# When Should We Believe Research Using ORBIS Firm Data?

**Chapter** · March 2023

**1 author:**

Lukas Arndt
Max Planck Institute for the Study of Societies
**11** PUBLICATIONS **42** CITATIONS

# 1st Article:

# When Should We Believe Research Using ORBIS Firm Data?[1]

**Abstract**

Since around 2010, large off-the-shelf firm datasets covering a wide variety of variables for millions of firms are for sale by commercial providers. Such datasets are the basis for high impact scientific works, and have also beyond that been used frequently e.g. in economics, finance, political economy, economic sociology, geography, and other disciplines. One example of such an offered dataset is the ORBIS firm data, claiming to cover more than 400m firms globally in 2022. After examining and using the dataset in a four-year research project, I make the case that there are significant problems to use the data for research, depending on the question to be answered, as well as the available capacity and skills to make the data usable. Some of these are already known from the literature, others are not. But the most pressing point is that the magnitude of these problems is still understated. I gather (references to) solutions and applications to overcome a few examples of the identified problems. However, overall, I argue that the complexity of the data, existing errors, and inconsistencies, missing relevant information, the interrelation of these problems, and insufficient documentation and support provided with the data only allow for confidence in the validity of results under very specific conditions. These conditions will in most cases only apply to a much smaller sample than the claimed hundreds of millions of firms covered which undermines the main reason to work with the data in the first place. Evaluating whether conditions for valid inferences are met requires extensive knowledge of the data. ORBIS in its current state should only be used for research if sufficient time, resources, and skills are available to fully understand the data, and overcome its problems – if possible - convincingly and demonstratable. Furthermore, these issues have serious implications for the feasibility to review and evaluate research using the ORBIS data, and therefore reproducibility, unless significant additional effort is made by researchers to document the data preparation process and its implied decisions. This article gives some insights to reviewers as well as potential authors to better evaluate the use and preparation of ORBIS data. This added required effort for researchers, together with the price of getting access to the data in the first place, further speaks against its scientific use.

**Submission status:**

To be submitted as MPIfG discussion paper after another round of comments by colleagues.

[1] Many thanks to Benjamin Braun and Sebastian Kohl for very helpful comments and critiques of earlier versions of this manuscript.

# 1. Introduction

The use of large off-the-shelf datasets for research contains promises and perils (Heemskerk et al. 2018; Liu 2020). This article focusses on the perils of using the ORBIS data for research (Bureau van Dijk 2020). It is motivated by the experiences and observations I made while working with the data in a four-year Ph.D. project. It aims to raise awareness and sensitivity for the complexity and related problems of ORBIS as a frequently used big dataset for firm research. This is important because the scientific method relies on trustworthy data and critical reflection of its sources. The larger the datasets used by scientists, the more difficult it becomes to oversee the data quality and validity of the complete dataset. Depending on the size of the sample used, ORBIS is a good example of this. One purpose of this article is to share knowledge about ORBIS as a data source, to enable researchers to choose whether they should use ORBIS to answer their research question at hand. Researchers expecting yet another external (in)validation of the ORBIS data will be disappointed - but such comparisons of ORBIS' coverage with that of other sources can be found elsewhere (Bajgar et al. 2020; Garcia-Bernardo and Takes 2018; Kalemli-Ozcan et al. 2022a, 2022b). I will focus here on internal issues and observations of the ORBIS data, and my experience with it.

There are several (large) financial datasets to choose from when sketching out a new research project and its potential data sources. Liu (2020) offers a critical overview and collects common problems mentioned in the literature with some widely used financial datasets. The most extensive among the options is the ORBIS global firm database. Its claimed size of covering several hundred million firms positions it in what is sometimes called "big data". Google scholar lists 11,700 research items at least mentioning the ORBIS data, 66 of which have more than 100 citations with the top article being cited 1013 times.[1] ORBIS invites itself mostly when interested in a transnational rather than a comparative perspective, when interested in the largest firms of the globally largest economies, when interested in non-listed firms as well, and when interested in specific combinations of variables such as financial, ownership, and management data.[2] Because of its unique characteristics, the data has been used and discussed widely by research communities broadly interested in the economy. Heemskerk et al. (2018) make an important start to a conversation which I think is very necessary among the research community. Namely, the authors provide a very helpful hands-on best practice procedure for big datasets of how to tackle problems of "the five Vs": volume, velocity, variety, veracity, and variability in large firm datasets in general. They structure their overview along four questions: "Are you clear about the appropriate unit of

[1] The search was conducted on 16 March 2023. Search terms were "ORBIS AND (BvD OR Dijk)".

[2] In many other cases alternatives may be much more suitable and clean such as Compustat for globally listed firms, Refinitiv for ownership and management data in some cases, or data sources only for individual countries such as the leak of the German Handelsregister.

analysis? Is there entity ambiguity in your data? How complete are the data? How accurate are the data?" (Heemskerk et al. 2018:8)

This article draws from this work but focuses more on the pessimistic side in conversation with it. More concretely, I challenge the implicit assumption that I think could be taken from such methodological articles applied to ORBIS, that it *is* possible to make valid inferences when just following the right data preparation steps (e.g. Heemskerk et al. 2018; Kalemli-Ozcan et al. 2022a, 2022b). In other words, the question at stake is whether the respective data quality is good enough and when this may not be the case.[3] In this spirit, I take up Heemskerk et al.'s (2018:26–28) initiation to "discussions among the research community when it comes to data quality and the means of addressing it" and "a vivid, candid and critical academic debate on the merits and pitfalls of using big corporate network data". The position I would like to take in this debate is that it so far is too concerned with making data work and thereby neglects the possibility that a data source may be so problematic that, on average over all applications, it could lead to wrong, i.e., not reproducible, or not valid, scientific knowledge in too many cases. I do understand the urge in Bajgar et al. (2020:10) "not to give a definite black and white statement as to when Orbis data should be used" or the concern in Heemskerk et al. (2018:27): "we are not claiming that work that does not give an indisputable answer to the proposed set of questions should remain unpublished.". However, I think that while the problems of the dataset are spelled out on hundreds of pages by researchers already (e.g. Bajgar et al. 2020; Garcia-Bernardo and Takes 2018; Kalemli-Ozcan et al. 2022a; Kalemli-Ozcan, Fan, and Penciakova n.d.), the dangers and inaccuracies in the data are in my view still understated. I do furthermore find that our experiences with ORBIS, e.g., the way I accessed the data and its consequences, is so far not presented correctly in existing literature, and some problems (e.g., inaccuracies in the ownership data) have to my knowledge not been presented or spelled out yet.

I make the case here that under many circumstances using ORBIS is not only a question of applying the right preparation steps, diagnostics, and fixes but the data should simply not be used for research. Ironically, this is the more the case the larger the sample in terms of number of firms, but also number of countries, which seemed the main advantage of using the ORBIS data in the first place. What is more, it is relatively hard to come to this conclusion, especially at the beginning of a research project when unfamiliar with the data. This article is therefore supposed to help researchers to decide against using the ORBIS data at an early stage of the research process if necessary. On the one hand, this could simply save a lot of time and not the least money for some

[3] This implies that they may invite researchers to overestimate themselves by thinking along the lines of "If paper X says it is possible to make the data work/representative/good enough it must be possible to do it". From my own experience, I can tell that significant data problems remain and must be tolerated without fully being able to oversee their magnitude. This aspect is not stressed enough in the existing literature.

research organizations. On the other hand, and in the worst case, investing a lot of time, effort, and money into preparing datasets as complex as ORBIS could be an incentive to publish results despite not being sure enough about their validity, i.e., bad scientific practice. This is the more dangerous in the case of ORBIS because obstacles are also exceptionally high for reviewers to evaluate data preparation and consequently results appropriately. The longer I worked with ORBIS, the surer I was that hardly any reviewer has the experience with the dataset to truly evaluate whether the sample I use, the data preparation steps, and the analyses I applied led to valid results or not. This realization can be dangerous for good scientific practice, even unconsciously.

I spell out the problems responsible for this alongside the structure suggested by Heemskerk et al. (2018). However, upfront I would like to pay attention to one important factor namely that of necessary IT skills to perform many of the necessary preparation steps and diagnostics. This is a point where I disagree with Heemskerk et al. (2018:7), who consider "it as a misperception that the integration between big data and social science is about technical capacities. Certainly, within the context of BCND, the volume is larger than before but manageable with current tools and techniques". I do have a degree in business and IT and consider myself interested in, and to some degree capable of, coding in Python in general, and data preparation specifically. Yet, I was not successful in applying several of the so far available suggested or necessary data preparation steps, at least not with reasonable time and hardware resources, to fully account for some problems in ORBIS. It is therefore important to realize that working with the ORBIS data and making it usable requires deeper knowledge of preparation of large datasets in some programming language, as well as efficient coding in terms of working memory and computation time. It may of course also be possible to hire data engineers or other people with training in this area, but this will likely not be affordable to most research projects especially given the current huge demand and low offer of these skills in the general labor market. [4] I do however speak more about the case where trained social scientists, even techy economists, attempt to make use of the data. It will in many cases not be feasible to use it as a single researcher, but a larger team of IT skilled and determined researchers or data scientists working closely together might be more promising.

Since I do not speak about big datasets in general but about ORBIS specifically, it is necessary to position this article a bit more in the ORBIS-specific existing methodological literature. Table 1 gives a succinct overview of existing literature dealing with ORBIS and how to

[4] In the business world, the tasks required to understand and clean messy big datasets are fulfilled by so called data engineers which comprises a whole fast growing and evolving field in computer science. However, this is a skill set that is to my knowledge not taught in any social science program and which requires months and years of training or experience in complex data engineering tasks. Tools commonly used in businesses to solve data problems of this magnitude are not familiar to social scientists at all.

position this contribution within it. In terms of access to the data, one relevant restriction of this article needs to be disclosed. There are several ways to access the ORBIS data which vary by the effort one needs to put into receiving the data, and not least by price. One main difference between these access options is the question of representative longitudinal vs. cross-sectional data. This is also one relevant axis along which to structure existing literature on the use of ORBIS for research, since the most extensive works are interested specifically in historical data, and many of the preparation steps only relate to the question of harmonization of different cross-sectional snapshots from ORBIS (Bajgar et al. 2020; Kalemli-Ozcan et al. 2022a). In addition to that, it should be noted that these works are mainly interested in aggregate outcomes of firm-level microdata such as market concentration, only focus on selected sectors, and mostly on European countries. Whereas my research interests in terms of time points were rather, first, to classify individual firms as either controlled by a super-rich family or not at a single point in time, to use this indicator as an independent variable to predict an outcome matched from an external data source (firm party contributions and lobbying) in a specific year. Second, to map firm networks based on the ORBIS management data in one year. Most contributions dealing with ORBIS have a specific research interest in mind and in all cases not all discovered and solved problems are relevant for all potential research projects one can pursue with ORBIS data.

We[5] accessed the ORBIS data in the following way. My institution had a paid contract with the data vendor Bureau van Dijk which allowed me to manually (only assisted by self-implemented web scraping) download the data through the ORBIS portal. This is the more affordable access option which also has some downsides in terms of data quality especially for longitudinal data, as will be spelled out in more detail in the next section.[6] It may be the case that some of the problems described here are only applicable to this download method (manual download through web scraping of the export function in the ORBIS portal). Researchers interested in using the ORBIS data should note that there are also other hands-on manuals for the ORBIS historical disks and the new "ORBIS historical product", which only relate to another method of access which is a lot more expensive to procure (Kalemli-Ozcan et al. 2022a, 2022b, n.d.). However, depending on the research questions of interest this access option is sufficient, especially when focusing on a cross-section of the most current ownership and management data – or 1-2 years before the download date for financial data to account for the reporting lag. Differences between the access options are further discussed below in section 3, and in Kalemli-Oczan et al. (2022b:2–3).

[5] More employees from my institution were involved in the communication with BvD and the selection of variables that could be relevant to us but I was responsible for the download and the only researcher really working with the downloaded sample so most of the experiences presented here are from my individual perspective.

