Organisation for Economic Co-operation and Development

COM/STD/WPTGS/DAF/WGIIS(2018)1

Unclassified **English - Or. English**

1 March 2018

STATISTICS DIRECTORATE
DIRECTORATE FOR FINANCIAL AND ENTERPRISE AFFAIRS
COMMITTEE ON STATISTICS AND STATISTICAL POLICY INVESTMENT COMMITTEE

Working Party on International Trade in Goods and Services Statistics
Working Group on International Investment Statistics

Measuring MNEs using Big Data: The OECD Analytical Database on Individual Multinationals and their Affiliates (ADIMA)

Despite their growing and significant importance, with implications across a range of policy areas, information on Multinational Enterprises (MNEs) remains at best patchy, hindering the formulation of robust evidence based policies.

The **OECD Analytical Database on Individual Multinationals and their Affiliates (ADIMA**) aims to help fill this data gap by providing statistics on the scale and scope of international activities of MNEs, taking a 'whole of the MNE' view. By using innovative methods and data sources (e.g. web-scraping and text analytics), and combining these with traditional financial reporting sources, ADIMA provides a ***Register*** of MNE parent-affiliate structures, a series of economic ***Indicators*** at the level of the MNE and individual countries in which its' affiliates operate; and a ***Monitoring tool*** that aims to provide a timely flow of information on MNEs' restructurings to aid the work of national compilers. International collaboration, in particular with Eurostat, underpins the overall approach. This paper provides an overview of the initiative, and the progress achieved thus far.

Contact: Diana.Doyle@oecd.org, Cecilia.Caliandro@oecd.org, Graham.Pilgrim@oecd.org, Fabienne.Fortanier@oecd.org

JT03427442

Measuring MNEs using Big Data: The OECD Analytical Database on Individual Multinationals and their Affiliates (ADIMA)

1. Introduction

Multinational enterprises (MNEs) have been key drivers of the increasing international fragmentation of production. MNEs account for nearly 30% of value added and trade, and about one-fifth of employment in OECD countries on average, and global Foreign Direct Investment (FDI) stocks now account for around 40% of world GDP (up from 25% in 2005). In developing countries, inward FDI has become the prime source of external funding, surpassing flows of official development assistance, and have provided important mechanisms for these countries to integrate into global value chains.

Whilst international fragmentation of production has helped to generate significant benefits to many, there are concerns that the benefits have not been felt by all, and indeed that part of the phenomena is driven by fiscal optimisation. Not surprisingly, therefore, the implications of MNEs for trade, economic growth, global value chains and inclusive globalisation have been high on the international policy agenda (see Annex A for a more detailed description).

Naturally, this has also led to growing demands for improved statistics that provide insights on the scale and complexity of international MNE activity, and indeed for timely information on any restructurings they may undertake – which was illustrated by the well-documented 26% increase in Irish GDP in 2015 due to the reorganisation of only a few MNEs.

However, despite their significant and growing importance, with implications across a range of policy areas, information on Multinational Enterprises (MNEs) remains at best patchy, hindering the ability to formulate robust evidence based policies. This is partly a function of complexity: by their very nature, MNEs are large, with a multitude of activities across a number of jurisdictions. However, for firms engaging in fiscal optimisation at least, it is also partly a function of design: some firms for example create elaborate chains of affiliates, holding companies and special purpose entities, designed to minimise taxes, but the consequence is also to obfuscate.

Another factor that complicates the measurement of MNEs is the limited possibility for National Statistical Institutes (NSIs) to obtain a holistic view of their activities, reflecting legislation that typically restricts data collections to activities within their economy or (and only very rarely) to firms headquartered in the economy (and even in these cases it is not clear that the coverage of the MNEs activities is exhaustive). The sharing of data across countries could provide a window to provide this holistic view but legal constraints aimed at preserving confidentiality and privacy of respondents within national borders in most countries mean that this is not possible. It is important to note in this respect that the absence of a 'whole of the MNE' view in any given country may generates ample scope for inconsistencies and asymmetries in MNE data (and, so, core macroeconomic statistics such as GDP) across countries.

To begin to address these challenges the OECD has begun to develop an analytical database of individual MNES and their affiliates(ADIMA), by compiling publicly

available statistics on the scale and scope of the international activities of MNEs, thus providing a unique 'whole of the MNE' view. More specifically, ADIMA involves the development of a ***Register*** of MNE parent-affiliate structures; a series of economic ***Indicators*** at both the level of the MNE and the individual countries in which it operates; and a ***Monitoring*** tool that aims to provide a timely flow of information on MNEs restructurings to aid the work of national compilers.

To develop these outputs, ADIMA combines traditional data sources, including commercial databases and company Annual Reports, with newly emerging sources (such as the Legal Entity Identifier (LEI)), and innovative data collection methods and Big Data analytics (such as XBRL, web-scraping and text analytics). The key principle is to leverage all possible data sources in order to effectively overcome limitations in terms of data quality, granularity, and restrictions on dissemination; while at the same time carefully validating results and ensuring maximum alignment with national statistical concepts.

This innovative approach is centred around the goal that, as far as possible, all data that are collected will be made publicly available. The approach has also been developed with a view to scalability. While currently in the 'proof of concept' phase, which will result in complete information for an initial sample of 100 of the largest MNEs by the end of 2018, the aim is to increase coverage to around 500 MNEs in the coming two years.

ADIMA is part of a growing international response in the area of MNE data, which ADIMA builds on and in turn complements. These include Eurostat's EuroGroups Register (EGR) and Early Warning System (EWS), and the work of the Global Legal Entity Identifier Foundation (GLEIF) to create a harmonised identification number (LEI) of all entities worldwide.

ADIMA is also an important step in operationalising the 2015 UN Statistical Commission decision to support the creation of a global register on MNEs and it is hoped will also prove an important resource for the development and improvement of statistics in other areas, such as the WTO's project to creating a Global Trade in Services database by Modes of Supply, and in the development of Extended Supply and Use tables (ESUTs).

This paper explains the objectives, methodology, current results and future applications of the OECD ADIMA initiative in more detail. First, in section 2, the overall structure of the ADIMA initiative is presented. Subsequently, the methodologies and results that have been developed so far are presented in sections 3 (on the register of parent-affiliate structures) and 4 (on the collection and geographical breakdown of indicators of MNE activity), using data collected for 37 US-headquartered MNEs as an example. Finally, section 5 provides an overview of the early conclusions and lessons learned, and outlines the next steps in the project; including amongst others, the further development of the Monitoring tool and geographical disaggregations of additional indicators - including jobs and activities - and the creation of a list of MNEs' Global Decision Centres.

2. The OECD MNE Analysis database (ADIMA): Objectives and Methodological approach

The challenges involved in measuring MNEs at the national and international level have to-date hampered the creation of statistics that provide a 'whole of the MNE' view, and have generated ample scope for inconsistencies and asymmetries in MNE data across countries. This section discusses how the Analytical Database on Individual Multinationals and their Affiliates (ADIMA) aims to address these core challenges, by exploiting Big Data sources and developing innovative methods in order to generate new indictors and insights on MNEs and their international activities. Importantly, as far as possible all data that are developed will be made publicly available, to facilitate their analytical use and further validation, e.g. by statistical offices and Eurostat (the EGR). Indeed, collaboration with existing efforts (Eurostat, RIAD, GLEIF, national Large Case Units) is one of the principles with which the OECD project has been conceived, ensuring a maximum of complementarity of activities, and optimally leveraged resources.

