



A guide to micro-CbCR databases

Summary

Country-by-Country Reporting (CbCR) has become one of the most important new sources of information on the global activities of multinational enterprises (MNEs). Originally introduced within the OECD/G20 BEPS framework for tax transparency and risk assessment purposes, micro-level CbCR data now offer unprecedented opportunities for both academic research and tax administrations to analyze the international organization of multinational firms, the location of profits and economic activity, and the functioning of the global tax system.

However, despite the increasing analytical importance of these data, firm-level CbCR datasets remain difficult to use in practice. Raw files frequently contain inconsistencies in identifiers, duplicates, reporting errors, sign reversals, unit inconsistencies, currency problems, and heterogeneous reporting practices across firms and jurisdictions. Moreover, the structure of the CbCR itself — based on jurisdiction-level aggregation of unconsolidated entity accounts — raises additional challenges related to intra-group transactions, permanent establishments, and the potential double counting of dividends.

This Note provides a practical methodological guide for cleaning and processing firm-level (“micro”) CbCR data. It proposes a set of harmonized procedures aimed at improving the internal consistency, longitudinal comparability, and analytical usability of the data. The document covers all major stages of the cleaning process, including:

- harmonization of multinational identifiers (TINs and names);
- detection of duplicates and filer switches;
- reconciliation between Table 1 and Table 2 of the CbCR;
- cleaning of affiliate-level information and business activities;
- correction of currency, sign, and unit inconsistencies;
- identification of subsidiary-level reporting errors;
- benchmarking with external firm-level datasets;
- comparison with OECD aggregated CbCR statistics;
- correction procedures for the double counting of dividends.

The guide is intended both for researchers working with micro-CbCR data and for statisticians or analysts within tax administrations responsible for processing exchanged CbCR files and producing aggregated statistics. Particular emphasis is placed on reproducibility, documentation of corrections, and the importance of maintaining clear audit trails throughout the cleaning process.

A central message of this guide is that careful cleaning is a prerequisite for any meaningful use of firm-level CbCR data. At the same time, the document shows that, once properly processed, micro-CbCR

data appear broadly consistent with other major sources of information on multinational enterprises. This confirms the considerable potential of CbCR data as a rich source for tax analysis, international macroeconomic statistics, and research on multinational firms.

Finally, the Note argues that the exchange of best practices across researchers and tax administrations is essential. Greater convergence in cleaning methodologies would improve the comparability of empirical studies, strengthen the quality of OECD aggregated CbCR statistics, and contribute to a more reliable use of these data for both analytical and administrative purposes.

GUIDE IN A NUTSHELL

Micro CbCR Cleaning:

1. Firm-level CbCR data are expected to become a central data source for understanding the activity of multinational enterprises, both for academic research and for tax administrations.
2. Despite gradual improvements in CbCR guidelines and reporting practices, firm-level data still contain a significant number of inconsistencies and reporting errors. Conducting a thorough data-cleaning process is therefore a sine qua non condition before any use of the data, whether for statistical analysis or tax audit purposes.
3. The exchange of best practices regarding data cleaning is essential. The greater the convergence in cleaning procedures across users of these data, the higher the comparability across studies and the better the quality of aggregated CbCR statistics will be.
4. After cleaning, CbCR data appear broadly consistent with other sources of information on multinational enterprises, highlighting the potential of CbCR as a rich and valuable source of data for both tax administrations and economic analysis.

This guide has received funding from the European Union (GA No. TAXUD 2026/DE/402) and Research Council of Norway, grants no. 341289 and 352981. The views expressed are those of the authors and do not necessarily reflect the views of the European Commission. Neither the European Union nor the granting authority can be held responsible for them.

We thank for their helpful comments and suggestions: Acciari Paolo, Tomas Boukal, Barbara Bratta, Thijs Busschots, Alice Chiochetti, Sébastien Lafitte, Idann Gidron, Vera Santomartino, Arjan Soede, Ruairi Sugrue.

Contents

1	Unique identification of MNE groups, subgroups and constituent entities	9
1.1	Basic cleaning	9
1.2	Reporting role and surrogate filing	10
1.3	Inconsistent Company Identifiers (TIN and company name)	10
1.4	Tracking companies across years (filer switches)	11
1.5	Duplicates	13
2	Affiliate-level cleaning and Business Activities	15
2.1	Constituent entities identification	15
2.2	Countries, functions and PEs	15
3	Cleaning financials at the subgroup level	18
3.1	Consolidation of affiliate-level reporting of financial information	18
3.2	Currency harmonization	18
3.3	Sign inconsistencies	19
3.3.1	Sign of tax paid and tax accrued	20
3.3.2	Sign of revenues	21
3.4	Standardizing reporting units	22
4	Benchmarking with external data sources	24
4.1	Matching with MNE-level data	24
4.1.1	Orbis	25
4.2	Comparison with aggregated CbCR	27
4.3	Correct remaining outliers	29
5	Biases, limitations and challenges	31
5.1	Double-counting of dividends	31
5.1.1	Cross-validation with national administrative data	33
5.2	MNE sample bias	33
	Glossary of key terms	37

Introduction

Country-by-Country Reporting (CbCR) is a standardized reporting framework introduced under Action 13 of the OECD/G20 Base Erosion and Profit Shifting (BEPS) project, requiring large multinational enterprise groups to report key financial and tax information on a jurisdiction-by-jurisdiction basis. By providing data on revenues, profits, taxes, employment, and tangible assets, CbCR has significantly improved transparency in international taxation and offers an unprecedented view of how multinational groups deploy their activities and profits across countries.

This information is a major asset both for researchers seeking to better understand the global organization of multinational firms and for tax administrations aiming to assess tax risks and audit compliance with international tax rules. However, despite its richness, CbCR data are not immediately ready for analysis. They raise a number of complex data-cleaning issues, reflecting reporting heterogeneity and inconsistencies. Addressing these challenges is a prerequisite for any meaningful empirical use of CbCR data, and is precisely the focus of the present guide.

This document provides readers with a cleaning guide for micro-CbCR data. It is intended both for researchers with access to country-by-country reporting data at the level of multinational enterprise (MNE) groups, once compiled by tax administrations, and for statisticians working within tax administrations who receive raw data from MNEs and seek either to analyze these data or to transmit them, in aggregated form, to the OECD Secretariat. It is also advisable to consult the issues identified by tax administrations regarding the quality of the reporting provided by multinational enterprises, as documented by the OECD¹.

Several recent academic studies have made use of firm-level CbCR data and have therefore gone through an extensive data-cleaning phase (Fuest and Neumeier, 2022; Bratta et al., 2024; Fuest et al., 2025; Chiocchetti et al., 2025; Boukal et al., 2026). However, because data cleaning was not the primary contribution of these papers, the associated methodological choices and challenges are discussed only briefly or implicitly. By contrast, this document focuses specifically on the cleaning of these data and on assessing their strengths and limitations, in the same spirit as Delpeuch et al. (2024) for the first years of available French multinational data.

This guide highlights potential issues in the micro-level CbCR data, proposes fixes to circumvent some potential data limitations (green boxes), suggests useful merges with external sources (pink boxes) and lists some useful descriptive statistics that can be drawn out of this data cleaning work (blue tables). It also gathers links to papers and reports that can be useful for the data work on micro-level CbCR.

The remainder of this guide is structured as follows. The first section focuses on inconsistencies in company identifiers, including tax identification numbers and firm names, and on the presence of duplicates. The second section examines the second table of the CbCR form and shows how to establish

¹<https://www.oecd.org/content/dam/oecd/en/topics/policy-sub-issues/cbcr/common-errors-mnes-cbc-reports.pdf>

a clean list of MNEs' constituent entities with their country of residence and their main business activity. Section 3 turns to the cleaning of financial variables, discussing currency conversions, outliers, unit and reporting inconsistencies and specific challenges related to tax variables. Section 4 lists best practices when benchmarking the cleaned firm-level data against external sources, in order to assess internal consistency and external validity. Finally, Section 5 discusses the remaining biases, limitations, and unresolved challenges inherent in the use of CbCR data for analytical and administrative purposes.

A note on scope. The procedures described in this guide are designed for large datasets settings, where systematic cleaning is the only tractable approach and a full manual check of every observation is infeasible. This is not the relevant setting for every user. Tax administrations receiving exchanged CbCR from a comparatively small number of foreign MNE groups may find that the population of subgroups to clean is small enough for a full hand check of group identifiers, sector assignments, and time-consistency across years — and such a targeted, manual review should then be preferred over the automated rules proposed below. In that setting, this guide is best read as a checklist of issues to verify by hand rather than as a set of rules to encode.

Data structure

The CbCR dataset is based on two standardized forms (Table 1 and Table 2) that are filed annually by each eligible MNE group. The CbC-reporting is mandatory for each MNE with an annual turnover above 750 M EUR and whose parent jurisdiction is part of the OECD/G20 inclusive framework.

In Table 1, a CbCR-reporting MNE group consolidates the financial statements of all its constituent entities on a country-by-country basis and reports, for each jurisdiction in which it operates, total revenues (distinguishing between related-party and unrelated-party transactions), profit before tax, taxes accrued and paid, tangible assets, number of employees, and stated capital. Each row in Table 1 therefore corresponds to the activity of a given MNE group in a specific country and year. This unit of observation is referred to as a sub-group. The list of the financial variables given at the subgroup level is given in Table 1, together with a corresponding currency variable indicating the currency used for the reporting (typically indexed by the "_currcode" pattern).

