for banks and insurance companies;

  2) more detailed financial variables for a subset of firms (largely listed companies);

  3) a brief version of the combined data set with both industry firms, and banks and insurance companies;

  4) Files with the same names as those from the folder *Financials Histo Dec text.* According to BVD representatives, the difference is that the files under folder *Financials Histo Dec text* are available for more financial years.

1. DATA CONTENTS

This section first explains the content of the folder *Financials Histo Dec text*

1. **Industry - Global financials and ratios, Industry - Global financials and ratios –USD, and Industry - Global financials and ratios – EUR:**
   a. All three files have the same underlying data, but differ in reporting currency: the first file reports in original currency in which the companies file financial information or that is available at BvD data providers, the second in USD, and the third in EUR. In the case when EUR or USD is used, a time-varying exchange rate used to convert currency is also reported, so it is straightforward to convert USD or EUR into the original currency unit. Our understanding is the files in common currency are provided for user convenience.
   b. The sample in these three files include all "industrial firms" from the universe of countries covered by the database. Industrial firms are broadly defined to include both manufacturing and non-manufacturing firms. Entities excluded from this data set are financial firms, such as banks or insurance companies. Because these entities tend to have a different set of key financial indicators, they are reported separately.
   c. In terms of country representation, Orbis Historic product spans around 200 countries, although the coverage differs vastly across countries. European countries, for example, tend to be better represented. In terms of time representation, the sample extends until 2016, with the earliest observation in the data set dates back to the 1970's. There is again heterogeneity across countries: European countries, for example, are better covered since the mid to late 1990's; many other countries, on the other hand, do not have a significant coverage until around 2005-2007. Overall, for most countries, the sample expands over the period of 1995-2005, and becomes more or less a stable panel afterwards.
   d. The data set contains primarily financial accounting information. Variables include balance sheet items, income statement items, and some derivative financial ratios. Variables from balance sheet and income statement are presented in several levels of aggregation. For example, total assets item is decomposed into fixed and current assets. The former is

further decomposed into tangible and intangible fixed assets, but no additional details on the composition of tangible fixed assets (such as plant, property, and equipment) are provided.

The rest of this section details the contents of the 20 financial related files, which consist of only 7 independent data sets. The reason why we have 20 files for 7 independent data sets is that for 6 of the 7 data sets, three versions are provided (local currency, USD, and EUR). We now explain these 7 data sets. Since the sample expands over time, in the following, we also highlight the time when a global steady sample size is reached, without looking into country heterogeneity. It is thus important to bear in mind that for some countries, the sample coverage could be very good since the 1990's.

2. **Industry - Global financials and ratios, Industry - Global financials and ratios –USD, and Industry - Global financials and ratios – EUR**: as explained earlier, files with the same name also show up under the folder *Financials Histo Dec text*. The difference between these two sets of files is that the files under *Financials Histo Dec text* have more financial years available.
3. **Cash flow US industries, Cash flow US industries-USD, Cash flow US industries-EUR:** more detailed cash flow items for listed and delisted U.S. industrial firms, which excludes banks and insurance companies. There are about 63 thousand firm-year observations. The sample starts since the 1990's but the main body of data is between 2007 and 2015, with more than 5000 observations per year during this period. The structure of variables follows a typical cash flow statement: it starts with net income, and makes gradual adjustments for non-cash items from the net income, such as depreciation, depletion, account receivables, and eventually arrives at the net change in cash.
4. **Cash flow Non US industries, Cash flow Non US industries-USD, Cash flow Non US industries-EUR:** similar to 3, but this data set contains only non-US listed and delisted firms. In terms of time representation, there are around 40 thousand observations for each year during 2007-2015. In terms of country coverage: AU, CA, CN, GB, IN, JP, KR, TW are the countries with more than 10K observations (for the entire sample period). Most observations in this data have non-missing values.
5. **Detailed format - industries – USD, Detailed format - industries – EUR, Detailed format - industries:**
    a. More detailed balance sheet and income statement items for a small subset of global industrial firms, with industry firms defined similarly as before. This data set effectively is a more detailed version of 1, but only for listed and delisted companies.
    b. More specifically, the data has around 600 thousand firm-year observations in total. The coverage is best during 2006-2015, with the number of observations for each year ranging from 36 thousand to 52 thousand. The country representation is roughly in line with the master financial dataset (file 1 of this section). The main difference between this dataset and the main financial data set is that this one has much detailed balanced sheet statements and cash flow statements. E.g., in balance sheet, under the category of total current asset, it contains net stated inventory, raw materials, work in progress, finished goods, inventory payments, accounts receivables, doubtful accounts, etc. Under fixed assets, in addition to a breakdown of PPE, there is also leased assets. In income statement,