[6] However, from our impression the products and pricing offered by BvD are not set to stone or even published somewhere but seem to vary case by case.

| | **Bajgar et al. (2020)** | **Garcia-Bernardo and Takes (2018)** | **Heemskerk et al. (2018)** | **Kalemli-Oczan et al. (2022)** | **This work (2023)** |
|---|---|---|---|---|---|
| **Focus on ORBIS as the main data source?** | Yes | Yes | No | Yes | Yes |
| **Focus on longitudinal use of the data?** | Yes | No | No | Yes | No |
| **Access method** | 2017 historical disk (Bajgar et al. 2020:8) | "creating a snapshot of the Orbis database in November 2015" (Garcia-Bernardo and Takes 2018:166) | Not specified/ only using ORBIS as an example | Multiple historical disks from ORBIS and Amadeus (Kalemli-Ozcan et al. 2022b:9) and "historical product" (Kalemli-Ozcan et al. n.d.) | Two downloads (2020 and 2022) from the ORBIS web portal |
| **Research interest** | "applications of Orbis to the study of productivity and business dynamism." (Bajgar et al. 2020:9) | "corporate networks, in which links represent particular relationships between corporations" (Garcia-Bernardo and Takes 2018:165) | "big corporate networks" (Heemskerk et al. 2018:6) (e.g. interlocking directorates but also others) | Mainly "macroeconomic outcomes" (Kalemli-Ozcan et al. 2022a:1), SMEs (Kalemli-Ozcan et al. 2022a:9–12), and "industry concentration" (Kalemli-Ozcan et al. 2022a:12–19) | Identifying super-rich individuals and families among shareholders and managers; describing and analyzing their firms and firm behavior |

**Table 1:** Literature on ORBIS data quality and relevant differences.

To evaluate the claims made in this article, it may be useful to know more about how and with what purpose I approached the ORBIS data. I consider myself a computational social scientist who used the ORBIS data for three purposes: first, identifying super-rich (i.e., listed on public rich lists) individuals and their firms among the ORBIS shareholder data. In this step, I matched an external list of strings containing individual, family, and firm names to the ORBIS shareholder and manager names. Through this, I gathered in-depth knowledge of the ownership data and its problems. Second, I used other variables such as listing information, financial data, and management data as estimators in regression analysis. This way I gained some familiarity with missingness, errors and inconsistencies, entity ambiguity, and other problems with the ORBIS data in general. Finally, third, I worked with management data to analyze networks of interlocking directorates. This way I gained insights into problems with the management data, and especially its validity over time or duplicates. This article is written based on these experiences and perspectives.

One takeaway from working with ORBIS for four years is that the longer one works with the data and gets familiar with it, the less one trusts in it. Even after four years I constantly find new unexpected, incomplete, very much unintuitive, or simply erroneous data in our downloads. This led me to ask the question that drives this article and is motivated by concerns of good scientific practice: *Under what circumstances can we trust results based on the ORBIS firm data?* As a final disclaimer, I would like to stress that I published research results based on ORBIS myself. I did this because I am as confident as one could be in my results by examining them and testing them for robustness myself to a feasible extent. At the same time, I do strongly believe in peer-review and know there are probably only a few other researchers in academia with enough knowledge about ORBIS who could review my work, as well as any work using ORBIS data, to the extent it would be adequate. In the current academic system where the time available and acknowledgment for in-depth reviewing (not to speak of code review) are very restricted, the complexity of the data is a problem for the reviewing process, and also for my own research and published results of course as much as for any other research project working with the data. That is to say, I do not want to give the false impression of authority to have fully understood the fallacies of ORBIS. In the following, I begin by briefly presenting basic information about the ORBIS database in the next section. Next, I present a (likely incomplete) list of problems and shortcomings of the data, together with more and less successful solutions to some of the problems offered in the literature, or that I have tried myself. Finally, I offer some summarizing thoughts and additional reflections on the question of when we should believe research using the ORBIS firm data.

## 2. Background on the full ORBIS data and our used samples

The ORBIS database offered by Bureau van Dijk (BvD), founded in 1970 and since 2017 a subsidiary of Moody's, is the largest global firm database comprising data on firm financials such as revenue, EBITDA, etc. In addition, it includes ownership data including shares and shareholder information, and management data such as names of management. There is also other data included in the database that is even more sensitive, such as credit rating scores of individual companies, addresses of owners, etc.[7] The data was brought to the market by BvD already in 1987 as a CD-ROM (Group van Dijk 2013).

In January 2023, BvD claims ORBIS to cover data on almost 447m firms globally. Table 2 shows the distribution of coverage by region and available financial data. 427,399,056 of these companies are active in January 2023, 84,501 of them are listed companies, and only 4,928,754 companies are not subsidiaries, i.e., not owned by another company with 50% or more. A related issue is that companies may publish consolidated accounts for a whole corporate group, unconsolidated accounts only for the respective entity in the firm hierarchy, or both types of accounts. This fact and related problems are very relevant when interested in financial accounts and their aggregation to the market or country level and they are laid out together with potential solutions elsewhere (Bajgar et al. 2020:50–52; Kalemli-Ozcan et al. 2022a:78–81). However, related issues are also discussed below in the unit of analysis section. Bajgar et al. (2020) also present and discuss coverage by different account types. Kalemli-Oczan et al. (2022a) gather the information providers and legal regulations for European Countries in detail.

[7] According to the talks we had with representatives of the company, the only hard restriction to publish excerpts from the database, e.g., for review purposes, were the credit rating scores. We were allowed to publish everything else for scientific purposes in principle.

| World regions and countries | Companies with detailed financials | Companies with limited financials | Companies with no recent financials | Companies without financials | Total |
|---|---|---|---|---|---|
| **Total** | 25,657,896 | 119,041,365 | 73,526,348 | 228,674,634 | 446,900,243 |
| **North America** | 33,288 | 35,497,401 | 6,130,192 | 29,877,514 | 71,538,395 |
| **Western Europe** | 12,231,563 | 10,833,281 | 18,710,273 | 45,630,171 | 87,405,288 |
| **Eastern Europe** | 7,251,973 | 2,920,536 | 12,473,779 | 22,277,876 | 44,924,164 |
| **Middle East** | 3,220 | 1,681,911 | 1,047,266 | 4,689,310 | 7,421,707 |
| **Far East and Central Asia** | 3,538,960 | 21,193,207 | 28,092,054 | 64,443,982 | 117,268,203 |
| **South and Central America** | 2,352,097 | 44,122,988 | 2,274,382 | 13,630,177 | 62,379,644 |
| **Africa** | 219,200 | 2,782,364 | 521,115 | 17,044,118 | 20,566,797 |
| **Oceania** | 27,444 | 9,466 | 4,274,641 | 29,064,748 | 33,376,299 |
| **Supranational** | 48 | - | - | 16 | 64 |
| **No country specified** | 103 | 211 | 2,646 | 2,016,722 | 2,019,682 |

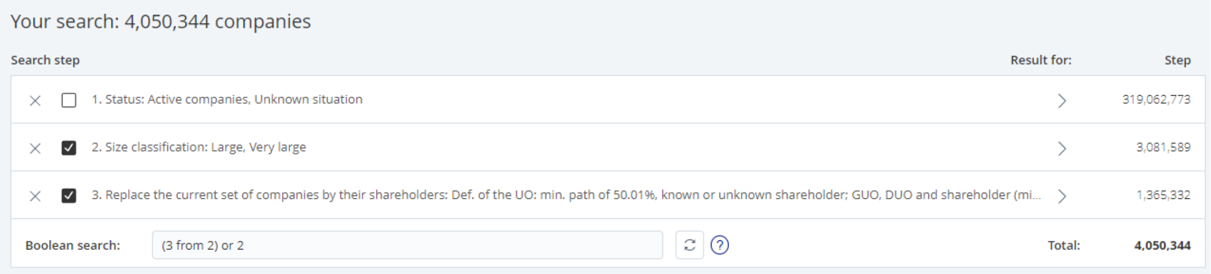
**Table 2:** Firms and financial data covered by region taken from the ORBIS portal as of 16 January 2023.

The thoughts and insights presented in this paper stem from the analysis of two samples (cf. Figure 1). The first was downloaded from the ORBIS portal by web scraping in 2020 and comprised the roughly 700,000 largest global firms in terms of operating revenue and the number of employees, as well as all their subsidiaries, and all their shareholders. In total, the sample comprised 3.2m firms. This sample was then extended by several additional downloads in 2022 tailored to our research interests, the largest of which included the 3m largest firms designated by ORBIS and their shareholders which summed up to a total of 4m firms. I also downloaded other samples such as all German firms which were not yet included in the sample. More precisely, I implemented a web scraper that clicked through the ORBIS portal in an automated manner with Selenium (2023) in Python. The script automatically logged into the ORBIS web portal, loaded a pre-defined search and a pre-defined selection of variables, and exported the results through the export function in chunks of around 80% of the maximum amount possible. This maximum amount depended on the number of variables chosen and was read in automatically by the scraper for every combination of variables I downloaded. Overall, I downloaded more than one hundred variables from the financials, ownership, management, industry, and stock sections for several million firms.

a) The first sample, which was downloaded in the first half of 2020 (own summary)

| Step | Selection | Years | Number of firms |
|---|---|---|---|
| | **Sample 1** | | |
| 1. | The largest firms in terms of operating revenue with a minimum of US$ 50m | Any of the years 2010-2018 | 385,028 |
| 2. | The largest firms in terms of number of employees with a minimum of 200 | Any of the years 2010-2018 | 425,605 |
| 3. | All shareholders of all firms from selection steps 1. and 2. included in the database | Any of the years 2010-2018 | 796,351 |
| 4. | All subsidiaries of any of the firms from selection steps 1. and 2. included in the database | Any of the years 2010-2018 | 2,406,289 |
| | | **Sample 1 total excl. overlaps:** | **3,207,095** |

b) Summary of the second sample downloaded in the first half of 2022 (screenshot from the ORBIS portal search)



Your search: 4,050,344 companies

| | | Search step | Result for: | Step |
|---|---|---|---|---|
| × | □ | 1. Status: Active companies, Unknown situation | > | 319,062,773 |
| × | ☑ | 2. Size classification: Large, Very large | > | 3,081,589 |
| × | ☑ | 3. Replace the current set of companies by their shareholders: Def. of the UO: min. path of 50.01%, known or unknown shareholder; GUO, DUO and shareholder (mi... | > | 1,365,332 |

Boolean search: (3 from 2) or 2 — Total: **4,050,344**

**Figure 1:** Summaries of samples downloaded from the ORBIS web portal a) in 2020 and b) in 2022.

## 3. Data access, support, and documentation

In this section, I will briefly report on the interaction our research institution had with the ORBIS data provider BvD, the different ways of data access, experiences with the documentation of the ORBIS data, and the support offered by BvD post-sale. We contacted BvD for the first time in 2019, to procure the first sample for my Ph.D. project. Two ways of access were offered to us. First, yearly snapshots of the full database since its creation (historical "disks"). Second, we got access to the portal and unlimited downloading and could download from the portal whatever we needed during an agreed period of a few months. According to BvD representatives, legal changes shortly before our request led to a steep increase in the pricing of the historical disks to a low six-figure amount. The second option could be offered much more affordable for an amount in the lower five-figure range. We chose the second option since the first was not affordable to us, as will assumably be the case for most except the best-funded research institutions in the world. We were made aware that full historical data would only be available on the historical disks but other than that there would be no limit. We could download as much as we managed to within the agreed period of around six months, and from the technical side only what is downloadable with the export function in the ORBIS web portal.

The way I accessed the data from the portal is claimed to lead to lower quality data, especially by Kalamli-Ozcan et al. (2022a, e.g. the footnote on p. 7). The main concern of the authors is the longitudinal dimension of the data. While it is true that the data from the web portal has several downsides towards the historical disks, most notably the shorter period of historical data included, I cannot confirm two of the most serious downsides of this access method presented by the authors. First, I could download historical data from the portal, e.g. operating revenue of the past 10 years, without systematically missing values ("download cap": e.g. Kalemli-Ozcan et al. 2022b:4). I am confident that there was no such cap in place for two reasons. First, no such download cap was mentioned by BvD sales representatives as our contract allowed us to download full information on millions of firms directly from the portal.[8] Second, I checked several cases by hand and each time the missing data was also missing in the web portal. Unless there is a distinct dataset distributed by BvD that has more complete information than the web portal, I think that our sample should show the same degree of missingness as the historical disks.[9] I encountered one case where I wanted to download historical ownership data and values existing in the web portal were capped during the process of our second download in 2022. However, after consultation with

[8] It is possible of course that our contacts were not aware of such a cap but in that case, I am pretty sure our organization would have legal claims against BvD because it would mean a breach of our contract.