In Table 2, a CbCR-reporting MNE group lists all its constituent entities, indicating their country of incorporation as well as their main business activity or activities, selected from a standardized list of 13 possible functions (see Table 2 below). The MNE-level identifier (TIN) is also indicated in Table 2, allowing for reconciliation between the two tables. However, for a given MNE, it is not possible to decompose country-level financial information by the constituent entities located in that country. The micro-CbCR dataset therefore remains, first and foremost, a subgroup-level dataset rather than an affiliate-level dataset, although the list of affiliates constituting each group is available in the CbCR.

TABLE 1**Overview of financial variables at the Subgroup Level (Jurisdiction × Year)**

Variable	Description
revenue_unrelated	Revenues from unrelated parties
revenue_related	Revenues from related parties
revenue_total	Total revenues
profit_or_loss	Profit (loss) before income tax
tax_paid	Income tax paid (cash basis)
tax_accrued	Income tax accrued (current year)
capital	Stated capital
earnings	Accumulated earnings
assets	Tangible assets other than cash and cash equivalents
employees	Number of employees

Figure 4, in appendix 5.2, gives a stylized illustration of how the three levels relate: the jurisdiction-level rows of Table 1, the entity-level listing of Table 2, and the underlying legal structure of the group. It shows why Table 1 cannot be decomposed to the entity level even though the entities themselves are observable.

TABLE 2**Overview of main business activities (entity Level)**

Variable	Description
R_D	Research and Development
IP	Holding or managing intellectual property
purchasing	Purchasing or procurement
manufacturing	Manufacturing or production
sales	Sales, marketing or distribution
admin	Administrative, management or support services
services	Provision of services to unrelated parties
intragroup_finance	Internal group finance
regulated_finance	Regulated financial services
insurance	Insurance
holding_shares	Holding shares or other equity instruments
dormant	Dormant entity
other	Other activities (not elsewhere classified)

Each multinational group is uniquely identified by a MNE group name and a MNE - Tax Identification Number (TIN). In theory, one MNE group should only have one name across year as well as a unique TIN. At the **Entity Level**, each entity (row) is identified by entity_name and entity_tin.

Note for tax administrations: from XML files to Tables 1–3

Tax administrations typically do not receive Tables 1–3 directly in the tabular form presented above, but rather as structured XML files following the OECD CbC XML Schema, from which Tables 1–3 must first be extracted. A few fields specific to this raw format are useful at this stage and are no longer visible once the data has been flattened into Table 1/2/3 form:

- `doc_ref_id` uniquely identifies each reported document, and `corr_doc_ref_id` references the `doc_ref_id` of a previously filed document that the current one corrects or replaces. Amended filings should be reconciled using this field *before* the deduplication steps of Section 1, rather than being treated as ordinary duplicates.
- `reporting_role` and `legal_address_type` are both available at this stage (see Section 1 for the use of `reporting_role` in detecting surrogate filing).
- The reporting period end date makes it possible to distinguish the MNE's *financial year* from the *calendar year* used for storage or exchange purposes; the two need not coincide, and MNE-years should be defined on the former.

Potential issue(s): A small number of jurisdictions have, in some years, transmitted files in a corrupted or non-standard format that precludes systematic parsing. These cases should be identified and set aside up front, rather than treated as ordinary data-cleaning issues to be fixed downstream.

General caveats and overarching confidentiality principles

Several general disclaimers apply to all stages of the data cleaning process described in this guide.

First, residual uncertainty may always remain. The CbCR database is inherently different in nature from administrative tax data. It is constructed for reporting and transparency purposes rather than for direct fiscal administration. As a result, certain inconsistencies, reporting biases, or structural ambiguities may persist even after careful cleaning.

Second, all corrections, adjustments, and methodological choices implemented during the cleaning process must be systematically documented. Maintaining a clear audit trail is essential in order to ensure reproducibility, allow for later revisions, and facilitate transparent discussion of data treatment decisions.

Third, no country-specific results referring to a particular headquarters jurisdiction should be published without prior authorization from the country concerned. More generally, outputs must be presented in a way that never allows the identification of an individual firm, whether directly or indirectly.

Finally, in the French case in particular, users must familiarize themselves with and strictly comply with the applicable confidentiality rules currently in force. These rules are detailed in the official confidentiality guidelines provided by the relevant data access authority.

HOW TO READ THIS GUIDE

Throughout the document, three recurring elements flag the type of content. They can be read independently depending on the reader's needs:

Fix

A concrete, actionable cleaning procedure addressing a specific data issue.

External Datasource

A description of an external dataset useful for benchmarking or enriching the micro-CbCR data, together with guidance on how to use it.

Descriptive statistics tables flag diagnostic questions and suggested statistics that can be produced at each stage to assess data quality:

Item	Questions
<i>n.m</i>	A suggested statistic or diagnostic question.

1 Unique identification of MNE groups, subgroups and constituent entities

The first challenge when working with the data is to ensure that the reporting format of the main table (Table 1) is properly followed. A row in Table 1 must correspond to the activity of a multinational company in a specific country for a given year, and that company's activity in that country during that period must be represented by only that single row.

Additionally, it is essential to ensure that the identification of companies reporting Country-by-Country Reports is usable. This means verifying that each Tax Identification Number (TIN) corresponds to a unique company name, and vice versa. The most common causes of inconsistency in this regard mainly stem from CbCRs being submitted by entities other than the group parent company ("surrogate filing") or from other forms of duplicates at the subgroup-year level.

1.1 Basic cleaning

We begin by describing the basic cleaning steps performed prior to applying corrections to ensure unique MNE TIN-name combinations:

- Remove MNE-year groups when all jurisdictions within the group have all financial variables equal to zero

- Remove rows with missing country and currency information
- Remove MNE-year observations with only one entity (one row)
- Harmonize name inconsistencies across entity countries within a TIN-year-declarant country combination by assigning the most frequently reported name.

Potential issue(s): Across reporting years, the names of multinational groups filing CbCRs may vary slightly, which complicates the longitudinal tracking of groups over time.

Fix — Clean MNEs' names

Company names should be normalized through a standardization procedure.

- Convert all characters to uppercase
- Strip special characters and punctuation
- Remove recurring non-informative expressions (e.g., CORP, INC, LTD, AND SUBSIDIARY, GROUP, etc.)

1.2 Reporting role and surrogate filing

Potential issue(s): Duplicates can arise when a report is sent twice by different entities. Typically it can be sent both by the headquarters and a surrogate Parent Entity.

Fix — Clean MNEs based on reporting roles

The micro-CbCR keeps track of the Reporting Role. A variable specifies the role of the Reporting Entity with respect to the filing of the CbC Report. Possible values are:

- CBC701 – Ultimate Parent Entity
- CBC702 or CBC703 – Surrogate Parent Entity

When the Reporting Role variable is unavailable, one solution is to aggregate the data at the MNExyear level to detect duplicated declarations with the same amount of financial variables :

- Distinguish duplicated declarations reported by different TINs with the same declarant country or with different declarant countries
- One solution is to keep the most frequent TIN among the duplicated declarations

1.3 Inconsistent Company Identifiers (TIN and company name)

In the raw data, companies may report different Tax Identification Numbers or use varying legal or trade names across years. This inconsistency makes it difficult to reliably track the same multinational over time, as no unique and persistent identifier exists in the source data.

Fix — Unique identification of MNE groups

Harmonization of TIN and company name:

- **Same TIN, different company names** Example: TIN 12345678 reported as ABC FINANCE in one year and ABC FINANCIAL SERVICES in another. Solution: create a column with “new name” where the same name is attributed to the same TIN (can be the most frequent name).
- **Same company name, different TINs** Example: ABC reported with TIN 12345678 in one year and 87654321 in another. Solution: create a column with “new TIN” where the same TIN is attributed to the same company name (can be the most frequent TIN).
- **Same company name, different TINs, and different declarant countries** Example: ABC reported with TIN 12345678 in one year with declarant country A, and with TIN 87654321 in another year with declarant country B. Solution: assign the most frequent TIN and flag these observations for manual review. The review should determine whether the discrepancy reflects an inversion of the headquarters jurisdiction or an error in the reported declarant country.
- **Slight variations in TINs across years due to country-specific conventions** Example: In Sweden, TINs may be reported with a country prefix (SE1234567890) or with additional trailing characters (SE123456789001). Solution: apply a country specific cleaning step, e.g. remove the SE prefix, and truncate the last two characters if necessary.
- **Keep a column with “old TIN”** to spot mistakes in next steps (e.g., matching with Orbis, see Section 4).
- **Missing TIN (NOTIN)** A substantial share of observations report a placeholder value (e.g. NOTIN) instead of an actual TIN. Since this cannot be used as an identifier, these observations require additional harmonization measures based on company name alone (see Section 1.1). This should be kept in mind consistently wherever a TIN-based join is used elsewhere in this guide, including for panelization (Section 1.4).
- **Companies changing TIN and NAME over the years** Solution: See Section 1.4

1.4 Tracking companies across years (filer switches)

Potential issue(s): Beyond the inconsistencies in TIN and name described above, a more subtle case arises when the entity designated to file the CbCR return changes from one year to the next. In this situation, both the TIN and the company name change simultaneously: the previous TIN disappears from the panel and a new one appears, with no shared identifier linking them. Standard harmonization by name or by TIN cannot resolve these cases, since neither is preserved across the transition.