there are small items such as minority interest profit/loss, extraordinary items after tax, preferred dividends, etc.

6. **Banks- global financials and ratios, Banks- global financials and ratios-USD, Banks- global financials and ratios-EUR:** This data set focuses on the banks. The primary information is banks' balance sheet variables, including loans, other earning assets, derivative assets, other securities, total earning assets, fixed assets, total assets, deposit and short-term funding, total customer deposit, deposit from banks, other interest bearing liabilities, loan loss reserve, etc. In addition, the data set also contains off balance sheet items, including hybrid capital, subordinated debt. Income statement: net interest revenue, other operating income, loan loss provisions, dividend paid, etc. In terms of time coverage, the number of observations range from 14 thousand to 21 thousand during 2006-2015. In terms of country coverage: U.S. has a total of 200 thousand observations, all other countries combined have 40 thousand observations.
7. **Insurances - Global financials and ratios, Insurances - Global financials and ratios-USD, Insurances - Global financials and ratios-EUR:** this data set focuses on insurance companies. It includes some general balance sheet item, such as total asset, total liability, etc., and some insurance company specific items, such as total investment, total reinsurers' share of tech provisions, total debts, capital surplus, gross premiums written, earned premium, total underwriting income, total underwriting expenses, etc. In terms of coverage, the data set is a steady sample between 2006 and 2014, with each year having around 11 thousand observations. In terms of country representation, the U.S. accounts for about 40% of the total firm-year observations.
8. **Key financials – EUR, Key financials – USD:** this data set includes industry firm, banks, and insurance company data worldwide, with a total of 236 million firm-year observation. It focuses on a few key financial variables. The complete list of financial variable includes: operation revenue, PI before taxes, PI for the period, cash flow, total assets, shareholders' funds, current ratio, profit margin, ROE using PL before tax, ROCE using PL after tax, Solvency ratio, number of employees, market capitalization. Coverage is best from 2006 to 2015, with about 10 to 16 million observations in each year. This data set effectively combines observations from data sets 1, 6, and 7, but with fewer variables.

In all above data sets, each observation is a firm BVDID-year observation, with BVDID is the firm-level identifier in the database. There are a small number of duplicates in terms of BVDID-year. The reason is threefold.

1. Some firms report both consolidated and unconsolidated accounting reports. Focusing on only on consolidated results (there is a variable indicating consolidation status) can significantly reduce the number of duplicates.

2. Some firms change reporting month in a given year. Reports might be given for both months during the year of the change.

3. A very small number of firms file their reports through two different channels (such as through local registry or in annual reports, indicated by the variable 'filing type' in the data).It is important to note that, due to different reporting rules, the same financial variables delivered by different channels could

have different values. In addition, when the data comes from the annual reports, detailed financials are provided.

2. PROCESSING DATA FILES

This section explains briefly how the financial data sets are processed.

Depending on the size of a file, the data can either be directly processed, or processed by pieces. Since most financial files has three copies based on different currencies, we use only the USD versions as an example.

