ORBIS Basic Clean Process and Overview

Sixun(Bill) Tang

October 28, 2019

Abstract

This report describes the detailed cleaning process of the raw historical data from `ftp.bvdep.com` and gives an overview of each data section contained in ORBIS.

Contents

1 Preparation Work for Cleaning ORBIS

1.1 Creating Sub Folders for Each Section

There are four file folders in the raw data, leading to natural division of the data into four parts: *Firm Description*, *Global Financial*, *Detailed Financial* and *Ownership*. In the first three sections, I create *Txt* for unzipped raw txt files, *Codes* for code and *Dta* for STATA data sets.

In the *Txt* file folder, I create another folder called *chunky*, which is used to save smaller chunks of each big raw txt file. This is important because STATA would turn a large txt file into an even larger dta file and slow down the whole process. So the strategy is to make use of the STATA command **chunky** to split and then reshape a large txt file. Note that this folder is not created manually but by code at the beginning of each code block so as to let multiple codes run simultaneously on the server.

In the *Dta* file folder, I create *Intermediate* for intermediate data and *Final* for final data. Note that in each sub-folder in *Dta*, there are sub-folders for each country, represented as the two-digit country ISO code in the ORBIS data set. The creation of these folders are described in the next section. So the extraction of data by countries could be found in respective folders.

In the section *Ownership*, I create two sub-folders *entity_information* and *ownership_structure* inside *Dta*. And the structure inside these two folders are the same as the other Dta folder in previous three sections.

1.2 Creation of Country Sub File Folders

There is no pre-compiled country ISO code list from ORBIS and according to the STATA manual, there are three files containing country ISO code as a variable: *Entities.txt* from Ownership, and *All_addresses.txt*, *Contact_info.txt* from firm descriptions. So I create a separate country ISO list from these three files and make directory using STATA inside each *Dta* folder. Note here that the first two digit of BvDID(The unique firm identifier by BvD) refers to the country and later the observations would be saved in respective country folders as a strategy to avoid large files.

Note that in these three files, the original country ISO code may differ from the first two digit of BvDID, which does not make any sense according to documentation.

2 Split Raw Data in Country Folders and Description of Each Data Set

2.1 Firm Description

The descriptive files contain the latest year for which information is available, and do not have a time dimension. All of these files contain the firm ID(BVDID) in the first column and can be linked by merging on this identifier

2.1.1 Additional_company_info.txt

This file contains additional firm information like previous company name, name change date. Main variables are:

- BvD ID number.
- Previous company name both in original language and English. Name change date.
- Also known as name in original language and English.
- Name of social networks and link to these social networks. Number of social networks account.

I dropped the link to social network accounts in this file.

2.1.2 All_addresses.txt

This txt file is super large because it contains many string characters. The variables included are:

- BvD ID number
- Address divided in four separate lines, English and native language respectively
- Postcode, City, City in native language, Country, Country ISO code
- Region in country, type of region in country
- fax and telephone number

- Type of address

To save storage, I drop anything in native language first, and then fax and telephone number are not very useful. Country is just a detailed information for CountryISOcode and also could be dropped. Note here two countries of interests are **DE** for Germany and **BR** for Brazil.

In this data set, one BvD ID could have multiple entries because there are four types of address: incorporation address, previous address, postal address and branch address. If we want unique observations for each BvD ID, we could drop previous address first. For firms which do not report incorporation address, use office instead of postal address. And always use incorporation address when it is available.

Also, in this file, I use the original CountryISOcode to split observations.

2.1.3 Auditors-current.txt

This file includes a firm's auditors' information, including auditing company name, legal function, appointment and resignation dates, address, country location, and the latest audit date. There are also individual information for auditors like gender, age and nationality. The main variables are:

- BvD ID number.
- Auditors' full name, original advisor function, appointment date/resignation date (year or exact date.), address, country, telephone number.
- Information source, providers, confirmation date and dates last received from IP.
- Auditor representatives' name, gender, birth date, age, nationality, place of birth, address and country.

Drop original advisor function in native language and telephone number, keep other initial data.

2.1.4 Bankers-current.txt

This file contains current firm-bank relationships. Main variables are:

- BvD ID number.
- Bank full name, country, address(mostly missing), telephone number and bank BvD ID when applicable(mostly missing).
- Bank appointment date(mostly missing) and original advisor function (mostly bankers).
- Information sources, information providers, confirmation dates(mostly missing) and date last received from IP.

First I drop telephone number as in previous section and also drop original advisor function because it is a duplicated variable of the same in English. Then, since most appointment dates are missing, it is impossible to construct a time measure for firm-bank relation. What we know is that such relation exists. This file is larger compared to Bankers_previous.txt.