This pattern is observationally indistinguishable in the raw data from a genuine entry or exit (a group crossing the EUR 750M threshold for the first time, or being dissolved or fully absorbed). Misclassifying a filer switch as an exit fragments the longitudinal panel; misclassifying a true entry as a filer switch

artificially merges two distinct groups. Both errors bias any analysis that depends on consistent group identity over time.

Fix — Panelization

We implement a dedicated linkage routine to detect filer switches across adjacent years. The procedure is restricted to the relevant subset of the panel, identifies candidate transitions, scores them on observable characteristics, and produces a stable group identifier for longitudinal analysis. The detection of filer switches relies on comparing three key variables (revenues, employees, and tangible assets), and should therefore be performed after the cleaning of financial variables detailed in Sections 3 and 4.

Step 1: restrict the search. Classify each TIN by its presence pattern in the panel: always-present, entrant only, exiter only, both, or irregular (with gaps). TINs present every year cannot, by construction, be a filer switch and are excluded from candidate matching.

Step 2: build candidate pairs For each TIN's exit event (last observed year) and entry event (first observed year), retain the pair only if:

- the two TINs have the same HQ country (enforced as a join key, no exception);
- the entry year follows the exit year by at most two years (allowing for a single missing filing or a temporary drop below the reporting threshold);
- revenues, employees, and tangible assets are each within a reasonable ratio of one another (default minima: 0.40, 0.35, 0.30 respectively).

As noted in Section 1.1, TIN is sometimes reported as a placeholder value (e.g. NOTIN) rather than a genuine identifier; when this affects one side of a candidate pair, the HQ-country join key above must instead be recovered from company name and country of residence.

Step 3: score retained candidates A composite similarity score in $[0, 1]$ is computed from four weighted components: the Jaccard similarity of the two CbCR jurisdiction footprints (weight 0.45), the revenue ratio (0.25), the employee ratio (0.15), and the tangible-asset ratio (0.15). The footprint receives the largest weight because, for groups operating in tens of jurisdictions, the set of reported countries is a near-unique fingerprint that drifts only gradually from year to year.

Step 4: tier the candidates. Each candidate is assigned to one of three tiers:

- **Auto-merge** when the composite score and each individual component exceed stringent thresholds simultaneously (typically $\text{score} \geq 0.85$, $\text{footprint} \geq 0.80$, $\text{revenue ratio} \geq 0.85$, $\text{employee ratio} \geq 0.75$, $\text{asset ratio} \geq 0.70$);
- **Review** when the composite score is moderate ($0.60 \leq \text{score} < \text{auto threshold}$). These pairs are written to a CSV for manual adjudication and are not merged unless explicitly confirmed;
- **Reject** when the composite score falls below 0.60.

The double condition for auto-merge (composite score *and* each component individually) prevents a single very strong signal from compensating for a weak one. The thresholds reported

here are starting values and should be calibrated on each panel by inspecting the score distribution and a sample of auto-merged pairs in the audit report; tighter thresholds reduce false positives at the cost of more pairs falling into the review tier.

Step 5: assemble the final mapping. All confirmed links (auto-merges and reviewed pairs marked as the same group) are combined to build the final TIN-to-group mapping. For instance, if TIN *A* is linked to *B* and *B* is linked to *C*, all three receive the same `mne_id`, even if *A* and *C* were never directly compared. Each group is labelled using its earliest TIN, providing a stable identifier.

Alternative when good external-database coverage is available. When a reliable external firm-level panel is available for most of the sample (e.g. CapitalIQ or Orbis, see Section 4), teams may prefer to reverse the order of operations above: build the group-level panel primarily from the external database's own linkage of legal entities to ultimate owners (i.e. starting from what is Step 5 here), and use Steps 2–4 as an internal validation check on the resulting CbCR panel rather than as the primary source of identification. This is particularly advisable when reported financial variables are unstable across years for reasons unrelated to filer switches (e.g. currency or unit changes, see Section 3), since such instability weakens the ratio-based components of the composite score in Step 3 regardless of whether a genuine filer switch occurred.

Suggested descriptive statistics

Item	Questions
2.1	How many MNE groups are reporting CbCR every year?
2.2	How many CbCR groups are present across all years ("stayers")? How many are entering ("entrants") every year? How many are exiting ("exiters")?
2.3	How many CbCR per multinational (defined as TIN or NAME) are in the sample?

1.5 Duplicates

Potential issue(s): Remove duplicates at the MNE-year-country level when needed. Duplicated reports can be due to:

- Various end of reporting date within a MNE-country-year
- Various currencies
- Reporting at the affiliate level rather than the country level

Fix — Deduplication of Table 1

In case of duplicates, respectively:

- **Keep the longest reporting period:** For instance, if multiple rows exist for the same MNE-country-year combination with different period end dates (e.g., 2019-08-31 and 2019-12-31), keep only the row with the most recent reporting date (2019-12-31).

- **Keep USD or EUR**
- Use the number of affiliates from Table 2 to detect potential reporting at the affiliate level and aggregate the declarations (see Section 3.1)
- For the remaining duplicates, keep only one observation MNE-year-country (for instance, the one with the highest unrelated revenue)

2 Affiliate-level cleaning and Business Activities

In Table 2, all MNEs list their constituent entities and their main activity or activities (i.e. each affiliate can conduct more than one function) out of 13 possible business functions. This information is valuable in itself, as it helps describe the global activity of multinational companies. Counting the subsidiaries of multinational companies also improves the identification of multinational groups over time and makes it easier to build a panel at the Table 1 level. For this reason, we recommend cleaning this table at this stage of the data cleaning process, even if the final use focuses only on the first table.

2.1 Constituent entities identification

For each year, Table 2 lists the constituent entities of multinational groups that file Country-by-Country Reports. The first step is to reconcile the identity (name and TIN) of the multinational groups in Table 1 with those in Table 2. In theory, all multinational groups in Table 1 should have their list of subsidiaries in Table 2. Conversely, every subsidiary listed in Table 2 should belong to a multinational group whose identity can be found in Table 1. A simple matching based on group-level TINs in table 1 and 2 should be the first step with Table 2.

Potential issue: Inconsistencies often arise in the identity of groups between Table 1 and Table 2.

Fix — Correspondence between table 1 and 2

String matching based on MNE names with Table 1: For Table 2 lines for which the identity was not found in Table 1 based on TIN matching, a string matching based on MNE name can be performed. To improve the quality of the string matching, the name filter applied on Table 1 (see Section 1) shall also be applied on Table 2's MNE name before proceeding to the matching.

Potential issue(s): The variable in Table 2 that should correspond to the multinational group's TIN is sometimes confused with the subsidiary's identification. This complicates the matching process with Table 1, for which the MNE TIN is the pivotal variable.

Fix — Table 2 cleaning

Recurring patterns in affiliates names: For unmatched affiliate in Table 2, clean affiliates names using the name filter described in section 1.1 and identify the word(s) in common in all affiliates. These groups of affiliate sharing a core naming pattern can then be reconciled with cleaned MNE names from Table 1.

2.2 Countries, functions and PEs

Potential issue(s): In CbCR Table 2, a constituent entity and its permanent establishment(s) (PEs) in other jurisdictions may share the same identifier within a given year. This reflects the reporting standard itself: according to OECD (2022), a PE should be listed in Table 2 by reference to the tax

jurisdiction in which it is situated, with the legal entity of which it is a PE noted separately — typically under a name such as “XYZ Corp – Tax jurisdiction A PE”.

Fix — PE and branches in Table 2

Construct a unique `establishment_tin` identifier that disambiguates upper affiliates from their PEs.

- **If entities are reported in long format**, use a combination of textual signals in the entity name and identifier (presence of keywords such as *PE*, *Permanent*, *Branch*, or *notin*) and the multiplicity of observations sharing the same `entity_tin` within a year. Within each group of duplicates, the parent row is selected by prioritising rows with no PE signal, then rows with the broadest functional profile as measured by the number of business activities declared.
- **if entities are reported in wide format** (multiple tax jurisdictions in separate columns), the country of incorporation column directly identifies the parent affiliate. In both cases, PE rows receive a suffixed identifier (`entity_tin-1`, `entity_tin-2`, etc.), and the two datasets are appended to produce a single harmonized table.

The ‘Other’ business activity and Table 3

The `other` category in Table 2 (“activities not elsewhere classified”) is used very unevenly across reporting jurisdictions: some tax administrations rarely see it used at all, while others — some low-tax jurisdictions in particular — see it used pervasively, at times with vague or uninformative free-text descriptions attached in Table 3. Since Table 3 is the free-text section in which MNE groups may clarify entries in Tables 1 and 2 (see Glossary), a heavy or vague use of `other` should not be taken at face value.