1. **Industry - Global financials and ratios - USD:** the unzipped text file is too large (around 200GB after converting to Stata data format) to be processed directly by most workstations. We recommending breaking it to smaller pieces following the general recommendations given earlier. For each of these pieces, we do the following:
   a. Read the data into Stata and convert it to dta (the Stata command chunky convert the original txt file into 10 or 20 txt files)
   b. Correct the difference between the country of a firm, and the country code indicated by the first 2 digits of BVDID, using the table "discrepancy". A detailed account of this procedure is provided in the ownership data processing section.
   c. After step b, in each small file, the country of origin for each firm in the data is known. Save these observations by their country in separate country-specific files.
   d. Process the small files one by one, and save observations for each countries.
   e. Combine all observations for a given country from different small pieces.
   The output of this process is a list of dta files, each corresponding to one country. All firms in a file are located in the same country. This way, when a data from any particular country is needed, the corresponding file can be conveniently loaded and potentially merged with other information.

2. **All other financial files**: other financial files are relatively small, so it is straightforward to load the entire file into memory, and convert it to dta.

IV. OWNERSHIP DATA

The ownership data are located in the *Ownership histo Dec text* folder on the historical disk, which contains 11 RAR files that need to be processed. The folder also contains a PDF that describes the type of entities covered and types of linkages between firms and shareholders.

1.DATA CONTENTS

This section first describes the contents of the 11 RAR files in folder *Ownership histo Dec text:*

1. **Entities.txt:** This file contains the information on the "entities" which are featured as either target (referred to as subsidiary in the data) company or the entity that owns the target (the shareholder). For all shareholders and target companies, contains the BVDID, name, country ISO code, and entity type.
    a. The full list of entity types is:
        i. Bank (B)
        ii. Financial company (F)
        iii. Insurance company (A)
        iv. Industrial company (C)
        v. Mutual & Pension Fund/Nominee/Trust/Trustee (E)
        vi. Foundation/Research Institute (J)
        vii. Public authorities (S)
        viii. One or more known individuals or families (I)
        ix. Employees/managers/directors (M)
        x. Self-ownership (H)
        xi. Private equity firm (P)
        xii. Listed (Z)
        xiii. Unnamed private shareholder, aggregated (D)
        xiv. Other unnamed shareholders, aggregated (L)
        xv. Hedge fund (Y)
        xvi. Branch (Q)
        xvii. Marine vessels (W)
    b. In 99.99% of cases this file only has one row for each BVDID. It is useful for processing the ownership data to create a file with one observation per BVDID.
        i. First, unzip and save the **Entities** text file in *Ownereship/Txt.* This file does not need to be separated into chunk and can be saved as a single data file using *import delimited*. Save the file in *Ownership/Dta*.
        ii. Open the data file and identify cases with multiple observations per BVDID using the *duplicates tag* command.
        iii. Preserve the data and in the (*preserve, restore*) step keep only the observations with multiple entries. Reshape so that each entity type appears as its own column, and denote the first of these as just *entitytype* (and the second as *entitytype2*). Save resulting data as a temp file
        iv. Restore the data and drop the observations with multiple rows per BVDID, append the small file generated in the (*preserve, restore*) block, and save the resulting data file.
2. **Links_YEAR.txt** files contain information on the links between a target firm and its owner(s) (shareholder(s)) in a given vintage year (specified by YEAR in the file name). A separate link (separate row of data) is provided for each different type of relation identified between the target firm and shareholder. The types of relations include simple shareholder, domestic ultimate owner (DUO) and global ultimate owner (GUO) – possibly following different definitions of ultimate owner. In the data, the first column contains the BVDID of the target company, the target company independence indicator (explained in the BvD PDF), shareholder

BVDID, shareholder independence indicator, the type of relation, the percentage of ownership (both direct and total), the source of the information, and the information date. Below is a description of all possible relation types between a target firm and its shareholder(s) – note that one target firm, shareholder pair can be classified as having multiple relation types.