[9] But this needs to be verified by comparing the two versions. I provide insights into our missingness structure that can be used for comparison purposes in section 4.

our BvD representative and his internal conversations with the data team, the download then worked as expected for all cases, and it was reconfirmed to us that there should not be a cap.[10] A second point I cannot confirm which is raised in Kalamli-Ozcan et al. (2022a:6) is that downloading historical ownership information was not possible for a large number of firms from the web portal. I downloaded historical ownership data successfully even for longer periods than 10 years by simply choosing direct and total shares at the respective past date as columns in the view search results view, and then exporting the output. However, indeed, the extent of the survivorship bias when retrieving historical ownership information in this way remains unclear. While we have no external data source to check attrition systematically, our download includes more than 100,000 shareholders that only held a share in some company in 2010 and no share after. Shareholders, therefore, do not seem to be excluded systematically even if there is no current share held by them in the data for the past 10 years. Disregarding these two points where I have to object to Kalamli-Ozcan et al. (2022a), it seems plausible that, as the authors report, firms are not included e.g. if they were inactive for five years or more. I can also confirm there is a lag in reporting of about 1 to 2 years on average which depends on attributes such as country, type of firm, size of the firm, etc.

The problem of available historical data in the web portal seems to vary significantly by the kind of variables used and downloaded. The non-availability of historical information in the portal is a lot more pressing for the management data. Figure 2 shows the number of management positions by year (how many managers were in office in the respective year) when using historical management data (dates of appointment and end dates) from the web portal. This data is the basis of the vast research on interlocking directorates. I already applied several steps of imputation in case of missing begin or end dates of the management appointment. A historical analysis of management positions seems impossible based on data from this download, most likely because historical management positions are not included in the portal anymore – or only in comparably few cases. The development mostly reflects which of the current directors were already in office in the respective years. The selection criteria for which historical positions are kept, and which ones are dropped are unknown to me and are also not included in the documentation. It also remains unclear whether the data is more complete in the historical vintages, and it would be worthwhile for researchers with access to the historical vintages of the management data could test and report this.

[10] This anecdote does however suggest that a download cap might depend on the exact agreement with BvD.

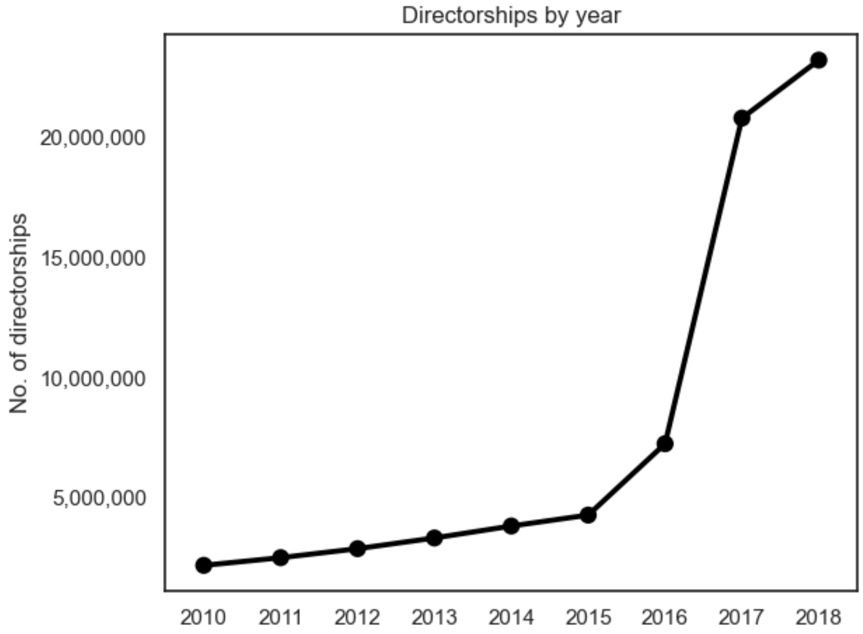


**Figure 2:** How many managers are in office according to the sample? The number of management positions by year in the ORBIS data from our first download in 2020. The curve suggests either attrition of management positions or a strongly increasing sample of firms for which data is available.

Changing BvD numbers over time is another shortcoming raised by Kalamli-Ozcan et al. (2022a). This may for example be the case if companies changed addresses or for other legal reasons depending on the country (Kalemli-Ozcan et al. 2022b:5). I would like to add to this that the problem needs to be segregated into two distinct problems: changing BvD IDs of companies and changing BvD IDs of individuals as shareholders. As concerns the former, we were made aware by a colleague from another institution that this can be overcome by an integrated tool in the ORBIS portal through which a list of company BvD numbers can be uploaded, and which then provides a list of old and new BvD IDs as a result which can be downloaded.[11] However, in our research when combining two downloads of ownership information, the most pressing problem was that of changing BvD ID numbers of shareholders which, in combination with our extrapolation of ownership shares, leads to a legacy of duplicate shareholders with old and new BvD IDs. Often, these are individuals or families as shareholders. One cannot simply search and download changed BvD IDs for individuals in the ORBIS portal, and therefore the only solution for this problem we see is to receive a correspondence table of changed BvD IDs of individual shareholders from a BvD representative. But this piece of information and how we retrieved it brings us to our experience with the ORBIS documentation and support. For a lot of questions, it is not possible to find information in the documentation.

[11] To find the tool just choose to search for companies by BvD ID numbers in the portal and upload e.g. a CSV file with BvD numbers.

A fundamental problem with the BvD support that I faced during our download phase and figuring out what exactly I needed was that the sales team does not seem to be closely connected to the data team. In summary, I was not successful in receiving support or additional information from the BvD Orbis support – or at least not the correct information. I was especially persistent in asking about two things. One is the question of whether it is possible to download old BvD IDs or to receive a translation table from BvD with old and new BvD IDs. Kalamli-Ozcan et al. (2022b:5) report that it should be possible to retrieve such a table form a BvD representative. The second request I had was to ask whether there was any way of acquiring a table that assigns Shareholder BvD ID numbers to unique contact identifiers (UCIs) of managers. Imagine a use case of looking for an individual majority shareholder's management board membership. Say, which corporate boards have super-rich individual shareholders such as Jeff Bezos as members. I asked our sales representatives, who claimed to have asked the data team, numerous times how to acquire such a table. I was sure that some sort of database or translation table must exist because when clicking through shareholders manually in the portal, the respective UCIs are shown together with an individual's shareholder BvD ID. In both cases, I was not successful in getting help, and most likely not even successful to make the other side understand what exactly I need. In both cases, I also did get the wrong information as a response. In the first case of changing BvD IDs, the correct or helpful answer would have been to point us to the tool integrated into the ORBIS portal. Instead, I only got the information that the old BvD ID cannot be seen in the portal. In the second case, I received the answer after weeks of continuous inquiry that it is possible to show the BvD ID number of shareholders of a company together with the UCIs of its managers. This did of course not answer our question and is not helpful. I gave up after this answer to my seventh email.

The point of these anecdotes is that I did not receive any helpful information from our contact with BvD support post-sale. I heard about other research teams working more closely together with different departments of BvD and making better experiences. Researchers thinking about working with the data should at least be aware of the possibility that no further support is offered with the sold data. Considering the numerous data problems, inconsistencies, and unknown sources and definitions of the data, this is a big problem. Researchers have to rely on the few manuals and guidelines published by other researchers cited here, and on asking each other for help and clarification. I can only speculate from our experiences, and what I heard from other experiences, that likely our financial weight by procuring the cheaper access option was not large enough to secure support with the data. The purpose of the web portal' documentation from my impression seems to be to enable customers in the business to use the basic functions of the portal. A use case might be to do some sort of small-scale analysis of businesses or competitors in their industry or area of interest. It is mostly about how to use the different functions of the ORBIS portal. Only a small share of the documentation (certainly not sufficient for research purposes) is about the data, its sources, and its definitions, which is of course relevant and necessary information when using the data for public research. Questions such as the one about changing BvD IDs or how exactly variables are defined can be found in a few cases but not systematically. For this reason, most of the background information about the variables, how they are defined and generated, and any additional information that is not included in the data needs to be requested from the BvD ORBIS support. The next section presents the problems of the data that I discovered while working with the data.

## 4. Data problems and some questions to ask

As mentioned before, criteria of data quality of large firm datasets as ORBIS, e.g. representativity (Bajgar et al. 2020; Kalemli-Ozcan et al. 2022a) and the "five Vs" (Heemskerk et al. 2018), were already suggested in the literature. Figure 3 shows the best practice procedure developed by Heemskerk et al. (2018). Following this procedure, I discuss known and new problems of application in ORBIS for the four mentioned problem classes: unit of analysis, entity ambiguity, completeness, and accuracy.

### Unit of analysis

The scientific applications to which this article speaks use ORBIS because they are interested in firms as the unit of analysis. However, ORBIS is the best case I know to illustrate that the "important ontological question: what is a firm?" (Heemskerk et al. 2018:11) is a very complex one. On the one hand, there are corporate groups including branches in different countries and possibly multiple other subsidiaries. These subsidiaries can either be understood as active parts of a company that differentiates into different activities possibly in different countries to exploit beneficial regulations and international competition for foreign direct investment (Reurink and Garcia-Bernardo 2020). They could also be solely administrative shell companies that only exist in a specific jurisdiction for tax purposes or other reasons, but do not produce anything (Garcia-Bernardo et al. 2017; Heemskerk and Takes 2016b:98). Which of these companies is *the* company?