2.1. Objectives and structure of ADIMA

ADIMA aims to contribute to this emerging international need to measure MNE affiliates and activities by developing statistics on the scale and scope of the international activities of MNEs. For a selected universe of the world's largest MNEs, the initiative develops three key outputs, each of which are discussed in more detail below:

- a ***Register*** of MNE parent-affiliate structures, including their geographical and industry classifications;
- a series of economic ***Indicators,*** including key components of balance sheet and income statement, at the level of the MNE and individual countries in which it operates; and
- a ***Monitoring tool*** that aims to provide a timely flow of information on MNEs' restructuring to aid the work of national compilers, as well as to the Eurostat Early Warning System (EWS).

2.1.1. Register of parent and affiliate relationships

The foundation of the ADIMA is formed by a register of MNE Parent-Affiliate structure. This dataset includes, for each parent MNE and all its affiliates, a variety of often-used enterprise identifiers, as well as information on the address (city, country) and industry of all affiliates. For the parent enterprise only, additional information is included on its year of incorporation (to establish MNE age), and consolidated industry classification (following ISIC Rev4), to allow for comparisons between MNEs within the same industry and align with industry classifications used in e.g. outward FATS statistics. The country of the Global Decision Center (i.e. the physical location of the headquarters/executive office) will also be included.

2.1.2. Economic indicators

The second component of ADIMA is a set of indicators that describe the economic structure and performance of MNEs via a selected set of variables reported in Balance Sheets and Income Statement. Examples include an overview of different types of assets (including intangibles), liabilities (short and long-term), revenues and cost structures (e.g. costs of goods sold, R&D expenditures, wages and salaries), different measures of profitability (before and after tax, for example) and employment. All these are collected at the consolidated MNE level.

To capture the extent of MNEs' international activities, ADIMA also includes geographical segment information (for e.g. sales, assets, profits, employment) as reported by MNEs. An important innovation is to disaggregate this data to the country level, using estimations derived from Big Data sources. For example, a country-by-country breakdown of MNEs' sales is developed using sophisticated textual and structural analyses of the geographical profile of MNEs' websites (see also section 4). Other breakdowns are currently being developed.

Finally, indicators derived from the ADIMA Register, such as the number of (foreign) affiliates, or countries in which an MNE is active, are included.

2.1.3. Monitoring tool

The Monitoring tool will provide a timely flow of information on MNEs' restructuring to aid the work of national compilers, again by exploiting the various innovative options that Big Data analytics provides. Using text analytics and other tools to provide structure to unstructured data sources such as news articles, MNE press releases, location changes or job vacancies, the monitoring tool can generate short and relevant information on the latest developments related to an MNE, classified by country and industry, supporting the work of statisticians that profile MNEs.

Given its use of public information and open source data, ADIMA Monitor can also serve as a key input to Eurostat's *Early Warning System*, an information and data sharing system to manage at the European level the organisational changes ongoing in many of these largest MNEs. Ultimately, the monitoring tool aims to facilitate the development of *national* statistics on MNEs and to improve international coherence of the treatment of MNEs and their related transactions in the accounts.

Table 3.1 provides an overview of the key variables and indicators included in each of the three components of ADIMA.

Table 2.1. Variables included in ADIMA: Register, Indicators and Monitor

Register	Indicators	Monitor
All enterprises (parents + affiliates):	Consolidated balance sheet	Changes in register variables, based on text analytics of unstructured big data sources (e.g. Open Street Map, GDELT, MNE IR and Jobs websites), classified by MNE, country and industry
Enterprise Identifiers (ISIN, LEI, CIK)	Assets (current; PPE; intangibles; …)	
Address/country of operation	Liabilities (current; long-term)	
Industry (ISIC rev 4)	Equity	
Immediate and ultimate parent IDs	Consolidated income statement	
Demographic events + dates	Revenues/turnover	
Ultimate Parents only:	Expenses (COGS, SG&A, subcomponents)	
Consolidated industry	Profitability (EBIT, EBITDA, net income)	
Year of incorporation	Employment	
Country of global decision centre	Share of int'l sales, assets, employment, …	
	Country-level sales, assets, employment, …	
	Register-derived indicators (# affiliates, # countries, spread)	
	Entropy measures of internationalisation	

2.2. MNEs covered by the ADIMA

2.2.1. Selection of MNEs

ADIMA aims to cover 100 of the largest global MNEs by the end of 2018, to be extended towards 500 by 2020. These sample sizes were selected to achieve a good balance between relevance (for example, the top 100 or 500 MNEs account for a significant share of FDI[1]) and feasibility (in terms or resources). Such sample sizes also align well with those currently chosen by LCUs in Statistical Offices (the long-established LCU in Statistics Netherlands for example covers ~300 enterprises, while the Italian and Irish LCUs cover 140 and 60 enterprises respectively).

Five selection criteria were created for MNEs to be included in the ADIMA . Enterprises have to rank among the largest firms globally in terms of revenues, to be publicly listed and not state-owned, to be a multinational, and have a certain web-presence:

- ***Revenues***. Companies were selected (top down) from the top largest enterprises globally, ranked by total revenues for 2016.
- ***Stock listing***. Publicly available Financial Reports and Accounts are a vital data source for ADIMA. Therefore, a listing on a stock exchange was considered a pre-requisite for enterprises to be included.
- ***Multinational***. Considering the focus of the project, only enterprises that met the internationally accepted definition of being a multinational were selected – i.e. to have an operating affiliate in at least one country outside the enterprise's home country.

[1] Already 30 years ago, a relatively small set of MNEs (~500) was responsible for the overall majority (80 percent) of global FDI (Rugman, 2005), and little seems to have changed since, as the sum of foreign assets of the 100 largest MNEs equals is estimated at 31% of global FDI in 2016.

- ***Not a State-Owned Enterprise***. SOEs, being as much instruments of government policy as profit-maximising entities, often behave very differently from privately owned firms, and have therefore for now been excluded from the analysis. Given however the growing international importance of these enterprises, particularly from emerging markets (notably China), this decision may be revisited going forward.
- ***Online presence***. Since webscraping and text analytics of company webpages are an important tool for the ADIMA, a significant online presence is required to allow for a meaningful analysis.

No *a priori* selection was made to exclude certain industries (such as finance or insurance). Annex A provides an overview of the initial 100 enterprises selected according to revenues. The selection of MNEs covers enterprises with headquarters in 16 countries.

2.2.2. Current proof of concept: focus on US MNEs

To-date, the affiliate structures and indicators as outlined above have been developed for all 37 US-headquartered MNEs that are part of the selected 100 large MNEs covered in the ADIMA. An important argument to focus on the US MNEs first is the very detailed US FATS statistics that are available, which provide a useful and readily available benchmark to assess the quality of the estimates of the geographical distribution of e.g. foreign affiliates and sales developed within ADIMA.