a. Type of relation:
    i. SHH: single shareholder of first level
    ii. CTP: first level shareholder identified via the Calculated Total Percentage
    iii. ISH: shareholder at first level who is the immediate shareholder
    iv. HQ: when the target company is a branch or foreign company, its single shareholder is its headquarter
    v. DUO 25: domestic ultimate owner with a definition min 25% ownership stake
    vi. GUO 25: global ultimate owner with a definition min 25% ownership stake
    vii. DUO 50: domestic ultimate owner with a definition min 50% ownership stake
    viii. GUO 50: global ultimate owner with a definition min 50% ownership stake
    ix. DUO 50C: domestic ultimate owner with a definition min 50% ownership stake, only with owner (shareholder) types B, C, A and F
    x. GUO 50C: global ultimate owner with a definition min 50% ownership stake, only with owner (shareholder) types B, C, A, and F
    xi. GUO 25C: global ultimate owner with a definition min 25% ownership stake, only with owner (shareholder) types B, C, A, and F

Below is an example of the data for a particular target firm in the U.S. (US169497636L) in 2010. This target firm has two unique shareholders (IT00825120157 and US169497636L). A total of six links are identified between the target firm and its Italian shareholder ((IT00825120157), representing six different relation types – GUO 25, GUO 25C, GUO 50C, GUO 50, SHH, and ISH. Additionally, a total of five links are identified between the target firm and its U.S. shareholder (US169497636L, which in this example happens to be the target firm itself), again representing five different relation types – DUO 25, DUO 25C, DUOO 50, and DUO 50C. As you can see, BvD also provides the BVDID of the GUO 25, GUO 50, GUO 50C, and GUO 25C as separate variables, which is done to help users identify target firms belonging to the same global ultimate owner.

bvdid	subsidiary~r	shareholderbvdid	shareholde~r	direct	total	informatio~e	source	typeofrela~n	guo25	guo50	guo50c	guo25c
US169497636L	D	IT00825120157	A+	100.00	100.00	20101231	VD	GUO 50C	IT00825120157	IT00825120157	IT00825120157	IT00825120157
US169497636L	D	IT00825120157	A+	100.00	100.00	20101231	VD	GUO 50	IT00825120157	IT00825120157	IT00825120157	IT00825120157
US169497636L	D	IT00825120157	A+	100.00	100.00	20101231	VD	GUO 25	IT00825120157	IT00825120157	IT00825120157	IT00825120157
US169497636L	D	IT00825120157	A+	100.00	100.00	20101231	HO	SHH	IT00825120157	IT00825120157	IT00825120157	IT00825120157
US169497636L	D	IT00825120157	A+	100.00	100.00	20101231	VD	GUO 25C	IT00825120157	IT00825120157	IT00825120157	IT00825120157
US169497636L	D	IT00825120157	A+	100.00	100.00	20101231	HO	ISH	IT00825120157	IT00825120157	IT00825120157	IT00825120157
US169497636L	D	US169497636L	D	100.00	100.00	20101231	VD	DUO 25	IT00825120157	IT00825120157	IT00825120157	IT00825120157
US169497636L	D	US169497636L	D	100.00	100.00	20101231	VD	DUO 25C	IT00825120157	IT00825120157	IT00825120157	IT00825120157
US169497636L	D	US169497636L	D	100.00	100.00	20101231	VD	DUO 50C	IT00825120157	IT00825120157	IT00825120157	IT00825120157
US169497636L	D	US169497636L	D	100.00	100.00	20101231	VD	DUO 50	IT00825120157	IT00825120157	IT00825120157	IT00825120157

2. PROCESSING DATA FILES

This section details the steps for processing the ownership data.