Fix — Checking the ‘Other’ category

- When an affiliate’s only reported function is `other`, check the corresponding Table 3 free-text entry, if any, for a more informative description of the activity before falling back on external sources (e.g. Orbis, see Section 4).
- Compare the prevalence of `other` across declarant countries (item 2.23 below): a share of `other` that is systematically higher in some jurisdictions than others is itself informative, and may signal that reporting entities there use the category to avoid disclosing a more specific activity, rather than reflecting a genuine difference in the underlying business.

Suggested descriptive statistics

Item	Questions
2.21	How many functions per affiliate do we typically observe in the data?
2.22	What is the prevalence of each function?
2.23	Do MNEs often report other activity? What do MNEs report in Table 3?
2.24	How often do affiliates have PE in other jurisdictions? And which are the most popular countries for such PEs?

3 Cleaning financials at the subgroup level

The core value and richness of Country-by-Country Reports lie in the financial information at the subgroup level: the activities of a multinational group within a country. However, raw data can contain inconsistencies and input errors that must be minimised to ensure the analysis remains relevant.

3.1 Consolidation of affiliate-level reporting of financial information

Certain firms may have improperly reported their activity at the subsidiary level rather than at the subgroup level. In Table 1, this effectively generates multiple rows for a single multinational in a given country and year. Unlike true duplicates, this repetition of rows within an MNE–country–year triplet cannot be corrected by simply keeping one of the entries; instead, the observations must be consolidated in order to recover the economic activity of the multinational group in the country concerned for that year.

Potential issue: Reporting financial and economic data at the subsidiary level creates inconsistencies in Table 1 and must therefore be consolidated. To address this, this type of reporting error must be distinguished from other forms of duplicates.

Fix — Detect subsidiary-level reporting in Table 1

Two indicators can help distinguish subsidiary-level reporting from other types of duplicates:

- A substantial discrepancy in the value of a given variable (for example, the number of employees) within an MNE–country–year triplet suggests subsidiary-level reporting.
- Using Table 2 makes it possible to identify the number of constituent entities of the multinational group in each country and year. If the number of repetitions of an MNE–country–year triplet in Table 1 matches the number of subsidiaries of that group in the given country and year, it is likely that the reporting was improperly conducted at the subsidiary level.

3.2 Currency harmonization

Currency codes in the dataset are sometimes reported incorrectly or inconsistently. As a general rule one CbCR should be reported only with one currency. In order to conduct coherent analysis the data need to be converted in one single currency.

Potential issue(s):

- **MNE reports value with the wrong currency** Example: Italian company reporting all years in EUR and one year in AED. Solution: identify how many companies report in several currencies, how many report in a currency different from the HQ currency.

- **Multiple currencies across financial variables within a single MNE-year-country entry** Example: For a given MNE-year-country observation, taxes are reported in AED while all other financial variables are in USD. Solution: handle manually if the number of observations is small.
- **Multiple currencies codes within the same CbCR** Example: Multinational reports each jurisdiction in its local currency (USD, GBP, JPY). Solution: flag and standardize.
- **Duplicate entries with different currency codes** Example: Same CbCR reported twice — once with USD and once with EUR, while values are identical. Solution: handle manually if the number of observations is small.

Fix — Cleaning currencies

Step 1: Correct currency reporting errors. Before converting all observations into a common currency, reported currencies should be checked for consistency across years and across financial variables.

Step 2: Convert all data into a single currency (euros). We recommend using end-of-period exchange rates. End-of-period exchange rates are preferred because they provide a more accurate valuation of balance-sheet positions and other stock variables at the reporting date, ensuring consistency between the timing of the observation and its currency conversion. For projects involving cross-country comparisons, we also recommend using exchange rate data from the International Monetary Fund (IMF), owing to its broad international coverage, methodological consistency across countries, and widespread use in empirical research. This enhances both the comparability and replicability of the results.

Suggested descriptive statistics

Item	Questions
3.21	How many MNE groups are reporting CbCR with more than one currency? How many with a single currency?
3.22	How many change currency from one year to the other?

3.3 Sign inconsistencies

Reported values in Table 1 are not always signed consistently. Because CbCR figures are typically extracted from accounting systems that follow their own conventions, the same underlying quantity can appear with opposite signs across MNE groups, and occasionally across jurisdictions within a single group. Sign errors are particularly damaging because they do not merely add noise: they invert the quantity of interest and propagate into any ratio built on it, such as effective tax rates or profit margins. The two subsections below treat the variables most affected: income tax and revenues.

3.3.1 Sign of tax paid and tax accrued

In CbCR data, reported income tax amounts do not always follow the same sign convention across companies. This is mainly because many groups extract CbCR figures directly from accounting systems, where tax is treated as an expense. As a result, some companies report tax payments using negative values and tax refunds using positive values. Under the CbCR framework, however, income tax paid (cash basis) is intended to show actual cash flows: positive amounts indicate cash paid to tax authorities, while negative amounts indicate cash received, such as refunds. Similar inconsistencies can also affect income tax accrued (current year) when current tax expense or benefit is transferred from financial statements without adjustment. To ensure that reported figures are comparable across companies and jurisdictions, the data are standardized to a common convention in which positive values represent tax charges or cash outflows and negative values represent tax benefits or cash inflows.

Fix — Sign of tax variables

The approach relies on the idea that profits and tax should generally have the same sign. In contrast, an opposite sign observed consistently across many countries is more likely to reflect a reporting convention. By combining an overall check at the multinational level with a country-level prevalence test, the method identifies systematic sign inversions.

- **Detection of problematic companies:** Multinationals are flagged for potential sign inversion when both total profit and total tax accrued at the multinational level are non-zero and the following two conditions are met:
 - **Multinational-level sign condition:** A sign mismatch is identified if either
 - * total profit is greater than 1 and total tax accrued is less than -1 , or
 - * total profit is less than -1 and total tax accrued is greater than 1.
 - **Row-level prevalence condition:** Among country-level observations where both profit and tax accrued are non-zero, more than 40% of records satisfy either
 - * profit is greater than 1 and tax accrued is less than -1 , or
 - * profit is less than -1 and tax accrued is greater than 1.
- **Inversion of the sign:** When both the multinational-level sign condition and the >40% row-level prevalence threshold are satisfied, the tax sign is reversed for the multinational.

Alternative detection approach. An alternative to the row/MNE-level rule above is to test whether profit and tax accrued exhibit an implausibly strong *negative* correlation within a country or MNE, which would be the signature of a mechanical sign convention rather than genuine tax variation. This is a useful complementary diagnostic, in particular for detecting inversions that are pervasive within a whole reporting jurisdiction rather than idiosyncratic to a single MNE. As with the rule above, it should be treated as a flag for manual review rather than an automatic correction: even after applying either check, some residual cases will still require a manual decision.

3.3.2 Sign of revenues

Unlike profits or taxes, revenues are conceptually non-negative in standard accounting frameworks. Negative revenue entries are therefore, in most cases, not economically meaningful but instead reflect reporting conventions or data-processing issues. In large firm-level panel datasets, such anomalies typically arise from sign reversals introduced during extraction from accounting systems, inconsistent treatment of revenue adjustments or mechanical parsing errors.

While isolated negative values may occasionally correspond to exceptional accounting events (such as major returns, restatements, or reclassifications) or specific sectoral accounting standard (banking and financial sector), a widespread presence of negative revenues within the same MNE and year is unlikely to reflect genuine economic activity. Instead, it usually signals a systematic inversion of the reporting convention for that firm-year.

It is worth noting that the OECD *Guidance on the Implementation of Country-by-Country Reporting* explicitly states that extraordinary income and gains from investment activities are to be included in “Revenues” (OECD, 2022), and separately clarifies that a jurisdiction-level revenue figure should be reported as *negative* whenever unrelated-party revenues, related-party revenues, or total revenues are negative in aggregate for that jurisdiction — for instance because previously reported sales are returned, or because a negative net item (such as a loss on an investment) exceeds other positive revenues — even where the MNE group’s overall revenues are positive. Isolated negative revenue observations can therefore be genuine, in particular for jurisdictions where a subgroup with substantial investment activity is reported; they need not, by themselves, indicate a sign-convention error. This is precisely why the correction below targets *systematic*, within-MNE-year sign reversals rather than isolated negative values.

Because revenue variables are frequently used to construct ratios (e.g. intra-group intensity, productivity measures, or profit-to-sales indicators), sign inconsistencies can mechanically distort empirical results and propagate errors throughout the dataset. The cleaning strategy therefore focuses on identifying systematic sign reversals at the MNE-year level rather than correcting individual observations in isolation.

Fix — Signs of revenues

The correction relies on a within-MNE-year prevalence rule designed to detect systematic sign inversions while preserving potentially meaningful isolated negatives.

- **Detection:** For each MNE-year, related and unrelated revenues are examined separately. An MNE-year is flagged when more than 70% of entity-country observations report negative values for a given revenue component.
- **Correction:** When the threshold is met, the sign of the corresponding revenue component is reversed for all observations within that MNE-year.
- **Consistency checks:** When both revenue components are predominantly negative, corrected totals are compared with consolidated revenues and, where available, external finan-

cial statements.

- **Residual cases:** Remaining negative consolidated revenues after correction are flagged for manual review.
- **Internal additivity check:** For a very small share of observations, reported `revenue_unrelated + revenue_related` does not equal reported `revenue_total`. Given the negligible incidence, we recommend flagging rather than attempting to reconstruct these observations, and documenting whether subsequent analyses use the reported total or the recomputed sum.