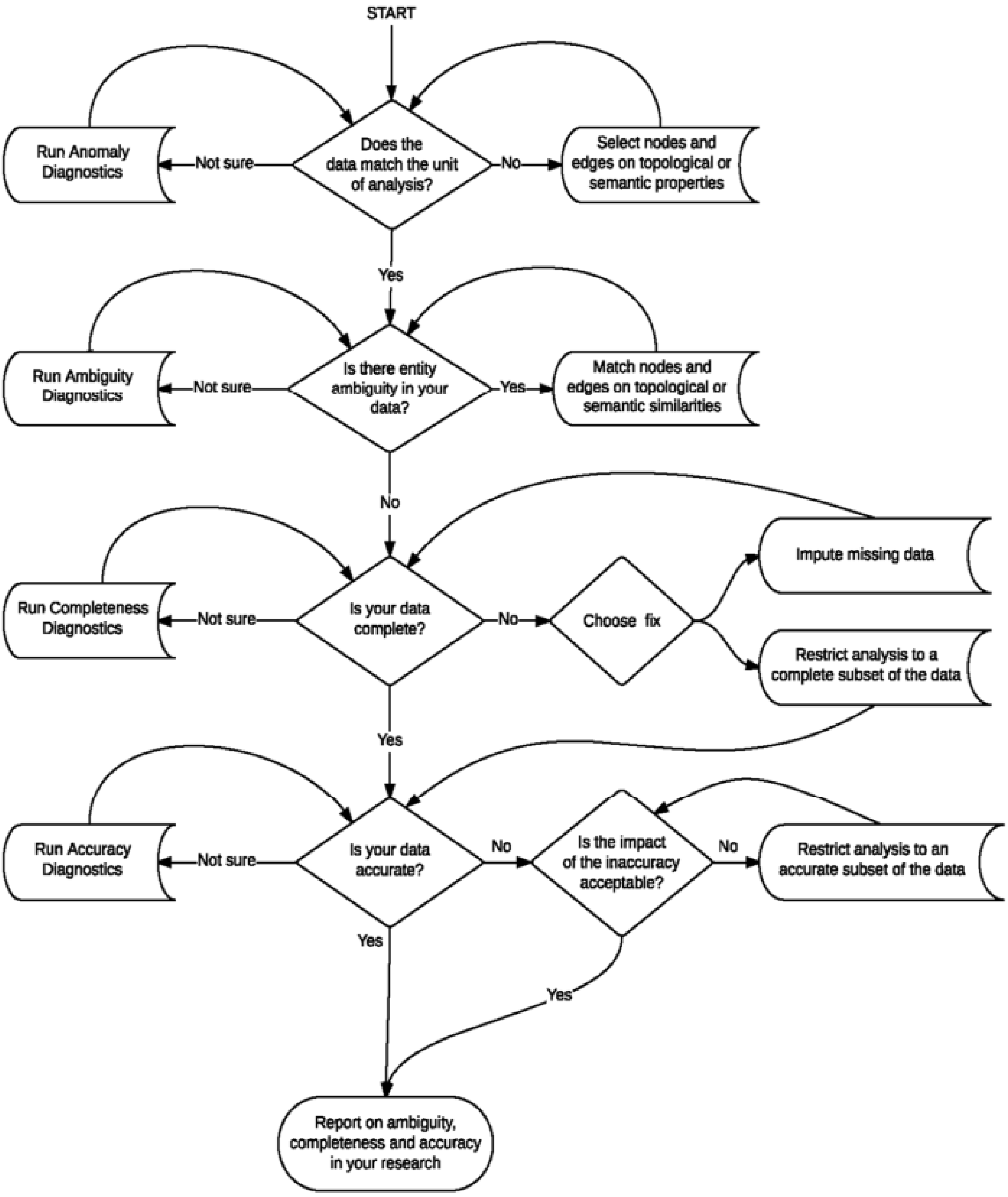
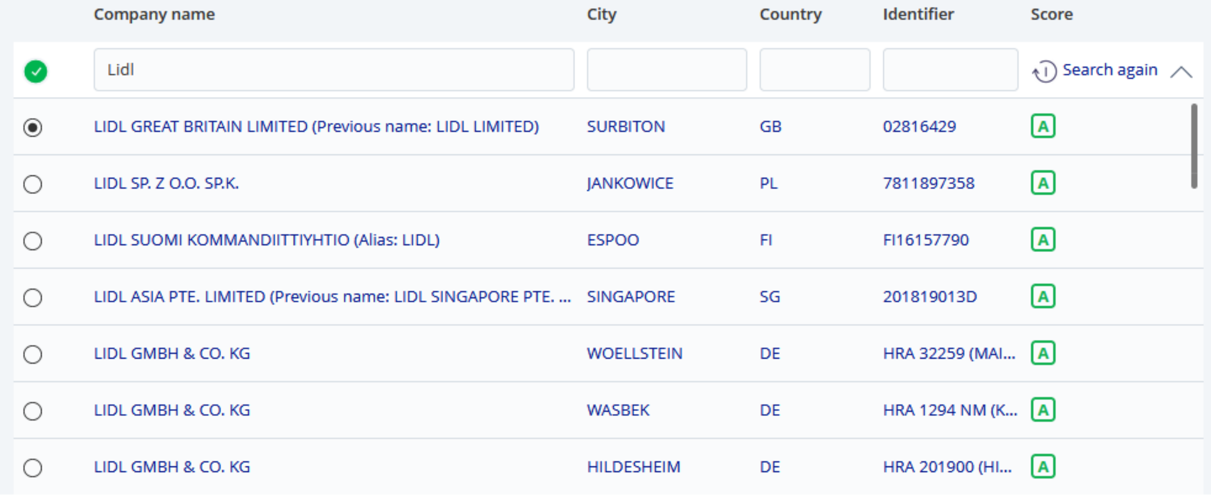

START
Does the data match the unit of analysis?
Not sure
Run Anomaly Diagnostics
No
Select nodes and edges on topological or semantic properties
Yes
Is there entity ambiguity in your data?
Not sure
Run Ambiguity Diagnostics
Yes
Match nodes and edges on topological or semantic similarities
No
Is your data complete?
Not sure
Run Completeness Diagnostics
No
Choose fix
Impute missing data
Restrict analysis to a complete subset of the data
Yes
Is your data accurate?
Not sure
Run Accuracy Diagnostics
No
Is the impact of the inaccuracy acceptable?
No
Restrict analysis to an accurate subset of the data
Yes
Yes
Report on ambiguity, completeness and accuracy in your research

**Figure 3:** How should big corporate network data be prepared for research? Decision tree from Heemskerk et al. (2018:10).

The larger the total assets of a firm, the more subsidiaries it has on average (Heemskerk et al. 2018:17). It might be intuitive trying to solve this problem by keeping only the head of a corporate group. Taking this fix may– depending on the research question - exclude firms of interest or that should be included to produce generalizable results. For example, Heemskerk and Takes (2016a) discuss that while only keeping the head of a corporate group will reduce administrative ties and shell companies which may not be relevant for several questions, they also exclude more relevant parts of the holding structures. The authors give a few examples. One is that excluding all subsidiaries in an ORBIS sample would exclude Volkswagen AG since it is a direct subsidiary of Porsche Holding. Depending on the research question this makes a relevant difference. In my research analyzing control over firms, I argue that excluding subsidiaries and with-it Volkswagen is the better choice since members of the management of Porsche Holding should have control over Volkswagen as a subsidiary. I, therefore, assume the power to rest on the Porsche Holding board. The benefit is that tight clusters of firms belonging to the same corporate group are dissolved. The consequence of this choice is that important ties between boards that go out from a subsidiary board are excluded. When interested in all ties of an elite community of managers in contrast, it is justifiable to include all subsidiaries and their management boards, because these boards individually may have important ties to other companies – and not only the board of the top holding company. In this case, an unknown error driven by administrative ties is accepted in return.

One of the things one needs to be aware of in this regard when using ORBIS is that there is no easy way in between these two extremes of excluding all subsidiaries or keeping them. While there is a variable provided in ORBIS for some corporate groups containing the "level" within a corporate hierarchy, this is not necessarily meaningful information to solve this problem and even if it was it is only available for a small share of corporate groups. Taking an example from my research on super-rich capitalist families, the question of what is *the* firm that makes up most of the fortune of a family is not trivial either. Asking someone on the street what company made Dieter Schwarz the richest German, might lead to the quick answer "No doubt, it's Lidl." But the supermarket chain Lidl is owned by a foundation in Germany with a revenue of more than 79bn € and 120,000 employees. This foundation is owned by five other firms and foundations, one of which is Lidl SB KG which is again held by D. Schwarz Beteiligungs KG which is again held by six firms and foundations, and this continues for a couple of levels until ending up at a level where the ultimately owning individual is located among some other companies that belong to him. The point of this is that corporate hierarchies are complex, and this full complexity is mercilessly reflected in ORBIS there is no easy way to identify a firm if you do not have an exact idea of what you mean by that. To make such an idea more concrete, one example could be "the entity with the largest revenue in their unconsolidated accounts within a corporate hierarchy". But even if you have an exact definition such as this one, the question of how to operationalize it considering other data problems such as mixed-up total and direct shares within corporate groups and incorrect GUO information is quite another question.

There may be cases where researchers have an external list of firm names that they want to find in the ORBIS database. At first sight, the good news is that there is a batch firm search implemented in the ORBIS portal that is supposed to make it easy to match such lists. The bad news is that depending on what exactly you are looking for, the result might not be the entity in ORBIS you are interested in but some other subsidiary of the corporate group.[12] Sticking with the example above, there are many companies containing the name Lidl in ORBIS. Using the batch search to find the firm gives as a first suggestion Lidl Great Britain Ltd. (Figure 4). Adding Germany as a country gives another branch, the Lidl GmbH & Co. KG which is not the head of the corporate group. Depending on your research question, this might be a problem. In some cases, it may help to simply include the whole corporate group, but in some, this will not help, not the least because the corporate group data is inaccurate as well, as will be discussed below in the accuracy section. This example also illustrates a general ironic point for the first time that is one of my main points in this text: while the database promises to enable large-scale analyses of millions of firms, often the devil is in the detail of individual cases. The definition of the correct unit of analysis might only be possible for every individual corporate group by hand – or by accepting a certain unknown error that is impossible to oversee at a large scale.



| | Company name | City | Country | Identifier | Score |
|---|---|---|---|---|---|
| ✓ | Lidl | | | | Search again |
| ● | LIDL GREAT BRITAIN LIMITED (Previous name: LIDL LIMITED) | SURBITON | GB | 02816429 | A |
| ○ | LIDL SP. Z O.O. SP.K. | JANKOWICE | PL | 7811897358 | A |
| ○ | LIDL SUOMI KOMMANDIITTIYHTIO (Alias: LIDL) | ESPOO | FI | FI16157790 | A |
| ○ | LIDL ASIA PTE. LIMITED (Previous name: LIDL SINGAPORE PTE. ... | SINGAPORE | SG | 201819013D | A |
| ○ | LIDL GMBH & CO. KG | WOELLSTEIN | DE | HRA 32259 (MAI... | A |
| ○ | LIDL GMBH & CO. KG | WASBEK | DE | HRA 1294 NM (K... | A |
| ○ | LIDL GMBH & CO. KG | HILDESHEIM | DE | HRA 201900 (HI... | A |

**Figure 4:** Suggestions from the ORBIS batch search implemented in the web portal when searching for the company name "Lidl" (screenshot).

A final problem I would like to point out concerning the unit of analysis is the problem of financial account types and consolidation of financial data in ORBIS. This is something that works interested in macroeconomic aggregates are very concerned about (e.g. Bajgar et al. 2020; Kalemli-Ozcan et al. 2022b). Generally, when interested in financial accounts it makes a relevant difference

[12] Or a completely different firm since this is only based on string matching and it therefore depends on how common a name is.

which accounts types to include. Recalling Table 2, only half of the firms contained in ORBIS have recent financial data at all, roughly one-fourth have limited financials, and about 25m (5%) have detailed financial data. Of the available account data 2.5m are consolidated, i.e., including subsidiary accounts, and 42.7m are unconsolidated, i.e., only counting the respective entity.[13] For example, simply adding up unconsolidated accounts does not lead to the value of the holding company. This does not have to be problematic for all research questions, e.g. when double counting subsidiaries and holding companies is not a concern. However, for example, when interested in combining financial accounts with ownership data, this *is* a problem. How do you value shares held by a shareholder? One way is to use firm financials as a proxy for firm value and multiply it by the share held. Researchers for example have used a similar method based on the ORBIS data to identify wealthy individual shareholders for survey research (Schröder et al. 2020). However, it becomes a very difficult task in cases such as the one by Dieter Schwarz illustrated above. There are two related sets of problems: One concerns the accuracy (see below) and the complexity of the ownership data. The other is more related to the unit of analysis and concerns which accounts to use as a proxy to estimate the value of the shares. The problem is that to my knowledge there is no reliable default way to determine which accounts to use if one does not want to count double accounts of subsidiaries.[14] A feasible option might be to consolidate all accounts of firms owned by firms owned by the shareholder of interest, but the problem just mentioned that consolidating all accounts of subsidiaries does not lead to the correct consolidated account of the head of the corporate group. This way you may, but not necessarily, get a correct ranking of wealthy shareholders, but the absolute numbers e.g. of share or revenue "owned" by an individual shareholder are not going to be correct. In individual cases, it might be possible to solve this by identifying the correct largest consolidated account, but in many cases, there are only unconsolidated accounts. Even more pressingly, this would require again to look at every individual case or at least the problematic ones which contradict the goal of a large-scale analysis of millions of firms or shareholders.

[13] If not specified otherwise the cited numbers from the ORBIS portal stem were valid in January 2023.

[14] Some suggestions in this regard are made by Kalemli-Ozcan et al. (e.g. 2022b:78–81).

For their longitudinal analysis of market concentration in Europe, Kalemli-Ozcan et al. (2022a:4–5) discuss that may lead to biased results by only using a certain type of accounting statements:

> "A priori there is no reason for focusing on a certain set of accounting statements as opposed to combining all statements, as long as one is careful about not double counting the same firm reporting both statements. Focusing on a selected set of statements will lead to focusing on a selected set of firms such as listed firms, business groups, foreign firms, and will give misleading trends in concentration. This practice of selecting certain groups will also deliver biased results due to changing regulation. For example, we show a sharp increase in concentration around 2007, which coincides with a change in the European accounting legislation."

Only taking certain account types to avoid double counting subsidiaries and heads of corporate groups therefore also does not seem to be a solution.