2.3. Data Sources

To mitigate the paucity of data that has hitherto prevented the development of an international database on the activities of individual MNEs by country and industry, ADIMA leverages *all* available data sources on MNEs. This includes traditional data sources (e.g. commercial databases such as ORBIS and company Annual Reports), newly emerging official sources (such as the LEI and the LEI relationship records), and a whole range of Big Data sources that have only become available in very recent years and that have to date not yet been explored for economic statistical purposes. It is exactly the integration and combination of different sources that is key to maintaining granularity and comparability, whilst maximising dissemination possibilities.

While discussed in greater detail in sections 3 and 4, the analysis of texts and weblinks scraped from MNE webpages with e.g. Natural Language Processing software, in order to obtain more detailed information on MNEs' most important national sales markets, is a first example of how Big Data can be used to provide more granular insights on MNE activities. Likewise, the textual analysis of annual reports for subsidiaries, in combination with LEI information and commercial data, provided the foundation to develop the register of MNE affiliates.

While Annual Reports are not a *new* data source on MNE activities, the requirement by securities regulators in, for example, the United States, Japan and Korea to file these reports in an XBRL format presents an important solution to scaling up the collection of information they contain, while avoiding reliance on the proprietary data from commercial sources. The upcoming change in US regulations that non-US owned MNEs that operate in the US will also be required to submit 20-F forms to the SEC will further accelerate take-up of XBRL.

The techniques presented in this paper are merely scratching the surface of what is possible. Digitalisation may pose measurement problems for statistics, but also presents new solutions. Social media, Open Street Map, new services such as GDELT (global news), job vacancy websites, and many more are currently being explored to develop more insights on MNEs, their international activities, jobs and business functions, and (supplier) relationships with other MNEs.

2.4. Comparison and collaboration with existing efforts

Although new and innovative in its approach, the ADIMA is not the first or only international initiative targeting the improved measurement of MNEs. For example, UNCTAD has for a long time developed a Top 100 of most transnational companies in its World Investment Report, with information on the foreign-to-total ratios of MNE's assets, sales and employment, derived from annual reports. No further breakdowns by country or affiliate industry are provided however, limiting our understanding of the spread or nature of MNEs' activities. Similar databases have in the past also been developed by academics (see e.g. Rugman, 2005; Fortanier and Van Tulder, 2007, 2009). While these included regional breakdowns as well as trends over time, today's technologies that help automate collection and provide additional information on MNEs were not yet available.

The EuroGroups Register is a long-standing initiative by Eurostat aimed at providing the European NSIs with a yearly population frame on MNEs with at least one legal unit in the territory of the European Union or EFTA countries. To create the EGR, Eurostat collects input information on MNE group members and on their relationships from the national statistical business registers of EU countries and participating EFTA countries, and from commercial sources. While an increasingly helpful tool for European FATS compilers, and contributing to the harmonisation of the treatment of MNEs across European statistical offices, the use of the EGR for analytical purposes is unfortunately limited due to the strong confidentiality of the data, and the geographical focus on Europe.

To leverage the work and advantages of both initiatives, OECD and Eurostat have strongly intensified collaboration on this topic in the past year, e.g. by comparing the results of the OECD pilot study with tailor-made higher level aggregates from the EGR, the exchange of technical expertise related to the use of big data, and the coordinated development of the OECD Monitoring tool and the Eurostat Early Warning System.

3. Developing MNE affiliate structures

3.1. Introduction

A full overview of MNEs' affiliates, their location and activities forms the foundation for all further statistical development towards better understanding the international scale and scope of the activities of MNEs. ADIMA therefore maintains a register of affiliates and their immediate and ultimate parents, including information on their geographical location and industry activities, following as closely as possible the international concepts related to classifications and ownerships thresholds. This section discusses these conceptual considerations in more detail before describing how they are dealt with empirically in ADIMA, and presenting the first results on the pilot sample of 37 US MNEs.

3.1.1. MNEs and affiliates in the register

The register of multinationals, their affiliates and relationships is constructed to align as much as possible with the 2008 SNA and BMD4 accounting frameworks. It includes all domestic and foreign controlled affiliates of the MNE Ultimate Controlling Parent, using an ownership threshold of 50% (defined as control) to decide whether to include an affiliate in the register.

As such, the multinational enterprise groups included in the database are more narrowly defined than in BMD4, which also includes enterprises under the influence (i.e. minority holdings) of the same ultimate owner.

Whilst acknowledging that minority shareholdings may be of analytical interest, and may, in combination with other constructions (e.g. large-scale contracts; processing activities; licencing agreements related to the use of intellectual property), still de facto result in control, the 50% threshold is both a practical measure and ensures full alignment with official FATS statistics, which are recommended to be compiled on this basis (see BMD4).[2]

There is no official recommendation in BMD4 as to whether the enterprise or establishment is the preferred statistical unit for data collection; instead it is recommended that this is aligned with other national statistics. It is clear however that either approach will influence the interpretation of results, with particular implications for industry attribution, for example. For ADIMA, the data were compiled at the enterprise level, partly due to a lack of data at the establishment level, and partly to reflect that the enterprise is the level at which data are or can be compiled in most OECD (and non-OECD) countries.

[2] Note that this is required in BMD4 at all levels of the chain.

3.1.2. Attribution of geographical location and industry

One important component of the analysis of MNEs is the attribution of the MNE, as well as of each of its affiliates, to a certain country and industry, which combine to form a register describing the ownership and operating structure of the MNE. Annex B provides a list of the MNEs including their geographical and industry classification.

Geographical classification

According to the 2008 SNA, BPM6 and BMD4, the geographical classification of institutional units should be made according to their residency, i.e. their centre of predominant economic interest or the economic territory with which a unit has the strongest connection.

The geographical attribution of the residency of the ultimate parent (the Ultimate Investing Country, UIC) is again determined following international standards, including in particular the recommendations by the Eurostat FATS Working Group on four types of 'special cases'[3], aimed at harmonising treatment within the Euro-Groups Register. There are cases in which the UIC may differ from the country of registration as well as from the country of the principle executive office (Global Decision Centre). These addresses are currently maintained within the database, although at the country level these coincided for the 37 US MNEs included in the pilot.

Activity classification

The MNEs and affiliates that are included in ADIMA are classified to activities following the ISIC industry classification, and the 2008 SNA recommendation that the principal activity of a unit is 'the activity whose value added exceeds that of any other activity carried out within the same unit'[4]. In practice, product sales are often used instead as value added by product is more difficult to achieve, and in a hierarchical fashion (e.g. affiliates are not assigned to a four-digit industry outside either its primary two or three digit sector, see for example BEA, 2014). As described in more detail below, this approach (i.e. using the primary source of revenue as the determinant) is also used in the main data sources that are used to populate the parent-affiliate structures of ADIMA.