3.4 Standardizing reporting units

In some cases, MNEs report financial data in different units of measurement across years (for example, switching from euros to thousands or millions of euros). To address this issue, a unit-detection and correction procedure is applied.

Potential issue(s): Not all MNEs report their financial variables and number of employees using the same units, which limits the interpretability of the data and reduces comparability across firms.

Fix — Cleaning units

The procedure identifies suspicious year-to-year jumps in key financial variables that resemble changes by powers of ten.

- **Detect "jumps":** Lead and lag ratios are computed for each MNE over time and compared against potential scale factors (10, 100, 1,000, etc.), allowing for a small tolerance to account for rounding or reporting noise. The procedure searches both for unusually low values (ratios close to the inverse of a scale factor, e.g. 1/10, 1/100) and unusually high values (ratios close to a scale factor, e.g. 10, 100). When a ratio closely matches one of these thresholds, it indicates a possible unit shift between consecutive years. From this detection step, a scaling multiplier is derived for each MNE-year observation. This multiplier takes values greater than one when the reported figures appear too small and values below one when they appear too large.
- **Rescaling:** The MNE-year multiplier is then applied to all country-level rows belonging to that MNE in that year. The multiplier rescales the affected values so that all financial variables are expressed in consistent units over time. The process is repeated across multiple sets of financial indicators to refine the correction and ensure internal coherence.
- **Ex-post validation with external data:** This step should be added to the list of checks once the matching with the enterprise database has been completed (see Section 4). Each year, by consolidating the results at the group level, it is possible to compare the aggregates with other MNE-level databases (e.g., Orbis, Compustat). If the comparison of these aggregated figures deviates significantly from 1, approaching a power of ten, either positive or negative,

this suggests a different unit of reference and may require realigning the units to ensure consistency both within the CbCR data and between the CbCR and external databases.

Suggested descriptive statistics

Item	Questions
3.41	Are some numbers more frequent than others?
3.42	Is there a pattern for specific reporting countries?
3.43	How many MNE-year observations are rescaled, and by which scale factors?
3.44	Do unit shifts occur at the same point in time for many MNEs at once, or are they idiosyncratic?
3.45	Are rescaled observations concentrated among MNEs reporting in particular currencies or from particular declarant countries?
3.46	After rescaling, do consolidated CbCR aggregates fall within a plausible range given the EUR 750 million filing threshold?

4 Benchmarking with external data sources

To ensure good data quality, it is important to compare firm-level CbCR data with external databases. This makes it possible to identify residual erroneous values that may have escaped the cleaning process described previously. External benchmarking catches a class of error the within-dataset rules of Section 3 cannot: values that are internally consistent, and therefore invisible to prevalence or ratio tests, but wrong in level for example, a unit shift applied uniformly across all years. It also provides an indication of the representativeness of the sample used.

Two distinct exercises are described below, serving different purposes. The first compares consolidated CbCR aggregates against commercial firm-level databases and is diagnostic of data quality at the level of individual MNE groups. The second compares the cleaned data against the OECD's aggregated CbCR statistics and speaks primarily to coverage and representativeness. The two are complementary and neither substitutes for the other.

A comparison at the level of consolidated multinational groups can first be used to detect potential inconsistencies at the company level. Once the firm-level CbCR data are consolidated by multinational group, they can be matched with commercial databases such as Orbis (Bureau Van Dijk) or Compustat in order to verify the internal consistency of reported figures and identify anomalies across firms. Aliprandi and von Zedlitz (2023) apply this approach to publicly disclosed CbCRs, benchmarking reported figures against the corresponding consolidated financial statements, and find that the two sources broadly agree while documenting differences in variable definitions and in the scope of the reports that complicate comparisons across groups.

At a more aggregated level, the cleaned firm-level data can then be compared with the aggregated CbCR database made available by the OECD. Although this database only provides aggregates by country pair and does not allow verification at the level of individual multinational groups, it makes it possible to identify country pairs that may still contain outlier values. Moreover, since the OECD's aggregated database corresponds to the Country-by-Country Reports of multinationals with annual revenues exceeding €750 million, this comparison also helps assess the proportion of multinationals covered by the dataset being cleaned.

A note on sequencing. Several checks described in Section 3 depend on the matching performed here. The ex-post unit validation of Section 3.4 explicitly requires it, and the filer-switch detection of Section 1.4 is more reliable once external identifiers are available. Sections 3 and 4 should therefore be iterated rather than executed once in sequence: an initial pass of Section 3 makes the matching possible, and the matching in turn refines Section 3.

4.1 Matching with MNE-level data

To obtain external information on the consolidated financial statements of MNEs filing CbCRs, the external database must first be restricted to groups plausibly within the CbCR scope. Note that the reporting obligation is triggered by consolidated revenue in the immediately preceding fiscal year, not

the current one. A strict contemporaneous cut at €750 million will therefore drop legitimate filers whose revenue fell below the threshold in the reporting year, and we recommend applying the filter with a margin (for instance, retaining groups above €600 million in any year of the panel) rather than year by year.

4.1.1 Orbis

External Datasource — Orbis - Bureau Van Dijk

The Orbis database, compiled by Bureau van Dijk, is a global firm-level dataset providing harmonized financial accounts, ownership structures, and detailed corporate information for millions of companies worldwide, and is widely used in research on multinational enterprises and cross-border activity.

Data preparation. The raw Orbis CSV files are first reshaped into a consistent firm-level panel dataset, with harmonized variable definitions and stable firm identifiers across years. The database is then cleaned to remove duplicates and correct inconsistencies in financial variables, using interpolation where appropriate. Particular attention is paid to changes in firm identifiers over time to preserve longitudinal consistency.

Ownership and MNE status. The cleaned data are enriched by merging detailed ownership and contact information to accurately reconstruct corporate group structures (Global Ultimate Owner (GUO) identifier). Additional refinements include the verification and correction of firms' nationality, specific jurisdictional treatments, and the identification of residual or complex duplicate cases to prevent double counting. Finally, the sample is restricted to multinational groups by retaining only firms with at least one foreign subsidiary, ensuring consistency with the CbCR.

Sectors. An important extension using external firm-level data sources is to incorporate information on the sectors in which MNE groups operate. In Orbis, several sectoral classifications are available and can be matched to the micro-CbCR data. However, even at the level of consolidated accounts, sector classifications sometimes reflect the headquarters entity rather than the activities of the group as a whole. As a result, categories such as "activities of holding companies" and "activities of head offices" tend to be substantially overrepresented when using a simple match. This issue should therefore be addressed by relying on alternative sectoral classifications that better capture the group's underlying economic activities.

Potential issue(s): The Orbis-CbCR matching based on TIN can be incomplete.

Fix — CbCR-Orbis matching

- **Proceed to MNE-name matching headquarters country by headquarters country.** Restricting the matching to MNEs sharing the same headquarters country substantially reduces the dimensionality of the problem, making fuzzy name matching faster. The fuzzy-matching procedure shall be applied on clean MNE names. The name filter described in section 1.1 is therefore applied before the name matching. Once the best candidates are retained, all matched pairs **must be manually verified** before being accepted.

- **For the remaining unmatched MNEs, proceed to general MNE-name matching.** The same name standardization filter described above should be applied before matching. Since matching is no longer restricted by headquarters country, the search space is considerably larger and the risk of false positives increases accordingly. Once the best candidates are retained, all matched pairs **must be manually verified** with particular care before being accepted.
- **If any MNEs remain unmatched, use the Orbis web research tool on an MNE-by-MNE basis.** When retrieving firm information from Orbis, **priority should be given to consolidated accounts**, which aggregate the financial activity of the entire corporate group and are thus the most appropriate counterpart to CbCR data. Unconsolidated accounts should only be used as a fallback when no consolidated filing is available and flags should be kept to make sure unconsolidated accounts are not used to benchmark CbCR values.

By way of illustration, Table 3 below reports the matching rate obtained at each step of this sequence on the French exchanged CbC reports for 2016–2019. TIN-based matching alone recovers almost half of the MNE groups. Restricting fuzzy name matching to MNEs sharing the same headquarters country adds the largest single increment, and relaxing that restriction adds a further nine percentage points at a higher risk of false positives. The final manual web search recovers the residual six percentage points and is labour-intensive; teams working under time constraints may reasonably stop after step 3, provided the unmatched groups are documented and their characteristics compared with the matched sample.

TABLE 3
Orbis-CbCR matching rate, by step

Step	Matching criterion	Match rate	Incremental gain
1	Tax identification number (TIN)	48.5%	—
2	Name matching — same headquarters country	67.0%	+18.5 pp
3	Name matching — any headquarters country	76.0%	+9.0 pp
4	Manual web search	82.0%	+6.0 pp

Notes: Rates based on the French-exchanged CbC-reportings (2016–2019). Steps are applied sequentially; each step recovers only firms unmatched by all prior steps. *pp* = percentage points gained over the previous step.

Suggested descriptive statistics

Item	Questions
4.1	Are consolidated profits, employees and assets consistent across datasources?
4.2	What share of MNE groups is matched at each step, and what is the final unmatched share?
4.3	Do unmatched groups differ systematically from matched ones in size, headquarters country or sector?