To sum up, the more researchers think about what they mean by a "firm", the less clear this may become. Excluding all subsidiaries of a corporate group may exclude important ties in the firm hierarchy, although the "firms" along the path will mainly exist on paper and their existence may have legal, tax, or other purposes. There are at least two opposing conclusions to draw from this. On the one hand, the complexity of the firm data in ORBIS could lead to the conclusion that it is not even meaningful to think in terms of "firms" or "corporate groups" when using the data but to treat the different entities as nodes in a network that contains a relevant extent of noise. The widespread use of the term firm does not resemble the empirical reality any longer since it is undercomplex. I would rather tend to the alternative conclusion which is that ORBIS is lacking a reliable indicator to identify ownership conglomerates that allow capturing what is widely understood as new forms of modern corporations (e.g. Reurink and Garcia-Bernardo 2020). The takeaways for researchers and reviewers concerning the unit of analysis, therefore, are the following questions:

1. **How do you/ does the reviewed work define the unit of analysis or answer the question "what is a firm"? How are subsidiaries treated?**
2. **Is it possible to operationalize this definition based on the ORBIS data?**
3. **What consequences does the definition have for the analysis, results, and possibly accepted errors, uncertainties, and shortcomings?**

### Entity ambiguity

The problem of entity ambiguity is closely related to the problem of the unit of analysis. In a way, it is the question of how to get from the definition of what a firm "is" to the operationalization of including all entities that should be covered (ambiguity) exactly once (duplicates). One could add that it is also about how to structure the entities, whether subsidiaries should be separate, or all consolidated. In my view, this might be the single most problematic aspect of the ORBIS database. The data is simply messy in this regard as is also documented elsewhere (Garcia-Bernardo and Takes 2018; Heemskerk et al. 2018). Frankly, in my view, the shortcomings in ORBIS regarding entity ambiguity cannot be stressed enough. To begin with, ORBIS simply does not only include data for firms but also other entities. The ownership information we downloaded for example also includes shares held in different kinds of museums, municipal utilities, non-profit associations or clubs, but also various central banks such as the ECB. It is not trivial to filter these out systematically, but it may to some extent be possible by excluding specific sectors and non-profit entities.

Entity ambiguity can further easily be illustrated by giving an example. Figure 5 illustrates the network of interlocking directorates between the Global Ultimate Owners of all firms containing the word "blackrock". Ties mean that at least one manager from one corporate group also sits on the board of another corporate group.  In an ideal data world, one might expect *one* Global Ultimate Owner (GUO) or head of the corporate group, namely Blackrock Inc. If this was the case, the GUO variable could easily be used to aggregate all entities that are economically owned by the Blackrock corporate group in this case. However, as the network illustrates this is not the case. At first sight, the availability of the GUO variable may be an apparent advantage over other databases such as Compustat which does not include a comparable variable about the GUO. It must be noted however that while the GUO variable promises a lot, namely, to identify the common global ultimate owner of a corporate group, the variable is often erroneous. This was confirmed to me also in conversations with other researchers working with the data who have made similar experiences. There are numerous examples one could give in addition to the Blackrock example. Although in this case, only part of this is due to missing links in the ORBIS database and therefore a matter of accuracy or data quality. In part, these "firms" are funds or other subsidiaries which may factually be shared among different shareholders not only from the BlackRock corporate group. A few of them are also completely different firms that are called "Blackrock" but do not have ownership ties to the asset management corporation. Again, depending on the research question this might be problematic. However, it is relatively clear that in many cases the GUO does not provide the actual global ultimate owner which is also why some researchers chose to gather ownership information from other sources (Aminadav and Papaioannou 2020:1196–97). This is yet another illustration of the ORBIS problem that while the database includes a large sample of firms, in many research applications individual cases have to be checked and additional information has to be researched (and not the least matched) by hand.

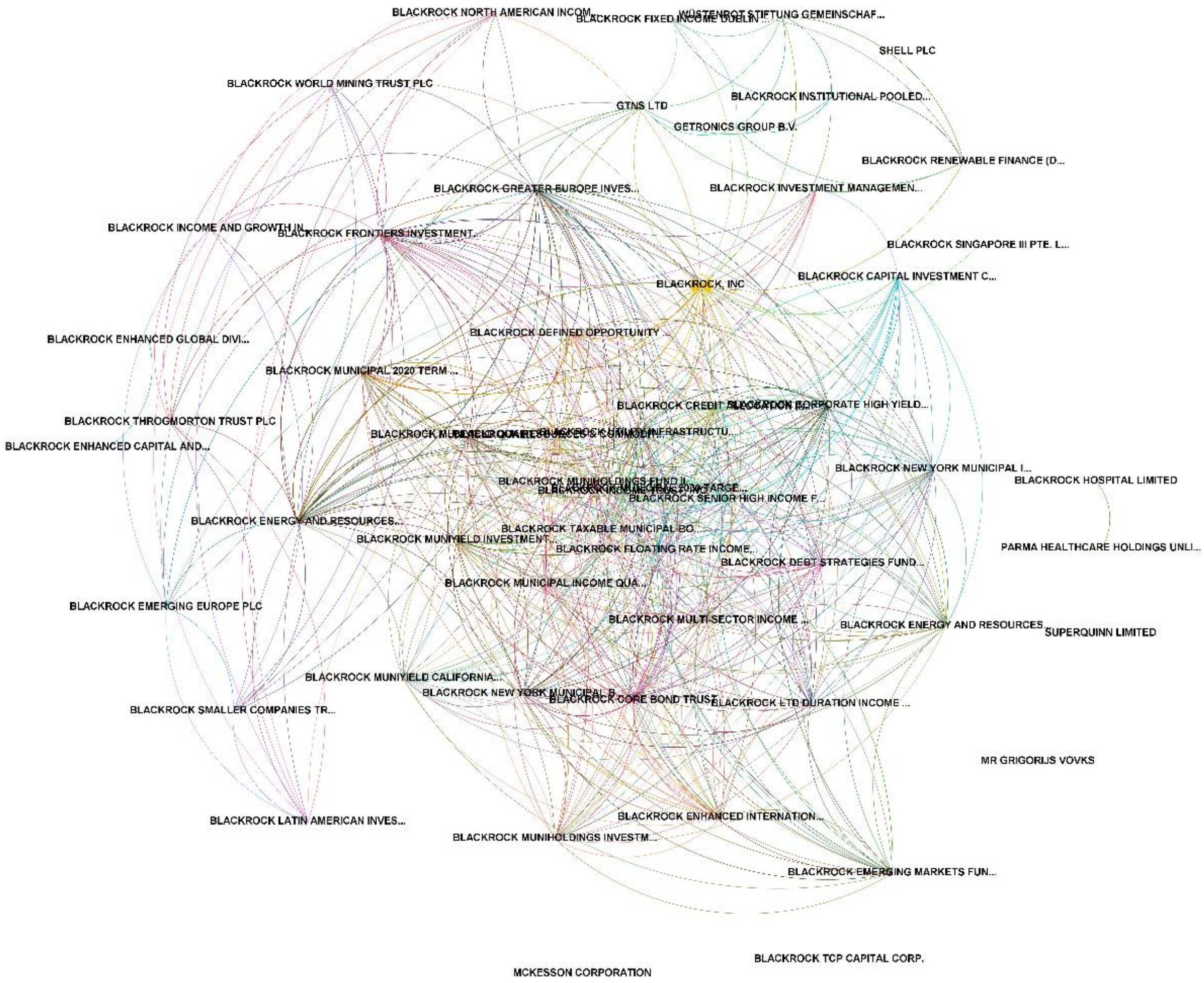

BLACKROCK NORTH AMERICAN INCOM...
BLACKROCK FIXED INCOME DUBLIN ...
WÜSTENROT STIFTUNG GEMEINSCHAF...
SHELL PLC
BLACKROCK WORLD MINING TRUST PLC
GTNS LTD
BLACKROCK INSTITUTIONAL POOLED...
GETRONICS GROUP B.V.
BLACKROCK RENEWABLE FINANCE (D...
BLACKROCK GREATER EUROPE INVES...
BLACKROCK INVESTMENT MANAGEMEN...
BLACKROCK INCOME AND GROWTH IN...
BLACKROCK FRONTIERS INVESTMENT...
BLACKROCK SINGAPORE III PTE. L...
BLACKROCK, INC
BLACKROCK CAPITAL INVESTMENT C...
BLACKROCK DEFINED OPPORTUNITY ...
BLACKROCK ENHANCED GLOBAL DIVI...
BLACKROCK MUNICIPAL 2020 TERM ...
BLACKROCK CREDIT ALLOCATION I...
BLACKROCK CORPORATE HIGH YIELD...
BLACKROCK THROGMORTON TRUST PLC
BLACKROCK RESOURCES & COMMODIT...
BLACKROCK UTILITY INFRASTRUCTU...
BLACKROCK ENHANCED CAPITAL AND...
BLACKROCK NEW YORK MUNICIPAL I...
BLACKROCK HOSPITAL LIMITED
BLACKROCK MUNIHOLDINGS FUND II...
BLACKROCK INCOME TRUST, INC
BLACKROCK MUNICIPAL 2030 TARGE...
BLACKROCK SENIOR HIGH INCOME F...
BLACKROCK ENERGY AND RESOURCES...
BLACKROCK TAXABLE MUNICIPAL BO...
BLACKROCK MUNIYIELD INVESTMENT...
BLACKROCK FLOATING RATE INCOME...
PARMA HEALTHCARE HOLDINGS UNLI...
BLACKROCK DEBT STRATEGIES FUND...
BLACKROCK MUNICIPAL INCOME QUA...
BLACKROCK EMERGING EUROPE PLC
BLACKROCK MULTI-SECTOR INCOME ...
BLACKROCK ENERGY AND RESOURCES
SUPERQUINN LIMITED
BLACKROCK MUNIYIELD CALIFORNIA...
BLACKROCK NEW YORK MUNICIPAL B...
BLACKROCK CORE BOND TRUST
BLACKROCK LTD DURATION INCOME ...
BLACKROCK SMALLER COMPANIES TR...
MR GRIGORIJS VOVKS
BLACKROCK LATIN AMERICAN INVES...
BLACKROCK ENHANCED INTERNATION...
BLACKROCK MUNIHOLDINGS INVESTM...
BLACKROCK EMERGING MARKETS FUN...
BLACKROCK TCP CAPITAL CORP.
MCKESSON CORPORATION

**Figure 5:** Who owns companies that are called “Blackrock”? The network of interlocking directorates between all Global Ultimate Owners (GUOs) owning any company with a name containing the word "blackrock".

But how to solve this issue of entity ambiguity of ORBIS? How to work around the fact that there are uncountable entities which somehow – depending on the definition – should belong to the same corporate group but according to the ORBIS data they do not? Once more, I cannot offer much hope for researchers looking for solutions completely accounting for this. One way to deal with the problem is to change the definition of the GUO to a lower threshold of shares of a company held. But this solves only part of the problem. Some literature suggests different forms of entity resolution methods based on string methods, exploiting additional data such as firm addresses, or topological methods based on networks such as the one illustrated in Figure 5 (Garcia-Bernardo and Takes 2018; Heemskerk et al. 2018). However, I was not successful in effectively improving the GUO data in ORBIS by applying any of these classes of solutions. In the following, I briefly summarize why.