3.2. MNE-affiliate structures: Data sources and methodology

To date, there is no publicly available database of corporate ownership linkages at a global level. At the national level, several countries have started to share firm-level records with a view to further their open data agenda. For example, INSEE (France) has opened its Sirene database[5] and in the United Kingdom, individuals and legal entities with significant control of UK companies should be identified on the register of Persons

[3] (i) natural persons and families; (ii) units in tax havens, in offshore financial centres, SPEs, non-profit institutions; (iii) dual-listed companies; and (iv) joint ventures.

[4] SNA 2008, para 5.8

[5] The database "Base Sirene" from INSEE is an open register on enterprises and establishments. The information on the headquarter location is detailed for French locations, but is limited to being classified as 'Foreign' otherwise. See also: https://www.sirene.fr/sirene/public/variable/rpen

with Significant Control.[6] It is expected that more countries will follow, particularly in Europe, in response to EU regulation[7], and, so, the pool of open data sources will progressively expand. However, at the time of drafting, the available datasets represented only a fraction of the global networks that Multinationals have built over the years.

Thus, while the overarching goal of ADIMA is to construct the multinational parent-affiliate register solely with publicly available data, the current limited data availability implies that information from a variety of sources (commercial data, regulatory data, as well as the newly-formed LEI relationship records) need to be combined in order to maintain granularity and comparability, whilst maximising dissemination possibilities.

3.2.1. Commercial sources

The Orbis database by Bureau van Dijk was one of the first data sources considered to develop the register of MNE affiliates. While the drawbacks of Orbis are well-known and documented (Ribeiro *et al.* (2010) and Gal (2013)), it remains one of the largest cross-country databases of enterprise information compiled from a large variety of private and public data providers, with near-global coverage (Kalemli-Ozcan *et al*., 2015). One disadvantage particularly pertinent to the construction of enterprise registers is the lack of harmonisation of the unit of observations. For example, data for some countries such as France is organised around enterprise units, which in turn manage establishments; data for other countries such as the United States do not allow for straightforward identification of the decision making units and their respective branches/establishments[8]. To mitigate this issue, a conservative approach was chosen and all types of branches (i.e. whether these could be identified as being dependent on a specific decision making enterprise unit within the MNE hierarchy, or directly to the ultimate parent) were excluded. Only those affiliates where the ultimate parent owned more than 50% were retained.

The affiliates were geographically attributed using the country of registration.[9] The industry classification in Orbis follows the NACE Rev.2 classification at the 4-digit level, which was converted to ISIC Rev.4.[10]

[6] Persons with Significant Control (Companies House): since mid-2016, UK companies are required to file information over individuals who have direct or indirect influence on them; in addition, UK companies and certain non-UK listed companies may also be registered as Relevant Legal Entities. See also: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/621566/170623_NON-STAT_Guidance_for_PSCs_4MLD.pdf

[7] Article 30 of the 4th EU Anti-Money Laundering Directive (4AMLD), for example, requiring all EU Member States to put into national law provisions around beneficial ownership information for corporate and legal entities.

[8] For example, for Walmart, the number of US subsidiaries dropped from 12,000 to 500 once the "branches" were removed.

[9] Note that residence is defined in official FATS recommendations as the country of registration, and can be different from nationality. Within Orbis, the country of registration can be inferred from the ISO code prefix in the entity identifier. In addition, address information is provided separately. The country inferred from the address is identical to the registration country for all but a handful (0.4%) of observations in the sample. These are cases for which the entity had to acquire a registration number to trade in that country, but no other demographic information such as address, or legal form can be attributed to that country. Further research is planned to assure the

3.2.2. Company reports and regulatory submissions

Listed companies are typically required by regulation to make all relevant information on their activities and performance publicly available on a regular basis, for example via independently audited quarterly and annual reports. These documents contain financial statements, segment information as well as subsidiary data at various levels of detail, and therefore form an excellent primary data source for statistics on MNEs.

The exact reporting requirements and standards differ across countries. For example, all companies listed on regulated markets within the European Union are required, according to the EU Transparency Directive[11], to submit, amongst others, consolidated financial statements according to accepted accounting standards on a yearly and half-yearly basis, as well as major (changes in) holdings of voting rights. Outside Europe, country-level jurisdiction also often imposes the complete listing of subsidiaries and related entities that form part of the scope of consolidation (in full or by equity method), including joint operations and joint ventures, and so called 'immaterial' subsidiaries, which are accounted at cost.

In the United States, listed companies are required to submit a number of regulatory reports to the SEC. Among these, the annual "10-K" submission to the SEC includes a list of 'significant subsidiaries', which lists all subsidiaries that account for more than 10% of assets or income[12]. These lists thereby form a high-quality, but not fully exhaustive source of data on US MNEs' affiliates, depending on consolidation strategies. For example, the data for the 37 US MNEs included in the initial analysis showed that Apple declared to the SEC only four significant subsidiaries in 2017 (one in the United States, and three in Ireland);[13] while General Motors submitted details on 250 significant subsidiaries, of which more than half are located in foreign countries.[14]

In addition to coverage, informing countries on the importance of the affiliate from the perspective of the MNE is a priority. The 'significant subsidiaries' is therefore of value in its own right as a mechanism by which to weight the role of those affiliates in the respective countries in which they operate.

However, it should be noted that information on SEC filings are also a prime source for basic business information (company name and address, industry SIC[15] code) as well as economic variables (e.g. assets, sales, see also section 4), which can all be automatically extracted from the SEC in XBRL format. ADIMA uses the SIC codes in SEC filings to

quality of geographical attribution, using alternative open data sources such as OpenCorporates, DBPedia, and Thomson Reuters PermID.

[10] More details on this are provided in Annex B.

[11] Directive 2004/109/EC, later amended by 2013/50/EU

[12] Conditions for significance are described in SEC Regulation S-X, Rule 1-02 (w).

[13] In contrast, Orbis data suggest Apple controls more than 140 entities in over 20 countries.

[14] An additional issue is the decrease in disclosed subsidiaries over time, which has been extensively documented, see Gramlich and Whiteaker-Poe (2013).

[15] The Edgar system still uses US SIC codes, in order to maintain continuity, but the US SIC classification was superseded by NAICS in 1997.

classify US parents by industry, converting these via several steps into the ISIC Rev. 4 classification, as explained in Annex D.

3.2.3. LEI relationship records

The third source consists of Legal Entity Identifier data. The Legal Entity Identifier (LEI) is a 20-character reference code to uniquely identify entities engaged in financial transactions. This identifier is supported by the 'Global Legal Entity Identifier Foundation' (GLEIF), an initiative launched in 2011 by the Financial Stability Board (FSB), mandated by the G20. In addition to providing firm-level identification information ('level 1 information'), the entities are required to declare the immediate and ultimate parent upon registration ('level 2 information'), and are expected to confirm or update this information at least on an annual basis. Within the LEI relationship information, the "ultimate accounting consolidating parent" is defined as the highest level legal entity preparing consolidated financial statements (LEI ROC, 2016), which in principle aligns with the concepts and definitions in BMD4.