4.2 Comparison with aggregated CbCR

Comparing the aggregated figures of the dataset with the aggregated CbCR is a useful check to assess the representativeness of the dataset. MNEs active in a given country might not be a random subsample of the universe of the CbCR (typically over-representing very large MNEs active in all countries in the world).

External Datasource — Aggregated CbCR

The aggregated Country-by-Country dataset is assembled by the OECD and based on the aggregation of the micro-CbCR provided by countries.

The data is broken down into three panels:

1. Only positive profit firms MNE subgroups
2. Only negative profit MNE subgroups
3. All subgroups

In addition, an adjusted profit variable is computed on top of the raw profit variable to limit the double-counting of dividends in various countries within the profit variable (see section 5.1)

For representativeness checks (turnover, employees, assets, etc.), use the **all subgroups** panel.

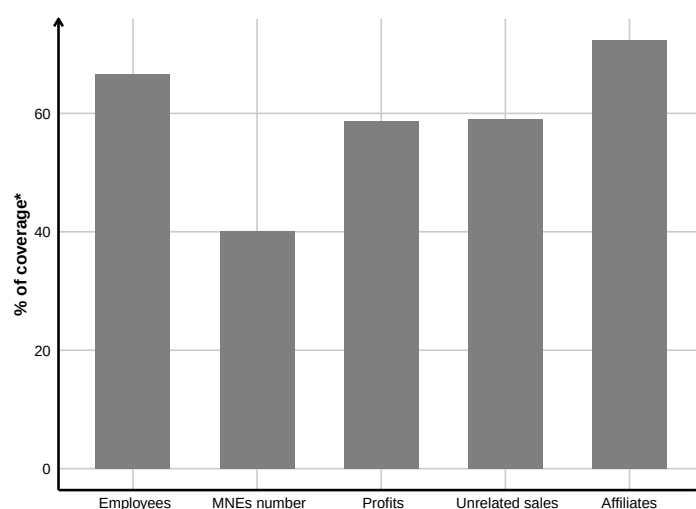
For profit checks, use the **positive profit subgroups** panel.

It is important to note that the OECD Secretariat does not itself consolidate CbCR data, but rather receives them already aggregated from individual countries for their respective MNEs. Consequently, if a reporting error has not been corrected by the national tax administration, the aggregated figures may still contain a degree of uncertainty. Although in the absence of systematic bias, such errors may offset each other through aggregation, this potential limitation should be kept in mind when comparing cleaned firm-level data with the aggregated CbCR statistics publicly released by the OECD.

Figures 1 and 2 illustrate these checks for the French case. Figure 1 compares the exchanged CbCR data with the OECD aggregated statistics across five dimensions, pooled at the world level across years 2016–2019. Coverage is uneven across dimensions: the share of MNE groups covered is the lowest of the five, while coverage of affiliates and employees is substantially higher. The exchanged subsample contains a minority of groups but a majority of activity, because the groups that appear are disproportionately the largest. Any statistic weighted by group counts will therefore behave differently from one weighted by sales or headcount.

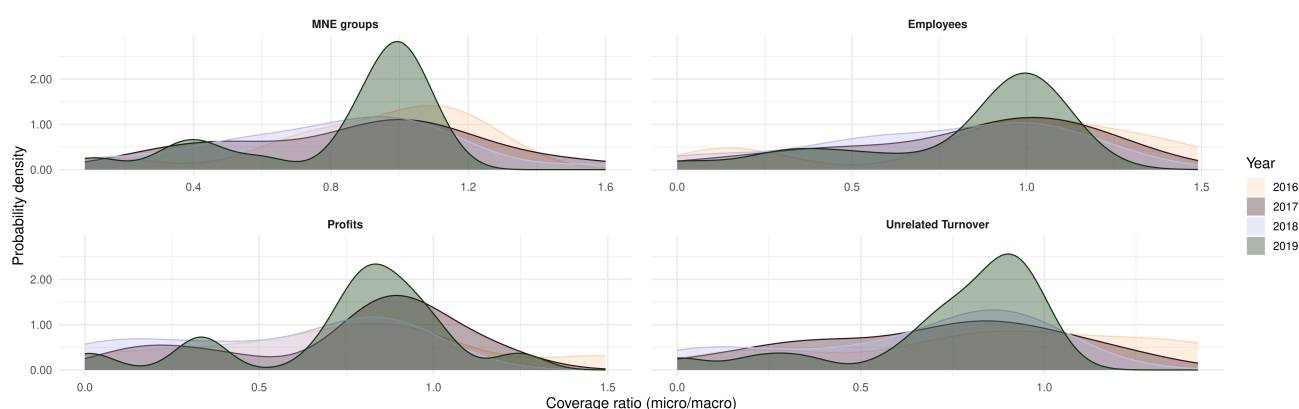
Figure 2 decomposes coverage at the country-pair level, for pairs in which France is the host country, and plots the distribution of the micro/macro coverage ratio by year. It is worth noting that the distributions tighten around a ratio of one over the period, consistent with improving reporting and exchange practices between 2016 and 2019.

FIGURE 1
Coverage of French exchange-CbCR at the world level



Notes: Coverage is measured as the ratio of the aggregate reconstructed from the French exchanged micro-CbCR to the corresponding figure in the OECD aggregated CbCR statistics, pooled across 2016–2019 at the world level. A value of 100% would indicate that the exchanged data cover the full CbCR universe on that dimension. Coverage in terms of the number of MNE groups is markedly lower than coverage in terms of activity, reflecting the over-representation of large, geographically dispersed groups in exchanged data.

FIGURE 2
Distribution of country-pair coverage of French exchange-CbCR, when France is the host country



Notes: Each panel plots the probability density of the coverage ratio (micro/macro) across country pairs in which France is the host country, by reporting year. The ratio compares the aggregate reconstructed from the French exchanged micro-CbCR with the corresponding OECD aggregated figure for the same country pair. Ratios above one cannot reflect coverage and indicate residual measurement error in the microdata, in the aggregated data, or both.

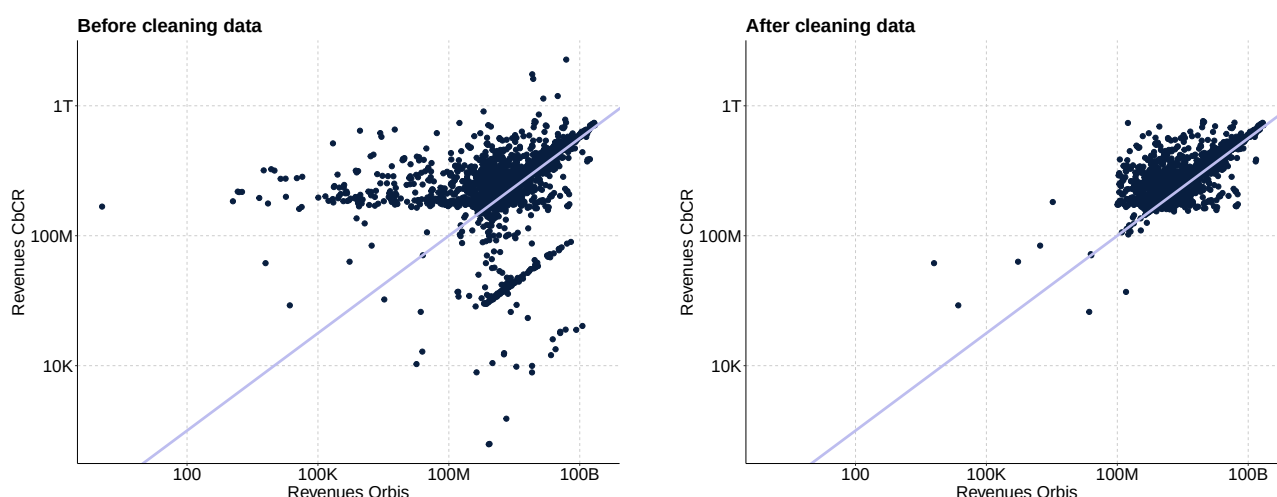
Suggested descriptive statistics

Item	Questions
4.21	What is the share of the universe of CbCR covered by the CbCR exchanged with the tax administration?
4.22	What is the distribution of the coverage of exchanged CbCR at the country-pair level?
4.23	How does coverage differ when measured in number of MNE groups versus in sales, profits or employees?
4.24	Are there country pairs where the reconstructed micro aggregate exceeds the OECD aggregate? What explains a ratio above one?
4.25	Does coverage improve over the years covered by the panel?
4.26	Which headquarters jurisdictions are least well covered in the exchanged data?

4.3 Correct remaining outliers

Comparisons with aggregated CbCR data, whether at the level of parent-company jurisdictions or at the level of country pairs (HQ × destination), make it possible to identify the most significant discrepancies and then verify that the underlying firm-level data do not contain input errors or remaining outliers. It should be kept in mind that the gap observed between the public aggregated data and the aggregates reconstructed from the microdata may also stem from statistical noise in the aggregated data received by the OECD. Therefore, discernment should be exercised on a case-by-case basis when making corrections to the microdata.

Finally, we use Orbis to perform a complementary check for remaining outliers at the MNE-year level. When consolidated CbCR values appear implausible, for instance when consolidated revenues are below €500 million, reported values are corrected using the corresponding Orbis figure where available. Figure 3 illustrates the resulting improvement in the alignment between CbCR and Orbis revenues before and after cleaning.