2.1.5 Bankers_previous.txt

This file has the same structure as Bankers-current.txt and is smaller. And as mentioned above, we could not construct a time dimension for previous firm-bank relations.

2.1.6 BvD_ID_and_Name.txt

This file contains BvD ID and corresponding firm name.

2.1.7 BVD9.txt

This file contains BvD ID and BvD9 for each firm. BvD9 is a random nine digit generated by BvD as another type of identifier, but is not used widely among data sets.

2.1.8 Contact_info.txt

This file is similar to All_addresses.txt, but only has one observation per BvD ID, which means that it only contains the latest address for each firm. Main variables are:

- BvD ID number.
- Firm name in native language and English.

- First four lines of street address, city, in native language and English. Postcode
- Country, country ISO code, 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

I drop street addresses and city in native language but keep the native firm name. Also, I drop country and country ISO code after splitting the file (Note that as All_addresses, I use country ISO code to directly split the observations.). And telephone, fax, email and website address are dropped for storage purpose.

2.1.9 DMC-current_only.txt

This file contains detailed information on the executives and Board of Directors of firms. Main variables are:

- BvD ID name.
- Full name, position, job title, type of position, level of responsibility, appointment and resignation date.
- Whether the individual is also a shareholder.
- Gender, date of birth, age, age bracket, nationality, education and compensation details.

I drop job title in local language, telephone and email and information providers, since these are not important in this data set.

2.1.10 DMC-previous.txt

This file has exactly the same variables as the previous one. The good news is that compared to bankers, the appointment date and resignation date are more complete and could used to compare board change between previous data and current data.

2.1.11 Identifiers.txt

This file contains various types of national identifiers for firms, including tax identifiers, LEI(Legal entity identifier), etc. Main variables are:

- BvD ID mumber.
- National ID number, ID type and ID label.
- Trade register number, VAT tax number, European VAT number, LEI, Statistical number, Other company ID number, Ticker symbol, ISIN number. (Popular types of national ID.)
- Information providers identification number.

Firstly, I drop IP information and an empty variable called v13. Then I drop other company ID number because each BvD ID has multiple observations if there are more than one type of ID recorded. So this variable is redundant.

2.1.12 Industry_classifications.txt

This file is the largest among firm descriptions because it contains many strings of description of industries. We could drop them because the standard industry code is well known. Main variables are:

- BvD ID number.
- National(Original) industry classification, primary code, secondary code and descriptions.
- NACE core, primary and secondary code and description.
- NAICS core, primary and secondary code and description.
- USSIC core, primary and secondary code and description.
- BvD major sector.

We would like to concentrate on standard industry classification. So we only keep NACE, NAICS and USSIC code and description. Also, for storage purpose, I dropped all the description variables, and then destring all the code variables. So the final data set is BvD ID with corresponding industry codes. The core industry code is 4-digit for NACE and NAICS, 3-digit for USSIC.

2.1.13 Legal_info.txt

This file contains one observation per BVDID, and legal information of each firm. Main variables are:

- BvD ID number.
- Previous firm name and change date. Also known as name.
- Status and status date. Standardized and national legal form.
- Same or similar name in Lexis-Nexis. Date of incorporation, state of incorporation in US.
- Type of entity, reason of filing exemption, category of the company, whether delisted, listed or unlisted.
- Delisted date and comment. Main exchange. IPO date. No recent financials flag.
- Historical record flag and historical record since date. Information provider.

As suggested, I dropped Also known as name, Delisted comment, Reason for file exemption, information provider and no recent financials flag to save space.

2.1.14 Other_advisors-current.txt

This file 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. The main variables are:

- BvD ID number.
- Advisor full name, original function, appointment/resignation date(mostly missing).
- Advisor address, country, telephone number.

- Information source, providers, confirmation date and dates last received from IP.

I dropped the function of advisor in native language and telephone number.

2.2 Overviews.txt

This file contains information found in the annual reports or websites of firms, including history, primary and secondary business line, main and secondary activity, etc. Main variables are:

- BvD ID number.
- Full overview, which is a long introduction of the firm.
- History (very short), primary and secondary business line, main and secondary activity.
- Main product sand services, size estimate, strategy in organization and policy, strategic alliances.
- Membership of the firms in a network (mostly missing), main brand name, domestic country, foreign countries and regions.
- Main production sites, distribution sites, sales representation sites and customers.

Most data in this file are long strings, which may take large storage space. (E.g, a 1GB txt chunk would turn into a 7GB dta.) Considering that the full overview is where all the other variables are drawn from, I drop this longest string to save space.