The LEI relationship data has been made public, gradually, since 2017 Q2. As of 12 February 2018, almost 1.1 million LEIs have been issued to legal entities globally, of which around 60 percent also reported information on direct and ultimate parents (up from 26 percent only a quarter prior).[16] The growth in uptake of LEI was particularly marked in 2017, as the LEI population nearly doubled in response to EU regulation. The latest information on renewal rates shows a 70% rate for the European Union and 52% for non-EU countries.

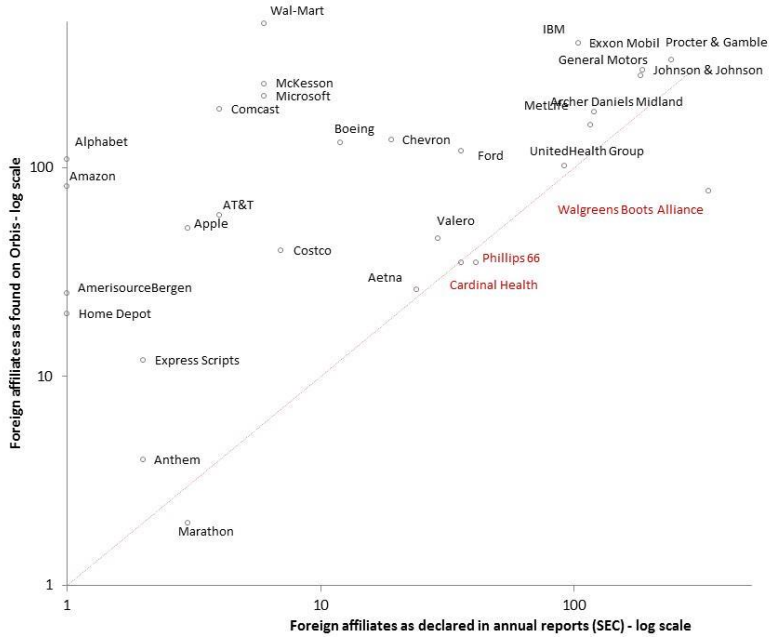
As the LEI and the LEI relationship datasets have only recently started to be collected, their coverage is still insufficient for the construction of company ownership hierarchies. It remains unclear at present how many affiliates will acquire an LEI, and to what extent the data can be used as the *sole* data source for affiliate hierarchies. For this reason, validation work incorporating the LEI data should be seen as an assessment of progress to date, as opposed to the overall quality of the source. Importantly, however, further disclosure requirements are being considered, including for example the inclusion of the LEI of subsidiaries in SEC submission, which would support further international adoption of the standards (SEC, 2016).

3.2.4. Comparing affiliates in Orbis, LEI and company annual reports

Comparing the three data sources discussed, Figure 4.1 shows that for most US MNEs included in the pilot study, commercial data reveal a far larger number of foreign subsidiaries than reported in annual reports submitted to the SEC, which can be attributed to the requirement to 'only' disclose 'significant subsidiaries'. For example, approximately 500 foreign affiliates of Walmart are identified in Orbis, whereas only 6 were declared to the SEC. However, some clustering is present around the 45°-line, implying that many multinationals disclosed most or all their foreign affiliates in their annual reports submitted to the SEC. There are three MNEs for which the Annual Reports appear to be a richer source of information than Orbis in terms of number of subsidiaries identified: Walgreen Boots Alliance, Phillips 66 and Cardinal Health.

[16] Information extracted from LEI website

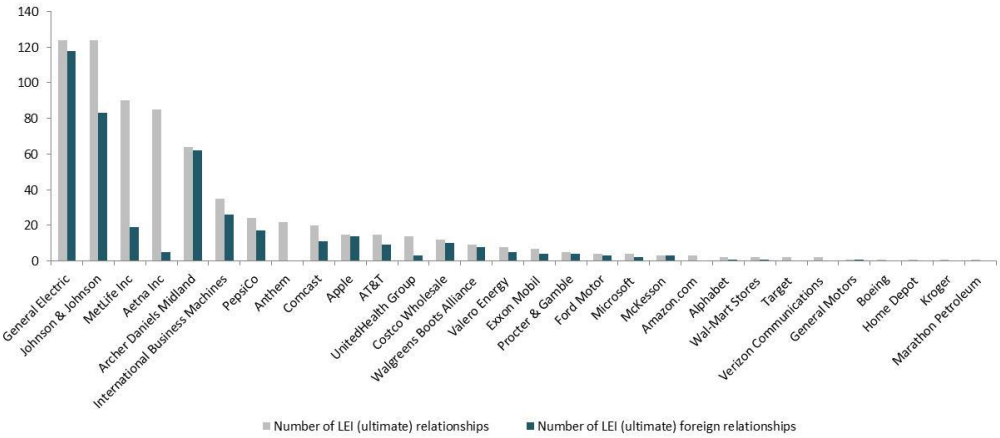
Figure 3.1. Comparison between Orbis and annual reports, number of foreign subsidiaries



Note: Verizon, Kroger, Target, CVS Health, and Lowe's are not included (0 foreign affiliates in SEC annual reports). General Electric is also not displayed (>1000 subsidiaries on Orbis/67 on SEC).
Source: Orbis and Annual Reports (as reported to SEC)

Figure 3.2 presents the numbers of affiliates per MNE with LEI relationship data. As mentioned above, the LEI is not yet a sufficient source for information on the entire network of MNE affiliates. Some of the largest global groups do not yet have any relationship filed at time of drafting, or very few, such as Wal-Mart and Amazon.

Figure 3.2. LEI ultimate consolidation relationship data: total and foreign affiliates



Note: Data as of February 2018. Zero values for AmerisourceBergen; CVS Health; Cardinal Health; Chevron; Express Scripts; Lowe's and Phillips 66 have been omitted.
Source: LEI

3.2.5. Combining affiliates from different sources

To fully leverage all available information, the affiliates identified by each of the three sources were combined into one single database, using a fuzzy matching procedure. Fuzzy matching (or more formally, approximate string matching) is a technique that matches records based on (their degree of) textual similarity. The OECD MNE-affiliate structures were created by using company names and the countries and cities of registration in each of the three sources.

One of the main challenges of fuzzy matching is to fine-tune the algorithm by deciding on the degree of record similarity that is required to consider the records matched. As Annex B describes in more detail, various cut-off values were tested, and a final 99% match was used as a threshold.

The matched values enabled the deduplication of records, while unmatched records from either source were retained.

For the sample of 37 US MNEs that are part of the current pilot, the integration of the three sources yielded an overall affiliate sample of more than 20,500 entities. Table 3.1 provides an overview of the sources.

Table 3.1. Overview of fuzzy matching results

Total affiliates	**20,500**
From Orbis	15,500
From company reports	1,200
From LEI	400
From Orbis and LEI	200
From Orbis and company reports	3,200
From Orbis, company reports and LEI	130

Note: Numbers provided are preliminary and subject to further quality assurance and addition of newly added LEI relationship data.
Source: OECD ADIMA.