FIGURE 3**Consolidated revenue in Orbis vs. French exchanged-CbCR, pre- and post-cleaning**

Notes: Each point is an MNE-year. The horizontal axis reports consolidated revenue from Orbis, the vertical axis the sum of revenues across all jurisdictions reported in the CbCR; both axes are on a logarithmic scale and the diagonal marks equality. The left panel shows the raw data before any correction, the right panel the same observations after the cleaning steps described in Sections 3 and 4. The bands of observations lying parallel to the diagonal in the left panel correspond to unit inconsistencies (differences by powers of ten) and largely disappear after correction. Sample: French exchanged CbC reports, 2016–2019, restricted to MNE groups matched to Orbis.

5 Biases, limitations and challenges

5.1 Double-counting of dividends

The issue of double counting is a well-documented limitation of CbCR data and can bias empirical analyses.² More broadly, CbCR relies on jurisdiction-level aggregation of unconsolidated entity accounts, meaning that intra-group transactions are not eliminated. Recent research further emphasizes that these aggregation practices and the inclusion of intra-company flows (e.g. dividends) can hamper interpretation by generating misleading ratios and overestimating economic activity.

This bias should diminish in more recent years. BEPS Action 13 already excluded intra-group dividends from *Revenue* but was silent on their treatment in *Profit (loss) before Income Tax*, leaving divergent practices across jurisdictions and multinationals. It was later clarified that such dividends must also be excluded from profit before tax, for fiscal years commencing on or after 1 January 2020, with a further refinement applying from 1 January 2025 (OECD, 2022). Post-2020 double-counting should thus be substantially reduced, if not largely resolved, though this is conditional on consistent compliance, which has been uneven.

Potential issue(s): Dividends can be double-counted, leading to inflated profit numbers in some sub-groups

Fix — 2-steps procedures to correct for double-counting of profits

The fix exploits a structural difference between the consolidated financial statements and CbCRs. Consolidated financial statements eliminate intra-group transactions, so intercompany dividends, which are distributions of profit already recognized within the group, are removed and not counted twice. Jurisdiction-level CbCR instead aggregates *unconsolidated*, separate-entity accounts and performs no such elimination, so a dividend paid out of a subsidiary's post-tax earnings can reappear in the recipient's profit before tax. Consolidated profit (from Orbis or Compustat) is therefore a "clean" benchmark. An aggregate that materially exceeds consolidated profit, beyond what coverage differences explain, signals potential double-counted dividends.

- **Step 1: Flag MNEs with potential double-counting at the consolidated accounts level**

The first step of the procedure to correct for the double-counting of dividends relies on a comparison between consolidated accounts of multinational enterprises (MNEs) from external data sources (mainly Orbis or Compustat; see Section 4) and the aggregation of financial information at the group level in the CbCR.

A flag is assigned to MNE groups for which double counting of dividends is suspected based on the following test. An MNE group is considered at risk of double counting if the ratio of profits observed in the CbCR relative to the external data source exceeds what would be expected given differences in coverage, as proxied by revenue or employment. Specifically,

²Corporate Tax Statistics, OECD, Box 7.2

a double-counting flag is assigned if:

$$\left(0.8 < \frac{\text{CbCR revenue}}{\text{Orbis revenue}} < 1.05\right) \quad \text{or} \quad \left(0.8 < \frac{\text{CbCR employment}}{\text{Orbis employment}} < 1.05\right)$$

and simultaneously

$$1.05 < \frac{\text{CbCR profits}}{\text{Orbis profits}} < 2$$

In other words, the procedure identifies MNEs for which the coverage of activity in the CbCR appears broadly comparable to that observed in external consolidated accounts (as measured by revenue or employment), while reported profits in the CbCR are substantially higher. Such a pattern is consistent with potential double counting of intra-group dividend income in the CbCR data.

- **Step 2: For flagged MNEs, reallocation of profits among subgroups**

For all MNEs flagged as being at risk of double counting of dividends, the geographical allocation of profits within the group is further analyzed to identify the jurisdictions where such distortions are likely to occur. To this end, effective tax rates (ETRs) and profit margins are computed for each subgroup within the flagged MNEs. The underlying assumption is that when dividends are mistakenly included in profits, they artificially inflate profitability while simultaneously lowering the ETR.

Within each multinational group, the procedure identifies subgroups that simultaneously exhibit unusually high profitability (measured as profits over unrelated revenue) and unusually low ETRs. Concretely, a subgroup is flagged as affected by double counting if it falls within the top 30% of the profitability distribution and the bottom 30% of the ETR distribution among all entities within the same multinational group.

A small but growing literature has proposed corrections for this type of aggregation bias. Working on French exchanged CbCR data, Chiocchetti et al. (2025) document the extent of dividend double-counting in the CbCR profit variable and implement the two-step correction procedure detailed above. In a related but distinct setting, the domestic accounting for the foreign affiliates of U.S. multinationals in BEA data, Blouin and Robinson (2025) show that indirect ownership chains can lead the same underlying income to be counted multiple times when aggregating affiliate-level accounts, and propose a correction that likewise reduces measured profit-shifting magnitudes. Readers adapting the procedure below to other settings may find it useful to compare their choice of thresholds and diagnostics against both references.

5.1.1 Cross-validation with national administrative data

External Datasource — Administrative data

As a third validation step, the corrected CbCR profit figures can be benchmarked against national administrative tax return data, when available.

An example of this cross-check has been done by (Chiocchetti et al., 2025). In their paper based on French CbCR, pre-tax book income reported in French tax returns (BIC data) is compared to both the raw and corrected CbCR profit amounts for French subsidiaries. Before applying the double-counting correction, CbCR profits reported for France are on average larger than the profits reported in domestic tax returns, consistent with the mechanical upward inflation induced by the inclusion of intra-group dividends. After implementing the correction, CbCR profits align closely with the tax return figures, providing strong support for the validity of the cleaning procedure.

5.2 MNE sample bias

Exchanged micro-level CbCR data never represent the full universe of CbCR filings. Indeed, the likelihood that a given host country receives data covering all multinational groups subject to CbCR requirements is low, and this probability decreases with the size of the host country. As a result, the exchanged CbCR data available to any given country constitute a biased subsample of the overall population of CbCR data. These biases can take various forms and should be carefully considered in any statistical analysis based on exchanged micro-CbCR data.

One important source of bias relates to MNE size. Since the largest MNEs tend to have a broader geographic footprint, they are more likely to appear in exchanged CbCR data and are therefore over-represented. This can be observed, for instance, by comparing the coverage rate of the exchanged subsample in terms of the number of multinational groups versus their share of total sales (or employment). Another potential source of bias is sectoral composition. Countries that are rich in natural resources, for example, may disproportionately represent extractive industries in the exchanged data.

Conclusion

This methodological guide proposes a series of procedures for cleaning and processing firm-level country-by-country data. Its central message is that careful cleaning is a prerequisite for any meaningful use of these data, and that the corrections required are systematic enough to be codified rather than improvised.

The guide also carries a more encouraging message. Once processed, micro-CbCR data align closely with other established sources of information on multinational enterprises, whether commercial firm-level databases or the OECD's aggregated statistics. The residual inconsistencies documented here are, for the most part, tractable. This supports the view that CbCR is a rich and reliable source for tax analysis, international macroeconomic statistics, and research on multinational firms.

Statistical analysis based on these data must nonetheless be approached with precaution. As with any dataset, errors and measurement imprecision remain after processing, and several of the corrections proposed here rest on assumptions and thresholds that should be kept in view during analysis rather than treated as settled. Depending on the purpose of the study, users should determine which processing steps are appropriate to apply, and document the choices they make.

For tax administrations, the case for cleaning extends beyond their own analytical use. Carrying out these procedures before aggregating micro-CbCR data for transmission to the OECD directly improves the quality of the aggregated statistics that the whole community relies on, including the benchmarks used in Section 4 of this guide.

This is a first step. The guide would benefit most from convergence in practice: shared thresholds calibrated on more than one country's data, common treatment of the harder cases such as filer switches and dividend double-counting, and feedback from administrations and researchers applying these procedures to panels other than the French one. Greater convergence would improve comparability across studies and strengthen the aggregated statistics in turn.