In my case, I was working with a very large sample from ORBIS of around 3 million global firms and in addition around 7 million shareholding entities. The problem with data of such size is that almost every imaginable name or word will occur multiple times in the data. Blackrock is even probably one of the easier cases because you do have a relatively unique identifying word "Blackrock" that you could use. In fact, by hand, the problem may be solved with an hour of work maximum for this case. But this does not help much when you need to prepare hundreds of thousands of corporate groups and you cannot rely on the corporate group membership variables. I tried different ways of identifying the most "significant" word among a firm name, e.g., in terms of how frequently the word occurs in the whole corpus. In this example, "Blackrock" is a meaningful string to find companies of the corporate group while "Inc." is not. To give another easy example, many firms are named after their founding family. A lot of founding families are called Smith, Jackson, or Becker. So you might have strings like the fictional "Smith Bakery Shops New York", and "Smith Baker Law Firm", but also "New York Bakery Smithtown". These firms would of course be distinct corporate groups. Also, there are many firms with indistinct names such as "International Service Corporation" or "Credit Limited Holding" where there is no chance of identifying a meaningful part of the string that will help you to structure firms into correct corporate groups. [15]

It should also be mentioned here that comparing distances between strings for a large set of strings requires huge adjacency matrices and can get very costly in terms of working memory very quickly. Trying to compute the Jaccard distance between all 1.3 million Chinese companies in my sample would have required more than two terabytes of working memory. Splitting up matrices into subgroups based on some other criteria suggesting that firms may be part of the same corporate group or not may then become costly in terms of computation time.[16] There are suggestions to preselect the number of firms for which string comparison is made to reduce these matrices. But first, this is maybe a good time to constate that things get very complicated in terms of implementation, and you need good knowledge of a programming language and linear algebra to handle a large number of strings, matrices, distance calculations, etc. Developing and comparing good methods to reduce entity ambiguity can quickly become a research project by itself with a significant chance of failure. Most researchers will not even have the time to implement something that sufficiently accounts for this. But assuming such skills and time resources are available, this

[15] I am aware that there are many different string distance measures and matching methods, e.g., based on n-grams. But I do insist that none of them will solve the fundamental issue that with a sample of this size, most firm name strings become uninformative because they are simply not sufficient to decide whether two firms are part of the same corporate group or not.

[16] All of this depends on sample size of course. It may be feasible for a couple of hundred or thousand firms, but it is not for the millions of firms ORBIS claims to cover.

brings us away from the suggestion to simply use string comparisons to combine these with other measures, e.g., based on network topology.

In theory, and this is also suggested in the literature, one could use string matching algorithms as sketched above in combination with other methods such as topological network methods, although this requires network data and thereby does not apply to many applications only working with the ORBIS financial data. A very helpful of such method to identify duplicates are presented in Garcia-Bernardo and Takes (2018). It seems intriguing to use networks of overlaps in management or ownership relations to identify corporate groups. But several problems remain even using these methods. First, in both cases, information is not available for all firms. So, it depends on your sample for what share of companies and corporate groups this is even applicable. Second, the method relies on the GUO variable to identify corporate group membership which is often inaccurate, as it may not cover all potential duplicates, as discussed above. Third, one must find suitable thresholds of closeness in the network within which firms can safely be assumed to be in the same corporate group. While this works well to identify duplicates (see below), I was not successful in reliably identifying other members, and only other members, of a corporate group with these methods. Especially based on ownership data, many companies may have very similar network positions but may not be part of the same corporate group. When adding string matching to this I ran tono problems on both sides: many firms do not share any part of the string but are part of the same corporate group, and some firms share a string but are not part of the same corporate group. Merging those entities that do have similar network positions and share a string should be fine but this way you do only solve a (small) part of the problem.

A second aspect of entity ambiguity illustrating this is simply duplicate firms, i.e., entities that occur multiple times but with differing unique IDs in ORBIS. This may be the case with the same firm names or slightly different firm names and also applies to shareholding entities and managers. As Kalemli-Oczan et al. (2022b:5) report, such duplicates may also be introduced when combining snapshots from different time points and firms' IDs changed in between. Duplicates are of course a problem that may distort results. BvD claims in the ORBIS online support documents that an external data science firm already removed duplicates as far as possible. Nevertheless, a relevant number of duplicates remains. Garcia-Bernardo and Takes (2018) developed an effective method to identify duplicate firms. Based on the network of interlocking directorates, they identify those firms whose overlap in management, that is the same people employed as managers, suggesting that they are referring to the same firm. Applying this method to the Swedish network of interlocking directorates, the authors reduce the original number of entities including subsidiaries by 47% (!) (Garcia-Bernardo and Takes 2018:169). In my work where I excluded all subsidiaries, I applied the procedure suggested by Garcia-Bernardo and Takes, as well as simple name duplicates. With these methods, I find 21,548 duplicate entities (7%) only among the 290,169 heads of corporate groups in my sample. One shortcoming of this method, however, is that it is only feasible at the national level (at least without adjustment), and for a manageable size of firms. Finding duplicates among all Chinese firms, for example, will exceed working memory capacities almost certainly with this method.

The problem of duplicates might even be more precarious among shareholders. It is easy to spot duplicate shareholding entities when simply checking individual cases by hand. For one of my research projects, I ran a supervised learning procedure to deduplicate the ORBIS shareholder data for my sample. I trained an algorithm for deduplication and matching based on multiple variables and their interactions (Dedupe.io 2022). The algorithm was trained to learn based on 1000 manually labeled cases whether two records are referring to the same shareholding entity based on their name, country, city, shareholder type, postcode, founding year (if a company), first name, and last name (if individuals). Among the 17,207,181 shareholders in my sample, I identify 186,737 (1%) duplicate entities with enough certainty.[17] However, this only captures a small share, and I spotted numerous additional duplicates among the remaining shareholders which could not be captured with enough certainty in this way. Again, the degree to which duplicates among firms, shareholders, and managers distorts results depends on the research question, sample, and research design. Yet, it is a problem that needs to be accounted for. In summary, the takeaways for researchers and reviewers concerning entity ambiguity are the following questions:

1. **How do you/ does the reviewed work operationalize the unit of analysis or classify firms as belonging to a firm or corporate group?**
2. **Does the work rely on the Global Ultimate Owner variable from ORBIS? Is it in line with the definition of "what is a firm"?**
3. **Which inaccuracies in corporate group membership does this introduce and tolerate? Is it relevant to the results?**
4. **How do you deal with duplicate firms, shareholders, and managers? Do they distort results?**

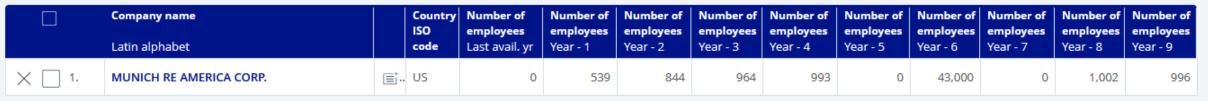
### Accuracy

The problem of accuracy is one of incorrect information. The ORBIS database includes numerous errors.[18] As was described along with the background information, the data comes from different

[17] The Shareholder – BvD ID numbers of these duplicates, a unique identifier, and the matching score can be downloaded from my GitHub page.

[18] I can only mention as a side note here that this and other problems may also be very relevant for business users of ORBIS and the decisions which are based on it. More specific examination of the ORBIS data for use cases from the business world are necessary to assess this more systematically.

data providers. This may be one reason for the inaccuracies. I have spotted several such examples. Again, I could not think of a method or illustration to oversee and describe the full extent. Heemskerk et al. (2018:25–27) suggest diagnosing and fixing inconsistencies by checking for outliers which is an intuitive and helpful first step. However, this does not account for all types of inaccuracies. Figure 6 shows an excerpt from the ORBIS portal showing the number of employees of an American Branch of Munich RE over ten years. It ranges between 0 and 43,000 but is on average around 900 employees with a jump from 0 to 40,000 and back to 0 within three years. Something is wrong with this record, but it remains unclear without further research into the database whether this error is due to some definitional problem or rather correct information for some unintuitive reason. Assuming it is an error: It is true, that problems such as these may be captured and fixed by looking for outliers, but here outlier detection needs to be looking at time series and from my assessment, it will be very difficult to accurately separate actual erroneous outliers from correct large fluctuations for a large number of cases in this way.

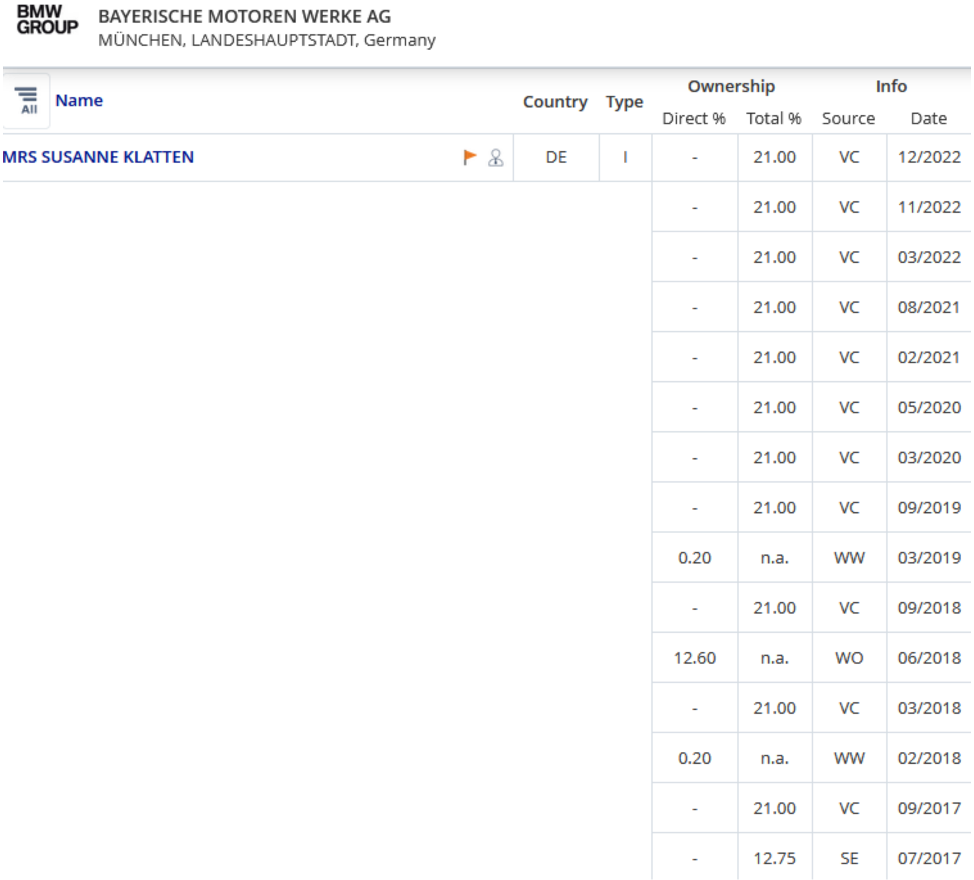


| □ | Company name<br>Latin alphabet | | Country ISO code | Number of employees Last avail. yr | Number of employees Year - 1 | Number of employees Year - 2 | Number of employees Year - 3 | Number of employees Year - 4 | Number of employees Year - 5 | Number of employees Year - 6 | Number of employees Year - 7 | Number of employees Year - 8 | Number of employees Year - 9 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| × □ 1. | MUNICH RE AMERICA CORP. | | US | 0 | 539 | 844 | 964 | 993 | 0 | 43,000 | 0 | 1,002 | 996 |

**Figure 6:** Inaccuracies in the number of employees of Munich RE America Corp (screenshot from the ORBIS portal).

I cannot elaborate here on all forms of inaccuracies in the data that I have spotted. A very illustrative example is that the two largest asset managers globally, BlackRock and Vanguard, are wrongly classified in the shareholder data as a bank, and as a non-financial corporation. Some other examples from the ORBIS financial information are also implied in Kalemli-Oczan et al.'s (2022b:22) suggestions to exclude outliers in terms of total assets and sales. I also do believe that there is not one systematic method to deal with them and whether they are relevant for research results depends on the exact sample used and the research question, as well as the kind of error. The only contribution I can make here in this regard is to draw attention to the fact that such errors exist, and one needs to be careful in research applications. However, I would like to mention one more example that might be relevant for researchers analyzing ownership data. It comes from my research on super-rich ownership and management of firms. Namely, the problem is the interchangeability of direct and total shares depending on the data provider. Figure 7 illustrates this for the case of Susanne Klatten, one of the richest Germans, and her history of shareholdings in car maker BMW. The column source denotes the data provider. One can see how entries from different data providers lead to inaccuracies, or at least potential for error in data handling, in the time series but also the meaning of the variables. The entry of total shares from September 2017 stemming from SE is lower than all other values reported by VC. It is much closer to Ms. Klatten's direct share reported by WO in 2018. This, however, stands in stark contrast to the other direct shares reported by WW for two-time points. The missing values may be distorting interpretation and interpolation attempts as well which needs to be taken into account carefully in data handling. Which of these values is the correct direct and indirect ownership information at the respective time points? It seems impossible to make this call without consulting external sources. And this is just one example of one of the largest and most famous shareholders, owning shares in one of the largest companies in one of the largest economies in the world which counters hopes articulated in the literature, that data quality in ORBIS is good enough the larger the companies and the larger the countries' economies – especially in Europe.