2.2.1 Status_history.txt

This file contains the status of each BvD ID. Main variables are:

- BvD ID number.
- Firm status, e.g., active, inactive, dissolved, in liquidation, etc.
- Firm status date(Mostly missing) and firm status update date in ORBIS.

This file is not big so I just keep them all.

2.2.2 Stock_exchanges_and_indexes.txt

This file includes all the information on where the firms are listed and what indices are they in. Note that one firm may be listed on several exchanges. And could be included in several indices in the same exchange. The variables are:

- BvD ID.
- Name of stock exchange listed
- Name of stock index.

2.2.3 Trade_description.txt

This file contains information on the type of filing, company description, etc. Main variables are:

- BvD ID number.
- Type of filing(mainly annual report and local registry filing). Company description (Trade Description) in native language and English.
- List of products/services. Company class (Insurance company only). Specialization (Banks only).
- Peer group name(industry). Peer group description. Peer group size.
- Language ID. Description and firm history(This variable is very long.).

For convenience, I drop trade description in native language.

2.3 Financial Data-Global Section

There are 12 raw txt files in this section, which contain many financial variables of firm-financial year. There are four different files in this section, each with three copies in different currencies (original currency, USD and EUR). For each raw file, I generate a year variable from closing date for each raw text and destring the number of months, referring to the months covered by the report.

2.3.1 Industry-Global_financials_and_ratios.txt

There are three versions of this text file, in original currency, USD and EUR. For USD and EUR version, a time-varying exchange rate is added so the data could be easily modified to get the data in original currency.

The detailed variables are recorded in *List of files and variables Orbis generic.xlsx*. Note that this file contains industrial firms, including manufacturing and non-manufacturing firms. Banks and insurance firms are incorporated in separate files.

2.3.2 (Banks/Insurances)-Global_financials_and_ratios.txt

As mentioned above, these files contain an abundant collection of financial indicators of banks and insurance firms respectively, and also have three copies each. The name of the variables could also be found in the excel.

2.3.3 Key_financials.txt

This file contains industry firms, banks and insurance firms data world wide. It focuses on a few key financial variables. Main variables are:

- Operation revenue, Pl before taxes, Pl 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.

This file is a collection of the three files before, but with fewer variables.

2.4 Financial Data-Detail Section

The raw files in this section are much smaller though there are 24 files in all. As before, each file has three copies so there are 8 different files. A short description of each is listed below:

- **Banks-Global_financials_and_ratios-Interim**: This file is like the global format, but obtaining data from interim reports.

- **Cash_flow_non_US-industries**: More detailed cash flow items for listed and delisted non-U.S. industrial firms. The structure of variables is like a typical cash flow statement.

- **Cash_flow_non_US-industries-Interim**: Similar as the above file.

- **Cash_flow_US-industries**: Similar as the one for non-US firms. This one has the same structure for U.S. firms.

- **Cash_flow_US-industries-Interim**: Similar as the above file.

- **Detailed_format-industries**: More detailed balance sheet and income statement items for a small subset of global industrial firms. It is a more detailed version, but only for listed and delisted firms.

- **Detailed_format-industries-Interim**: Similar as the previous one.

- **Industry-Global_financials_and_ratios-Interim**: Similar as the global version, with less coverage and also from the interim reports.

2.5 Ownership

There are two file folders in the dta folder and each representing a sub-section in the ownership data.

2.5.1 Entities.txt

This file records all the firms as either target or owners of other firms. A BvD ID may have multiple observations since it may be in several entity types. To assure the accuracy of merging with other information, I reshape so that each entity type appears as its own column. Please refer to the STATA guide for a full list of types. Main variables are:

- BvD ID number. Firm name.

- Entity type. Country ISO code.

Note here I use the first two digits of BvD ID to split the file.

2.5.2 Links_YEAR.txt

This part is the best one with a time-dimension firm-link information. First, I generate a year variable for each file. Then, I create two files for each country-year: SUB(all subsidiaries located in one country) and SHA(all shareholders located in one country). And note that for each shareholder-subsidiary pair, there might exist several types of link and each link is identified as an observation in the data. Main variables are:

- Subsidiary BvD ID and independence indicator.
- Shareholders BvD ID and independence indicator.
- Direct percentage of control. Total percentage of control.
- Information date and source.
- Type of relation.
- Whether active or archived.
- GUO25, GUO25c, GUO50, GUO50c.
- GUO25jo, GUO25cjo(Not sure what they mean, not in the documentation and mostly missing)

A full list of link type and independence indicator is in the STATA guide and also in the excel file attached.