1. **Ownership/Dta/entity_information:** this folder contains sub-folders for each country (identified by the first two digits of the BVDID). Each country folder contains an augmented *Entities* file composed of several data files.
    a. The core component of the augmented file is the original *Entities* file This file contains the BVDID name, country ISO and entity type.
    b. Merged with this is information of interest from the *contact info* country file. Some potentially important variables include *name* (from the contact info file, which can be different than the name in the Entities file), *street address, city, state, postal code*.
    c. Merged with this are the industry core codes obtained from the *industry classification* country file.
    d. Additional information can be brought in as desired from the *legal info* (original file is unique in BVDID) and *identifiers* (original file is not unique in BVDID) files.
    e. The resulting file has one observation per BVDID.
2. **Ownership/Dta/ownership_structure:** this folder contains sub-folders for each country. Each country folder contains two files.
    a. **SHARE_ISO_links_allyrs** (ISO is the country code): contains all target company – shareholder links where the *target company* is located in the country designated by the country ISO code.
    b. **SUB_ISO_links_allyrs:** contains all target company – shareholder links where the shareholder company is located in the country designated by the country ISO code.
    c. **Each of these files contains the following variables**: target and shareholder IDs, type of relation, direct and total control percentages, information date, file (identifies the *Links* file the data comes from), target company & subsidiary types, target company & subsidiary core industry codes (one each for NACE, NAICS and USSIC classifications), and target company & subsidiary ISO code
3. **The steps for creating the SUB and SHARE files are as follows:**
    a. Open the full *Entities* file and merge it on BVDID with a file that contains all BVDIDs covered in the *industry classifications* file and includes only their core industry codes. The resulting file is saved in *Ownership/Dta/entity_information,* and has one observation per BVDID. It will be referred to as *augment* in the remaining steps.
    b. For each of the *Links* RAR files:
        i. Unzip the RAR file and save the text file in *Ownership/Txt*.
        ii. Using Stata's *chunky* command, break the text file up into smaller 5 GB pieces and save these in *Ownership/Txt/chunky*.
        iii. Process each 5 GB text file and save as a Stata file in the *Ownership/Dta/intermediate* folder. At the end of this step, the user can erase the text files in *Ownership/Txt/chunky.*
    c. For each country ISO code and each year 2007 through 2016, create two blank files (SUB and SHARE) in the country sub-folder of *Ownership/Dta/ownership_structure/intermediate*. Each file contains one empty variable named *bvdid*
    d. Loop through all of the pieces of all of the *Links* files in *Ownership/Txt/chunky* and for each small file:

i. Merge in the *augment* file using the target company ID. The new variables are the target company type and its three core industry codes.
ii. Repeat the previous step, but merging the *augment* file using the shareholder ID.
iii. Using the *levelsof* Stata command, identify all of the target country ISO codes in the file.
    1. In a (*preserve, restore*) block, keep observations with the correct target country ISO code. Append the SHARE file for the appropriate year and save.
iv. Using the *levelsof* Stata command, identify all of the shareholder ISO codes in the file.
    1. In a (*preserve, restore*) block, keep observations with the correct target country ISO code. Append the SUB file for the appropriate year and save.

e. At the end of these steps you will have one SUB and one SHARE file for each *Links* file, for each country ISO code. The user can now erase the intermediate links files in *Ownership/Dta/intermediate*. The next series of steps creates one SUB and one SHARE file per country ISO code by appending the individual year files.
f. The data coverage in BvD has changed over time, and is different for each country. The process of creating a single SUB or SHARE file is the same so let's use the SUB file as an example. Loop through each country. For each country:
    i. First identify the first year where the country *Links* file has at least one non-missing observation.
    ii. Open that file and loop through all the years *t+1* through the end of the period. For each year check if the file contains at least one non-missing observation. If it does, append it.
    iii. Save the resulting fully appended SUB file in the appropriate country sub-folder of *Ownership/Dta/final*.
g. Repeat the process for the SHARE file.
h. The user can now erase all of the intermediate data files located in the country sub-folders of *Ownership/Dta/intermediate*.
i. The output are two files per country (SHARE and SUB), each containing all the relevant ownership links between 2007 and the end of the sample period (2016).