The most constructive comment I can make on this specific point is that I have made the best experiences for the research task of identifying large owners, i.e. super-rich individuals and families, by using mostly total shares, interpolating gaps between total shares, and if entries are still missing, complementing missing total shares with direct shares. This is because, in the matching procedure, it was a necessary precondition that not only the name of the super-rich individual match the shareholder's name but that it also actually holds a share in a matching firm name. Through this data preparation procedure, I maximized the number of identified super-rich individuals among the largest firms. However, when truly interested in the distinct meanings of direct and total shares and using the data for this, I can only advise against it based on the inconsistencies in the data in this regard.



BAYERISCHE MOTOREN WERKE AG
MÜNCHEN, LANDESHAUPTSTADT, Germany

| Name | Country | Type | Ownership | | Info | |
|---|---|---|---|---|---|---|
| | | | Direct % | Total % | Source | Date |
| MRS SUSANNE KLATTEN | DE | I | - | 21.00 | VC | 12/2022 |
| | | | - | 21.00 | VC | 11/2022 |
| | | | - | 21.00 | VC | 03/2022 |
| | | | - | 21.00 | VC | 08/2021 |
| | | | - | 21.00 | VC | 02/2021 |
| | | | - | 21.00 | VC | 05/2020 |
| | | | - | 21.00 | VC | 03/2020 |
| | | | - | 21.00 | VC | 09/2019 |
| | | | 0.20 | n.a. | WW | 03/2019 |
| | | | - | 21.00 | VC | 09/2018 |
| | | | 12.60 | n.a. | WO | 06/2018 |
| | | | - | 21.00 | VC | 03/2018 |
| | | | 0.20 | n.a. | WW | 02/2018 |
| | | | - | 21.00 | VC | 09/2017 |
| | | | - | 12.75 | SE | 07/2017 |

**Figure 7:** Shareholder history of Susanne Klatten's ownership in BMW in ORBIS stemming from different sources (screenshot from the ORBIS portal).

To close the section on accuracy, researchers should consider questions along the following:

1. **How do you/ does the reviewed work check for erroneous data?**
2. **How are values over time interpolated? Are these checked for inconsistencies?**
3. **Has a serious diagnosis of accuracy taken place? Can it be visualized or otherwise demonstrated?**

### Completeness

Two types of completeness must be separated and discussed distinctly. First, this is the completeness of the population, or at least the representativity of samples, from the ORBIS database. Second, this is missing information among variables provided in ORBIS. I will begin with only a few summarizing sentences on completeness and representativity since this is discussed in detail elsewhere (Bajgar et al. 2020; Garcia-Bernardo and Takes 2018; Kalemli-Ozcan et al. 2022a). I will then concentrate more on missing information after that.

Among other things, existing contributions have demonstrated that coverage in ORBIS is better for countries with higher GDP and varies by region (e.g. Garcia-Bernardo and Takes

2018:166–68). This might for example be due to varying legal requirements of limited liability companies to register, e.g. depending on their size (Kalemli-Ozcan et al. 2022a:6). Bajgar et al. (2020) report that representativity of the ORBIS data varies by year, by country, and even by industry. Coverage of large and top-performing firms seems to be good, however weighting of the sample to make it more representative does not seem to be possible since inclusion in ORBIS is non-random even when taking size into account (Bajgar et al. 2020:9). The authors conclude that "Overall, Orbis seems more suitable for studies that: i) take a global perspective rather than making comparisons across countries; ii) analyse top performers and multinationals rather than underperforming firms; iii) and focus on mean performance or within-firm changes rather than on the entire firm distribution or entry and exit" (Bajgar et al. 2020:3). All these contributions compare samples from ORBIS to external statistics and should be consulted for further details.

A distinct question is that of missing information. Financial data is generally only available for a small share of all companies in ORBIS. This was already illustrated in section two. Figure 8 visualizes the extent of missingness for some financial variables separately for six samples. First listed (a) vs. unlisted firms (b), second US S&P 500 listed firms (c) vs. Chinese Shanghai Stock Exchange listed firms (d), and third, the largest listed and unlisted firms in terms of revenue (e) vs. the largest global listed and unlisted firms in terms of their number of employees (f). All data comes from our second download in 2022. Since completeness in ORBIS depends on accounting regulations and there are more strict regulations for listed firms, it seems intuitive that listed firm data is more complete than non-listed firm data. For the latter, even the most widely available variables have a missing share of 75% among the ca. 3m largest non-listed firms globally. Also, the completeness of variables differs by listed and unlisted firms. Profit and loss for example seem to be more available for listed firms and the number of employees for unlisted firms.[19] Among listed firms, completeness of information seems to be similarly high at least for large economies as shown here for the US and China, and very similar numbers are also found for listed German firms. When going beyond listed firms, the completeness of information decreases rapidly as can be seen when comparing the completeness of data for the largest firms (Figures 8e and 8f) with data of listed firms only (Figure 8a).

[19] As discussed in section 3, I do not believe that the high share of missingness is due to the method of access and download. However, to fully clarify this it would be good to see a comparison with the missingness structure from the historical disks.

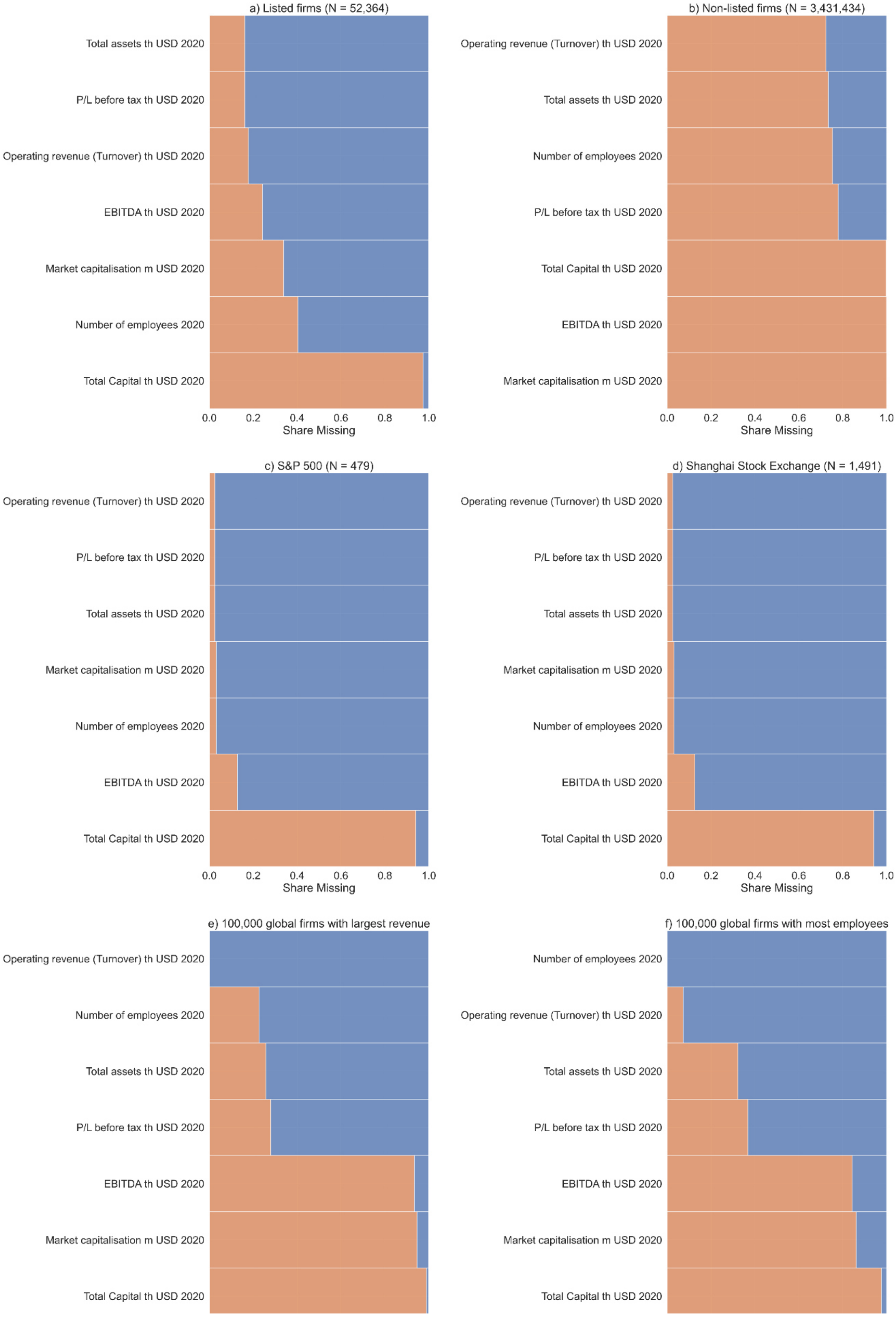


**Figure 8:** Degree of missingness of some key financial variables in 2020. Orange: Share of missing records. Blue: Share of non-missing records. a) All listed firms, b) all non-listed firms, c) S&P 500 listed firms, d) SSE listed firms, e) 100,000 largest firms in terms of revenue, and f) largest firms in terms of the number of employees. The download took place between January and June of 2022.

Bajgar et al. (2020), following Gal (2013), suggest imputing missing values and showing promising results for the imputation of value-added. However, as the authors acknowledge this requires the availability of other information such as the cost of employees which is in many cases also not available. Yet, they present evidence showing that "Internal imputation substantially increases coverage for about half the countries in the sample, and it reduces it for Norway. […] The internal imputation makes the firm distribution in Orbis more representative in terms of mean firm characteristics, and it moves the level of heterogeneity in the bottom half of the productivity distribution closer to that observed in the population […]. Using internal imputation moves average firm size and labor productivity closer to that observed in official microdata. It also increases the productivity dispersion in the lower half of the distribution" (Bajgar et al. 2020:32). Unfortunately, this conclusion conflicts with the fact that international comparison is often affected by differing reporting standards as gathered by Kalemli-Oczan et al. (2022b:73–76) or the meaning of e.g. management positions as constated by Heemskerk et al. (2018:8): "different countries have different governance structures, rules and regulations. A non-executive director in China is not the same as a non-executive in the UK. A big data approach easily allows for the study of, for instance, board interlocks across the globe, but decontextualizing boards and firms may lead to invalid conclusions."

From my experience imputation is a task that should not be underestimated and some of the imputations of market capitalization I performed based on random forests led to clearly wrong estimates. Most likely these were related to my sample choice for imputation, i.e. which firms and subsidiaries to include and which ones not, as well as account types and double counting of heads of corporate groups and subsidiaries. I then had to cut on further efforts due to the time restrictions of my project. For many tasks, imputation of small firms, branches, and subsidiaries may make decrease data quality because missingness itself says something about the firm in ORBIS (country, industry, firm type, reporting, and accounting requirements, transparency of companies such as potential shell companies, etc.). Not all of this information is available as variables that can be used for imputation which means the data is not missing at random and imputation will distort results to an unknown extent. Imputing the revenue of all companies in a corporate hierarchy for example including administrative companies will inflate the revenue of a corporate group. These kinds of problems possibly can be circumvented by choosing the right sample such as only heads of corporate groups and being aware of consolidated and unconsolidated account types, but this solution is dependent on the correct identification of subsidiaries as discussed above as entity ambiguity.

As a final problem of completeness to be mentioned here, I would like to refer to Bajgar et al. (2020) once more who underline that some reported financial data in ORBIS is only imputed by estimates. This leads to the problem that for firms with limited financial accounts, the number of employees and operating revenues are only estimated as the middle of a range especially "for the Czech Republic, Poland, Slovakia and the United States [and] a smaller, but non-negligible, role for Finland, Japan, Sweden and Italy" (Bajgar et al. 2020:49). To close the section on completeness, researchers should consider questions such as the following:

1. **Does representativity/ coverage matter for results?**
2. **If the answer to 1 is yes, which countries/industries/firm sizes does your work/ does the reviewed work include? What do the existing external validity checks of ORBIS coverage say about coverage in these countries (Bajgar et al. 2020; Garcia-Bernardo and Takes 2018; Kalemli-Ozcan et al. 2022a)?**
3. **Which account types are chosen and how are missing values dealt with?**

## 5. Reflections

In the following, I offer some reflections on, and defense of, the points made in this article. These thoughts deal with the questions of when to believe ORBIS data after all, the origins of the problems in the political economy of corporate corporate data, the question of when a dataset is just not good enough, and finally, the idea of public goods as an answer.