References

- G. Aliprandi and G. von Zedlitz. Benchmarking country-by-country reports. Working Paper 17, EU Tax Observatory, 2023. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4445272. HAL working paper halshs-04103852.
- J. Blouin and L. A. Robinson. Accounting for the profits of multinational enterprises: Double counting and misattribution of foreign affiliate income. *Journal of Public Economics*, 252:105512, 2025. doi: 10.1016/j.jpubeco.2025.105512.
- T. Boukal, P. Janský, and M. Palanský. Decomposing the revenue implications of the global minimum tax. *International Tax and Public Finance*, 2026. doi: 10.1007/s10797-026-09946-8. Forthcoming.
- B. Bratta, V. Santomartino, and P. Acciari. Assessing profit shifting using country-by-country reports: A nonlinear response to tax rate differentials. *National Tax Journal*, 77(2):349–380, 2024. doi: 10.1086/729293. URL <https://www.journals.uchicago.edu/doi/abs/10.1086/729293>.
- A. Chiocchetti, M. Francois, L. Heidmann, and G. Aliprandi. Shift or share? Anatomy of profit shifting and distributional effects on workers. Working Paper 5637574, SSRN, 2025. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5637574.
- S. Delpeuch, S. Laffitte, M. Parenti, H. Paris, B. Souillard, and F. Toubal. Quel reporting pays par pays pour les futures réformes ? Technical Report Focus 038, Conseil d'Analyse Économique, 2024. URL <https://cae-eco.fr/static/pdf/cae-focus038.pdf>.
- C. Fuest and F. Neumeier. The global minimum tax: Design and economic implications. *EconPol Forum*, 23(6):22–27, 2022. URL <https://www.ifo.de/DocDL/econpol-forum-2022-6-fuest-neumeier-global-minimum-tax.pdf>.
- C. Fuest, S. Greil, F. Hugger, and F. Neumeier. Global profit shifting of multinational companies: Evidence from country-by-country reporting micro data. *Journal of the European Economic Association*, 23(5):1773–1808, 10 2025. ISSN 1542-4766. doi: 10.1093/jeea/jvaf007. URL <https://doi.org/10.1093/jeea/jvaf007>.
- OECD. Guidance on the implementation of country-by-country reporting: BEPS action 13. Technical report, OECD Publishing, October 2022.



Co-funded by the
European Union

Appendix

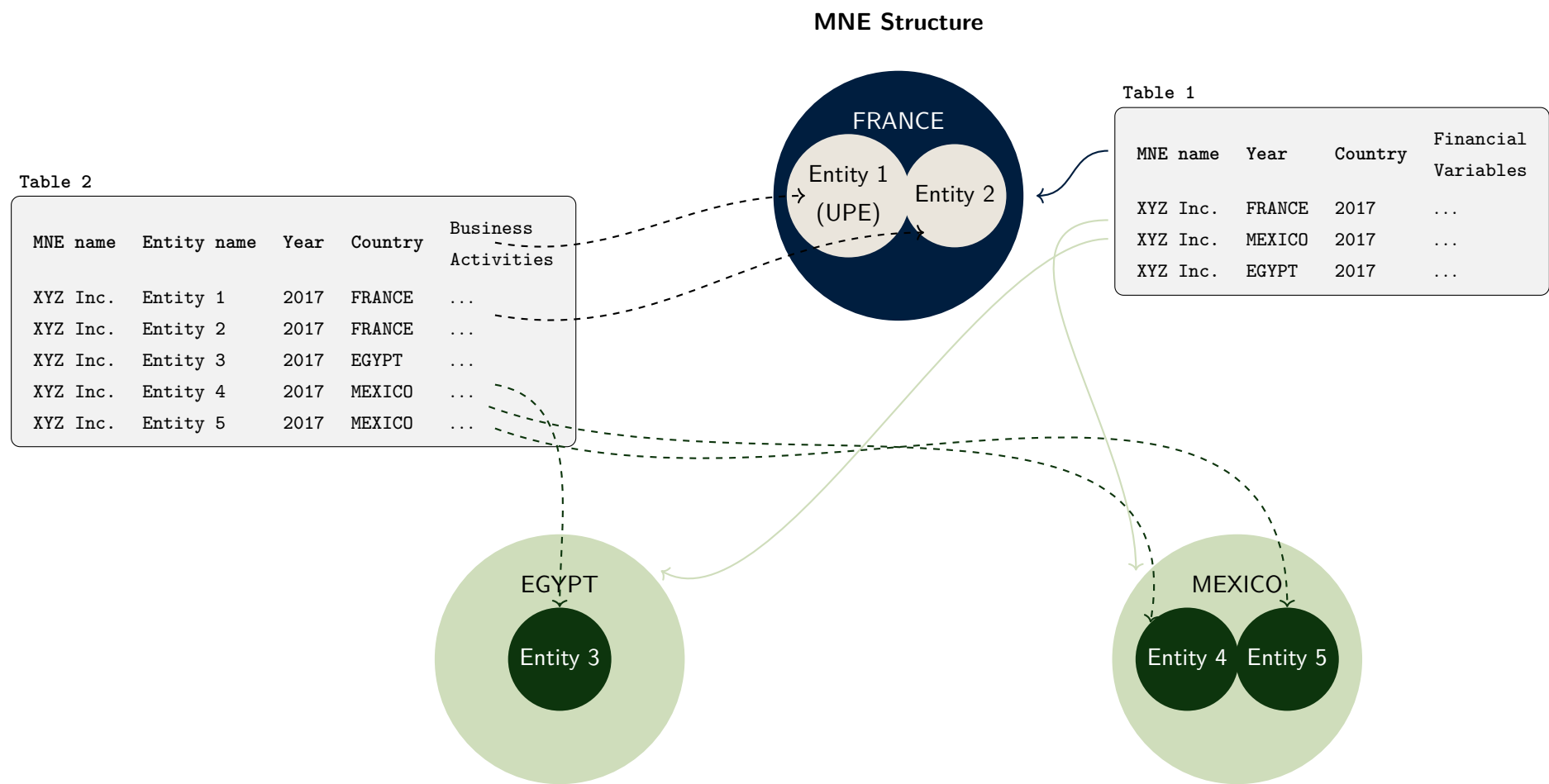


FIGURE 4

Mapping between CbCR country-level data, entity-level listings, and MNE structure. Source: OECD.

Glossary of key terms

The definitions below follow the BEPS Action 13 Final Report, the OECD Model Legislation for CbCR, and the OECD *Guidance on the Implementation of Country-by-Country Reporting*.

Term	Definition
BEPS Action 13	The OECD/G20 action that established the three-tiered transfer-pricing documentation standard (master file, local file, Country-by-Country Report), published in October 2015 as one of the four BEPS minimum standards.
CbCR	Country-by-Country Report. An annual filing by the reporting entity of an MNE group, providing aggregated information on the global allocation of income, taxes and economic activity across the tax jurisdictions in which the group operates. Intended for tax administrations' high-level transfer-pricing and BEPS risk assessment.
MNE group	A group of enterprises whose tax residences span at least two jurisdictions (including via permanent establishments), required to file a CbCR when total consolidated group revenue in the <i>immediately preceding</i> fiscal year exceeded EUR 750 million (or a near-equivalent in domestic currency, fixed at January 2015), as reflected in the group's consolidated financial statements.
Subgroup	The unit of observation in Table 1: the activity of a given MNE group in a given tax jurisdiction in a given reporting year (one row). Data are <i>aggregated</i> across the group's constituent entities in the jurisdiction, without elimination of intra-jurisdiction transactions.
Constituent entity	Any separate business unit of an MNE group included (or treated as included) in the group's consolidated financial statements, plus any permanent establishment of such a unit for which separate financial statements are prepared. Constituent entities are listed in Table 2. Used interchangeably in this guide with <i>affiliate</i> and <i>subsidiary</i> .
Permanent Establishment (PE)	A taxable presence of a constituent entity in a jurisdiction other than its jurisdiction of residence. Listed in Table 2 by reference to the jurisdiction where it is situated, with the legal entity to which it belongs noted separately (e.g. "XYZ Corp – Tax jurisdiction B PE").
Ultimate Parent Entity (UPE)	The constituent entity required to prepare consolidated financial statements for the group, and not itself owned by another constituent entity required to do so. The default filer of the CbCR.

Term	Definition
Surrogate Parent Entity (SPE)	A constituent entity appointed to file the CbCR on behalf of the group in place of the UPE. Where conditions are met, surrogate filing removes the need for local filing in other jurisdictions. The micro-CbCR <i>Reporting Role</i> variable distinguishes UPE filers (CBC701) from surrogate filers (CBC702/CBC703).
Table 1	The main CbCR form. Reports, for each jurisdiction in which the MNE group operates: revenues (unrelated and related party), profit (loss) before income tax, income tax paid (cash basis) and accrued (current year), stated capital, accumulated earnings, employees, and tangible assets other than cash.
Table 2	Lists each constituent entity of the group, its tax jurisdiction of residence (and of organisation if different), and its main business activity or activities, selected from a standardized list of functions.
Table 3	A free-text section in which the MNE group may provide any further information or explanation useful for interpreting Tables 1 and 2 (e.g. data sources, accounting standards, or explanations of apparent profit-substance misalignments).
Double counting of dividends	Inflation of reported profit when distributions received from other constituent entities are not excluded from profit before income tax. Action 13 excluded such dividends from <i>revenue</i> from the outset; explicit exclusion from <i>profit before tax</i> was only added in the November 2019 guidance update, so practice varied in earlier years. This is the source of the correction in Section 5.1.
Aggregated CbCR	Anonymised, country-aggregated CbCR statistics published annually by the OECD, built from the micro-CbCR by tax administrations and transmitted to the OECD Secretariat. Used here for representativeness checks (Section 4).
ETR	Effective Tax Rate; income tax over profit before tax. Used here as a diagnostic for dividend double-counting (Section 5.1).
Orbis	Commercial firm-level database (Bureau van Dijk) of harmonized financial accounts and ownership; used to benchmark consolidated CbCR figures (Section 4). The associated <i>Global Ultimate Owner</i> (GUO) identifier is used to reconstruct corporate group structures.