Geographical distribution of affiliates

The geographical distribution of the examined MNEs is displayed in Table C.2 in Annex C and Figure 3.3 below. Canada is the most prominent location for US affiliates abroad, with more than 800 US affiliates, of which more than 170 are from Wal-Mart. This is closely followed by the United Kingdom where the total number is just above 800, of which more than 200 are controlled by General Electric. The Netherlands and Ireland follow with around 400 enterprises each. Interestingly, there is some marked concentration of affiliates in certain countries for some firms: for example, the Netherlands are a key investment partner for Exxon, with more than 100 active subsidiaries, and IBM, with 42. Likewise, General Electric and McKesson have many affiliates in Ireland.

Due to the commercial nature of the data, granular industry information at subsidiary level cannot be published at present[17]. However, synthetic measures based on less

[17] Work is in progress to identify suitable open data alternatives, such as Thomson Reuters PermID data.

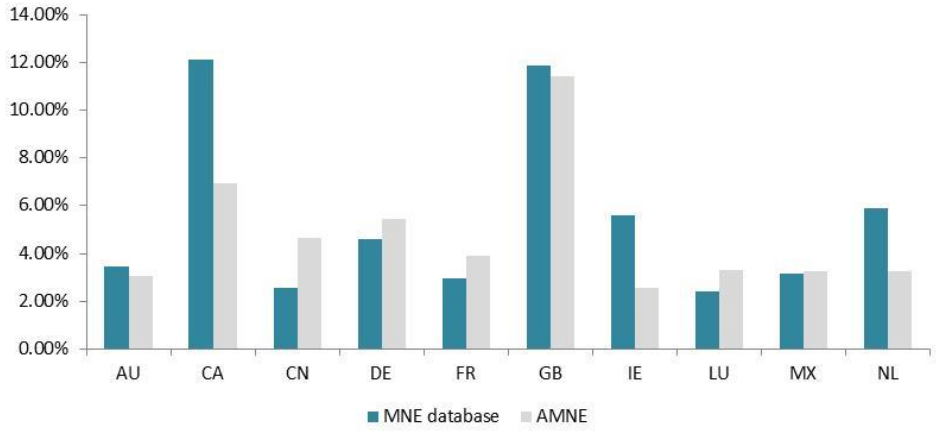
granular information can be created to obtain a view of the extent of integration and industry dispersion of the MNE group as a whole.

3.3. Data validation: Comparison with US outward FATS statistics

To provide a first assessment of the validity of the results obtained, the geographical distribution of the affiliates of the sample of 37 MNEs are compared to those provided by US FATS statistics, focusing on the ten most important partner countries in US outward (enterprises) in US FATS statistics, which alone account for 45% of foreign outward affiliates.

This set of 10 partner countries in FATS is exactly the same as for the 37 MNEs. However some countries' weight in the total are different. In particular, Canada, Ireland and the Netherlands are over-represented in the pilot set of 37 US MNEs, while the People's Republic of China (hereafter 'China') appears under-represented. Table C.2 in Annex C, already discussed above, indicates that this may be partly be due to only a handful of enterprises. For example, excluding Wal-Mart from the equation reduces the differences in the two sources with respect to the importance of Canada by more than a third. A larger sample, as planned for ADIMA going forward, would likely result in a closer match with US outward FATS statistics. In addition, it is important to note that US FATS statistics on affiliates only include affiliates with assets, sales, or net income greater than $25 million. ADIMA does not include such a cut-off, and it may be possible that smaller foreign affiliates, in particular, are located more 'closely to home' (Canada as opposed to China) as they can benefit from corporate support functions at headquarters. The high level of affiliates in Ireland and the Netherlands likely reflect at least in part the presence of SPEs.

Figure 3.3. Share of 10 key partner countries over total foreign affiliates, ADIMA vs US outward FATS



Source: OECD ADIMA and BEA US FATS 2015

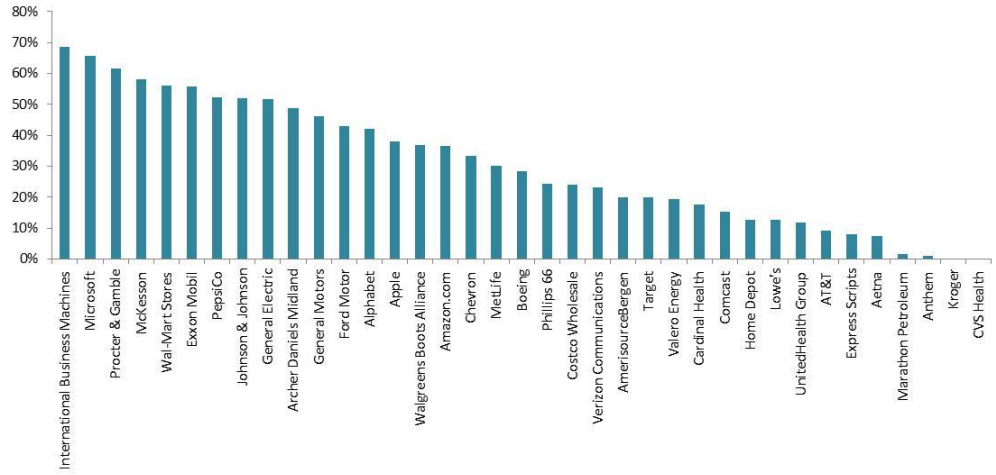
3.4. Indicators of MNE international activity

The internationalisation of MNE activities, including their affiliate networks, has long been studied in the academic literature on international business. While most measures were first operationalised in the mid-1990s (see e.g. Sullivan, 1994), they continue to be heavily used, as evident for example from the recently published meta-analysis by Marano *et al.* (2016). Many of these indicators aim to assess the *scale or degree* of internationalisation, which is typically calculated as the share of foreign activities in total activities, where activities may constitute sales or assets (most commonly), but also the number of employees or subsidiaries. These may be either combined in a composite index (Sullivan, 1994; or UNCTAD's TransNationality Index), or used as separate dimensions (Ramaswamy et al., 1996). Other indicators target the *scope or breadth* of international activities, reflecting a more or less sophisticated indicator of the number of countries or geographical regions in which MNEs are active, such as the Network Spread Index (Ietto-Gillies, 1998), or entropy measures of international diversification (Hitt et al., 1997). This section provides several illustrations of the types of indicators on the scale and scope of MNEs' international activities that can be calculated using the register of MNE affiliates, using the 37 US MNEs included in the pilot as an example.

3.4.1. The degree of internationalisation

One of the primary and simple measures of internationalisation is represented by the Degree of Internationalisation (DoI), calculated here as the share of foreign affiliates in total affiliates. Figure 3.4 shows the results for the 37 US MNEs currently included in the pilot study, indicating that even among the world's largest MNEs, important variation exists with respect to the extent of their international exposure and the relevance of international activities for their strategic decision making.