### When should we believe research using ORBIS firm data?

In my opinion, three things can increase trust to an acceptable level. First, authors need to show that they have engaged with the existing literature and the manifold known data problems. That is, at this point and in my view, especially Kalemli-Oczan (2022b, 2022a, n.d.)[20], Bajgar et al. (2020), and Heemskerk et al. (2018). Second, authors must deliver evidence that they have engaged in-depth with their specific sample, problems that are relevant to their research questions either already known from the literature or newly discovered by themselves. This includes being explicit about their access method of the ORBIS data, being transparent about their choices of data preparation, and at least how they tackled known existing problems. I hope that the questions gathered above may help to structure the necessary documentation of the use of ORBIS data for research. Finally, in this regard, authors should be explicit about potential errors known from the literature that they could not solve and that they, therefore, tolerate in their interpretation of results. Third, and possibly most demanding, research should ideally be reviewed by other researchers who have reasonable experience with or knowledge of ORBIS data. I do try to give a guideline here for

[20] But consider the diverging experiences we made concerning downloads from the web portal reported in section 3.

reviewers less familiar with the data of where to point the finger, but to fully evaluate which of the problems apply likely more knowledge of the data is necessary. This might be a difficult claim since the amount of researchers knowing the data as far as I can tell is very limited. It might be worthwhile to collect a public list of researchers, or set up a community some other way, of researchers familiar with the data. This could be a source from which journals could draw anonymous reviewers for the data side, or researchers working with the data could help each other out. While one could argue that this is also not done for most other datasets, I do think that the number and complexity of problems with the ORBIS data is larger and more pressing than for most other frequently used datasets for research and it is therefore necessary. I do believe (and have in fact witnessed) that in cases where researchers are not familiar to the downsides of the dataset and that there is a whole literature about it, the large promises of ORBIS – like Homer's sirens - may make blind for its perils and in the worst-case lead to wrong scientific knowledge that could have been prevented.[21]

## The political economy of corporate corporate data

One could object to my comments elaborated here that I am overly pessimistic about the data, as well as the research community's use of it, and not making a constructive contribution.[22] Also, similar problems as the ones of ORBIS are known from other big data sources, as is the potential messiness of big data in general. Is this just a situation of being skeptical towards technological and methodological progress as were people of railroads in the 19th century? After all, the alternative would be not to use the data at all. Therefore, I would like to defend and justify my position before closing on the most constructive note I can. I do believe that the exploitation of large data sources is the methodological future of social sciences. However, social sciences were always known for critically reflecting on their research methods and I believe this is one of its main strengths. The sources of the problems with ORBIS from my assessment are the following. A commercial enterprise gathers prepares and sells the data mainly to a target group that has very different needs and demands than the research community. Since the enterprise is commercial, it does matter that customers from the business world probably make up the majority share of BvD's revenue with the data. Selling the data to scientist rather seems to be a side business for BvD that came later in its company history. This origin leads to the fact that the higher demands of researchers in terms of e.g., representativity, harmonization and comparability across countries, historical data and its

[21] Which in the case of ORBIS is the more problematic because there is no other dataset available or in sight for many questions to disprove findings based on other more trustworthy sources.

[22] I do also have my problems with the imperative of constructiveness and this case made here exemplifies some of the reasons why, but debating this would be beside the point here.

consistency, entity ambiguity, etc. were not considered in building the database from the start. Existing efforts e.g., to reduce duplicate individuals in the database by external data engineering companies cannot fully account for this. At the same time, ORBIS holds a monopoly position, especially for the international perspective as well as unique combinations of information e.g., of shareholders, management, and financials. Since BvD is a commercial enterprise, they have an interest in making a profit and not in being proactively transparent about shortcomings of the data for the generation of reliable scientific knowledge. After all, which company is proactive in advertising the shortcomings of its products? This is of course a fundamental difference between publicly financed datasets such as many national panel surveys. In sum, therefore, I do believe that ORBIS' history and legacy, as well as the fact that it is owned and maintained by a commercial enterprise with a different target group, are the source of its specific problems. It is therefore more problematic in this sense than other data sources. But is it too problematic to use for scientific research? This is the final point I would like to offer some thoughts on.

## Just not good enough?

Heemskerk et al.'s (2018) ambition is "to begin a conversation about research process standards now if we are to advance the quality of the research community in the future." I very much agree that this is necessary, and my point here is that we need standards when a (e.g. new or particularly large and complex) data source is not good enough for research. Good enough is itself a matter of definition of course, and I do mean it here in the sense that on average over all applications it could lead to too many invalid inferences at least for very large and international samples – which is exactly the kind of applications the dataset invites itself most for in comparison to other data. In the case of ORBIS, I disagree with the authors that the necessary means to "assess the extent to which data quality issues exist and what it means for the meaning that we derive from the analysis of concern" (Heemskerk et al. 2018:6) are presented by them, are yet available, nor will ever be available as long as the data does not fundamentally improve. This, however, is an empirical question. What would be necessary is a large reproducibility initiative of trying to reproduce existing (if only the most influential) research results generated with ORBIS data. I would however not say much about their validity of course.

## Public goods

To close on a more constructive note, the wide use of the ORBIS data despite its problems shows that there is a high demand for researchers for large-scale international firm data. Questions that can be answered with such data are highly relevant to the lives of many. Possibly an initiative of well-funded international research institutions and their libraries or some other public institutions could collaborate to set up a non-commercial alternative to ORBIS. One that is constructed with the needs of public research and reliable scientific knowledge in the first place. Comparable examples although smaller in scale already exist, such as the open global firm database OpenCorporates (2023) or the related leaked German company register (OffeneRegister 2023). Until these or other alternatives can compete with ORBIS, the open question remains how many problems it needs for a data source not to invite itself for the generation of scientific knowledge – or at least only under very specific and highly costly circumstances which are especially years of experience and the necessary skills to use the data. With the comments and questions raised here, I hope to help reviewers and authors to conclude for themselves at least as long as there is no comparable alternative data source, no significant improvement of the data quality on the side of BvD, and no more systematic external checks of the data's validity and potential for reproducibility.

## References

Aminadav, Gur, and Elias Papaioannou. 2020. "Corporate Control around the World." *The Journal of Finance* LXXV(3):1191–1246.

Bajgar, Matej, Giuseppe Berlingieri, Sara Calligaris, Chiara Criscuolo, and Jonathan Timmis. 2020. *Coverage and Representativeness of Orbis Data*. Paris: OECD. doi: 10.1787/c7bdaa03-en.

Bureau van Dijk. 2020. "Orbis. Powering the Business of Certainty." Retrieved March 4, 2020 (https://www.bvdinfo.com/en-gb/our-products/data/international/orbis).

Dedupe.io. 2022. "Dedupe." Retrieved March 20, 2023 (https://github.com/dedupeio/dedupe).

Gal, Peter N. 2013. "Measuring Total Factor Productivity at the Firm Level Using OECD-ORBIS." *OECD Economics Department Working Papers*.

Garcia-Bernardo, Javier, Jan Fichtner, Frank W. Takes, and Eelke M. Heemskerk. 2017. "Uncovering Offshore Financial Centers: Conduits and Sinks in the Global Corporate Ownership Network." *Scientific Reports* 7(1):1–10.

Garcia-Bernardo, Javier, and Frank W. Takes. 2018. "The Effects of Data Quality on the Analysis of Corporate Board Interlock Networks." *Information Systems* 78:164–72.

Group van Dijk. 2013. "Presentation of Van Dijk Management Consultancy Companies." *Group van Dijk*. Retrieved March 20, 2023 (https://web.archive.org/web/20150708012732/http://www.vandijkmc.com/en/group-van-dijk_35.aspx).

Heemskerk, Eelke M., and Frank W. Takes. 2016a. "Supplementary Material for 'The Corporate Elite Community Structure of Global Capitalism' by Eelke Heemskerk & Frank Takes." Retrieved January 17, 2023 (https://pure.uva.nl/ws/files/67005080/cnpe_a_1041483_sm2199.pdf).

Heemskerk, Eelke M., and Frank W. Takes. 2016b. "The Corporate Elite Community Structure of Global Capitalism." *New Political Economy* 21(1):90–118. doi: 10.1080/13563467.2015.1041483.

Heemskerk, Eelke M., Kevin Young, Frank W. Takes, Bruce Cronin, Javier Garcia-Bernardo, Lasse F. Henriksen, William Kindred Winecoff, Vladimir Popov, and Audrey Laurin-Lamothe. 2018. "The Promise and Perils of Using Big Data in the Study of Corporate Networks: Problems, Diagnostics and Fixes." *Global Networks* 18(1):3–32. doi: 10.1111/glob.12183.

Kalemli-Ozcan, Sebnem, Jingting Fan, and Veronika Penciakova. n.d. "Processing ORBIS Historical Disk." Retrieved January 12, 2023 (http://econweb.umd.edu/~kalemli/assets/policy&blog/Processing%20Orbis%20Historical%20Disk%20FINAL.pdf).

Kalemli-Ozcan, Sebnem, Bent Sorensen, Carolina Villegas-Sanchez, Vadym Volosovych, and Sevcan Yesiltas. 2022a. """How to Construct Nationally Representative Firm Level Data from the ORBIS Global Database: New Facts and Aggregate Implications."""" National Bureau of Economic Research, Cambridge, MA.

Kalemli-Ozcan, Sebnem, Bent Sorensen, Carolina Villegas-Sanchez, Vadym Volosovych, and Sevcan Yesiltas. 2022b. """Online Appendix to How to Construct Nationally Representative Firm Level Data from the ORBIS Global Database: New Facts and Aggregate Implications."""" National Bureau of Economic Research, Cambridge, MA.

Liu, Grace. 2020. "Data Quality Problems Troubling Business and Financial Researchers: A Literature Review and Synthetic Analysis." *Journal of Business & Finance Librarianship* 25(3–4):315–71. doi: 10.1080/08963568.2020.1847555.

OffeneRegister. 2023. "Licht Ins Dunkel Der Firmengeflechte." Retrieved March 20, 2023 (https://offeneregister.de/).

OpenCorporates. 2023. "The Largest Open Database of Companies in the World." Retrieved March 20, 2023 (https://opencorporates.com/).

Reurink, Arjan, and Javier Garcia-Bernardo. 2020. "Competing for Capitals: The Great Fragmentation of the Firm and Varieties of FDI Attraction Profiles in the European Union." *Review of International Political Economy* 1–34.

Selenium. 2023. "The Selenium Browser Automation Project." Retrieved February 27, 2023 (https://www.selenium.dev/).

# Linking Wealth and Power

## Unity and Political Action of the World's Wealthiest Capitalist Families and the Corporate Elite

Inauguraldissertation
zur
Erlangung des Doktorgrades
der
Wirtschafts- und Sozialwissenschaftlichen Fakultät
der
Universität zu Köln

2023

presented
by

Hans Lukas Richard Arndt, MSc M.A.

from

Minden

| | |
|---|---|
| First reviewer: | Prof. Jens Beckert<br>(MPIfG and WiSo Uni Köln) |
| Second reviewer: | Prof. Olivier Godechot<br>(CNRS-Sciences Po Paris-CRIS) |
| Third reviewer: | Prof. Clemens Kroneberg<br>(ISS, WiSo Uni Köln) |
| Fourth reviewer: | Prof. André Kaiser<br>(CCCP, WiSo Uni Köln) |
| Fifth reviewer: | Prof. Lucas Chancel<br>(CNRS-Sciences Po Paris-CRIS) |
| First external reviewer: | Prof. Catherine Comet<br>(Université Paris 8 Vincennes-Saint-Denis) |
| Second external reviewer: | Prof. Mark S. Mizruchi<br>(University of Michigan) |
| Date of oral defense: | 26 June 2023 |

This dissertation was written in accordance with a joint cotutelle agreement between the University of Cologne and Sciences Po Paris. The exact same version with a differing title page was submitted to the doctoral school of Sciences Po Paris.

# Content