Figure 3.4. Degree of Internationalisation of affiliates

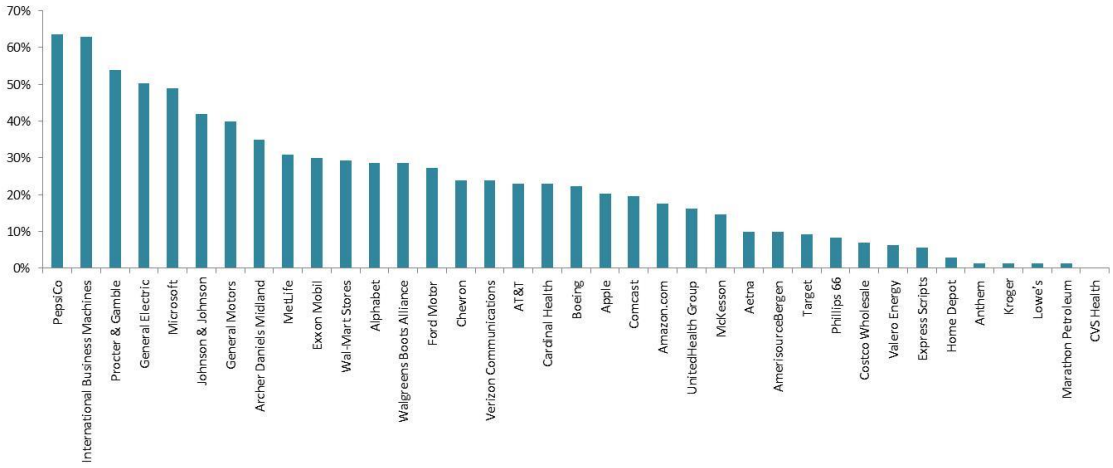


Note: CVS Health appears to own a number of subsidiaries in Brazil according to the Exhibit 21. This relationship is detected in Orbis but deemed inactive.
Source: OECD ADIMA.

3.4.2. Network Spread Index

A straightforward measure for the scope or spread of a multinational is the network spread, which is calculated as the number of countries where affiliates are present as a share of number of potential countries (Ietto-Gillies, 1996). The results are presented in Figure 3.5, showing that the most 'geographically spread' MNEs appear to be PepsiCo and IBM, with a presence in over 86 countries (or 60% of the total number of (144) countries in the dataset.

Figure 3.5. Foreign Affiliate Network Spread Indices, by country



Source: OECD ADIMA.

3.4.3. Entropy measures of geographical diversification

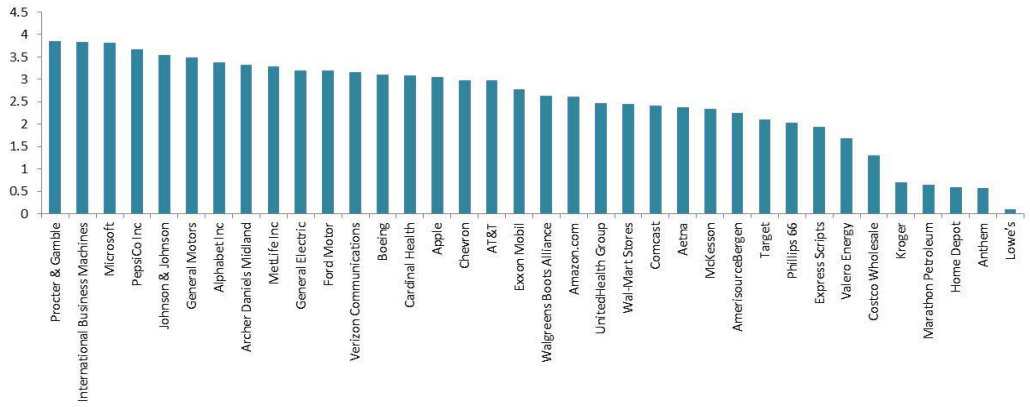
Geographical diversification was calculated using an entropy index as conceptualised by Jacquemin-Berry, which takes account of the number of countries in which an MNE operates as well as the relative importance of each country within the multinational group (see also Hitt et al., 1997). It is calculated as:

$$International\ Diversification = \sum_j p_j ln(1/p_j)$$

where p_j reflects the number of affiliates in country j.

Entropy measures have also been used extensively as a measure for product or industry diversification in similar literatures. Intuitively, the measure is a weighted average of each country-affiliate shares over the total affiliates and takes into account both the number of countries and the relative importance of each, in terms of affiliate counts. The results of the calculations (excluding domestic firms) are presented in Figure 3.6.

Figure 3.6. Geographical diversification

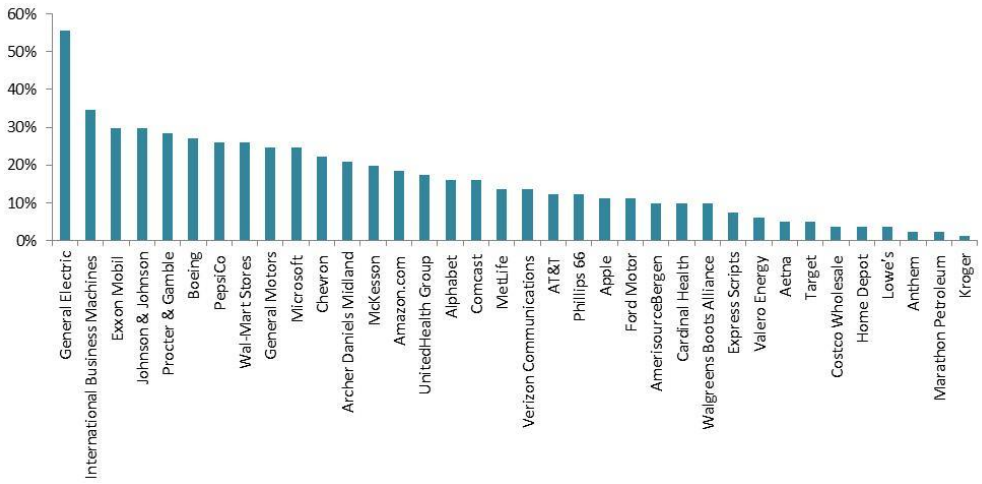


Source: OECD ADIMA.

3.5. Industrial diversification

Using the information on the activities (industries) of affiliates, a similar analysis on the scale and scope of international activity can be conducted to provide further insights on the extent to which MNEs' activities are diversified across industries. While currently still conducted at a relatively high (2-digit) level of aggregation, Figure 3.7 shows the number of industries in which an enterprise is active (as a share of the total number of 2-digit industries in the database), indicating that, in particular, General Electric, and, to a lesser extent, IBM appear to be the most dispersed, with operations in more than 40 industries.

Figure 3.7. Industry network spread of foreign affiliates



Source: OECD ADIMA.

4. Economic indicators on MNE activity

4.1. Introduction

In addition to providing a register of MNE-Affiliate relationships, the OECD ADIMA database aims to provide a series of key economic ***Indicators*** on MNEs, including the main components of profit and loss statements, balance sheet information, and employment. Ideally, many of these indicators are also broken down geographically - in any case between a 'foreign' and 'domestic' component, but preferably to a more granular (country) level - in order to provide insights into how the largest enterprises worldwide organise their production internationally and affect value added, trade and employment in the countries in which they are active.

This section first discusses the collection of these indicators from publicly available data sources, automated where possible (Section 4.2). Subsequently, a first example is presented to disaggregate these indicators to countries, focusing on sales. Using company websites – an important marketing channel – as a key source of information, a variety of Big Data analytical techniques is used to transform textual information and weblinks to a very plausible predictor of country-level sales for the 37 US MNEs in the current pilot study (Section 4.3). The estimated country level measures are validated amongst others via a comparison with official US outward FATS data (Section 4.4), before a first set of internationalisation indicators are presented (Section 4.5).

4.2. Economic indicators of MNE activities

The primary source of information on the consolidated activities of companies is Annual Reports. Disclosures of particular interest are the Balance Sheet Statement and Income Statement. The Balance Sheet provides an overview of the value of assets and liabilities at a given date, whereas the Income Statement shows the revenues and expenses (including non-cash items such as depreciation) and profits/losses over a given time period. For publically listed companies, the disclosures are in general independently audited, and the auditor gives a classification of whether the disclosures are a true and honest reflection of the underlying business.

The financial statements in Annual Reports are typically prepared following agreed accounting standards. The International Financial Reporting Standards (IFRS) provide the most widely used internationally comparable accounting standard; however companies may also use the Generally Accepted Accounting Principles (GAAP) (for example, US listed enterprises are required to submit their accounts to the SEC using US GAAP). IFRS and GAAP principles have seen convergence but differences remain (see e.g. EY, 2013). ADIMA classifies the reporting regime used, making users aware of the differences, however as of yet has not attempted to harmonise between accounting regimes. As recognised in the 2008 SNA, commercial accounting standards are typically well-aligned

with National Accounts concepts, even if some differences exist[18] or adjustments have to be made[19].

While clearly a rich source of information, the extraction of data from Annual Reports in an automated fashion is complicated, as companies may publish their reports in a variety of formats (PDF, websites) that are not optimised for this purpose. However, as noted in Section 2, securities regulators in several countries, including the United States, require that companies submit their annual reports using XBRL (which entails that data points are tagged with a harmonised coding), which facilitates the extraction of key data items. Some manual data entry remained necessary however for this pilot study, for example for variables on employment and certain parts of the data on geographical segments.

Table 4.1 presents a selection of the variables collected for the 37 US MNEs in ADIMA, with manually entered data classified in grey. All companies report following US GAAP standards, and, as indicated in the table, the data reflect reference year 2016. For MNEs whose Fiscal Year does not coincide with the calendar year, ADIMA treats the results for the Fiscal Year with the greatest overlap with the calendar year as reflecting that calendar year.

Finally, it should be noted that variation exists with respect to the types of assets or sales that are reported with a geographical disaggregation. For example, for assets, companies may report either Non-Current Assets; Property, Plant and Equipment; Fixed Assets, or another form. For sales, it is unfortunately much less clear if the geographical segments are reported following the location of the sale, or the location of the final consumer, and neither GAAP nor IFRS prescribe a specific method for the attribution of sales. MNEs are however required to declare which variables are reflected in geographic attributions (EY, 2017) and as a result the OECD database includes a flag that records this methodology for future analysis purposes.

Table 4.1. Selected economic Indicators for 37 US MNEs for 2016

	Revenues (USD bn)	Assets (USD bn)	Employment	International Sales	International Assets	International Employment
Wal-Mart Stores	485.87	198.83	2,300,000	24.3%	27.5%	34.8%
Exxon Mobil	218.61	330.31	73,200	66.4%	58.6%	Unavailable
Apple	215.64	321.69	116,000	64.9%	39.4%	Unavailable
McKesson	198.53	60.97	78,000	17.2%	39.7%	Unavailable
UnitedHealth Group	184.84	122.81	230,000	Unavailable	Unavailable	Unavailable
CVS Health	177.53	94.46	250,000	Unavailable	Unavailable	Unavailable
General Motors	166.38	221.69	225,000	28.9%	24.2%	53.3%
AT&T	163.79	403.82	268,000	6.0%	5.0%	Unavailable
Ford Motor	151.80	237.95	201,000	38.4%	29.5%	Unavailable
AmerisourceBergen	146.85	33.66	19,000	Unavailable	Unavailable	Unavailable
Amazon.com	135.99	83.40	341,400	33.6%	24.4%	Unavailable
Cardinal Health	129.98	40.11	40,400	3.8%	13.6%	30.7%

[18] For example, related to the treatment of holding gains and losses

[19] For example, turnover (output) is not typically recorded at basic prices, since it excludes VAT but often includes taxes on products. In contrast, subsidies are rarely included in turnover.

Verizon Communications	125.98	244.18	160,900	Unavailable	Unavailable	Unavailable
General Electric	123.69	365.18	295,000	56.9%	48.9%	64.7%
Costco Wholesale	118.72	33.16	218,000	27.1%	31.1%	Unavailable
Walgreens Boots Alliance	117.35	72.69	360,000	28.6%	23.8%	Unavailable
Kroger	115.34	36.51	443,000	Unavailable	Unavailable	Unavailable
Chevron	114.47	260.08	55,200	56.5%	72.7%	52.0%
Express Scripts	100.29	51.74	25,600	0.1%	0.1%	0.0%
Home Depot	94.60	42.97	406,000	8.5%	11.0%	Unavailable
Boeing	94.57	90.00	150,500	59.0%	Unavailable	Unavailable
Alphabet	90.27	167.50	72,053	52.6%	23.7%	Unavailable
Microsoft	89.95	241.09	124,000	49.7%	43.3%	41.1%
Anthem	84.86	65.08	53,000	Unavailable	Unavailable	Unavailable
Phillips 66	84.28	51.65	14,800	29.1%	5.2%	0.0%
Comcast	80.40	180.50	159,000	Unavailable	Unavailable	Unavailable
IBM	79.92	117.47	404,800	62.2%	54.4%	Unavailable
Valero Energy	75.66	46.17	9,996	32.0%	9.9%	Unavailable
Johnson & Johnson	71.89	141.21	126,400	47.4%	42.7%	Unavailable
Target	69.50	37.43	323,000	Unavailable	Unavailable	Unavailable
Procter & Gamble	65.06	120.41	95,000	58.0%	55.8%	Unavailable
Lowe's	65.02	34.41	290,000	Unavailable	Unavailable	Unavailable
MetLife	63.48	898.76	58,000	30.4%	Unavailable	Unavailable
Marathon Petroleum	63.34	44.41	44,460	Unavailable	Unavailable	Unavailable
Aetna	63.16	69.15	49,500	1.0%	Unavailable	Unavailable
PepsiCo	62.80	74.13	264,000	41.5%	38.8%	57.2%
Archer Daniels Midland	62.346	39.769	31800	52.8%	30.7%	Unavailable

Note: Manually entered data is shown in grey, with all other data sourced from XBRL. All companies report in US GAAP.
Source: OECD ADIMA.

4.3. Geographical breakdowns of economic variables: Example of international sales

A key objective of ADIMA, in light of the strong user demand as described in Annex A, is to develop more granular (country-level) data regarding the activities of MNEs. It is clear however that the information provided by MNEs via their Annual Reports, even if this may include 'geographical segment information', is insufficient to meet this objective: companies tend to refer to business regions and not individual countries.

In absence of detailed reported data, it is necessary to use alternative methods to develop further breakdowns of geographical segment information reported by companies, on a country-by-country basis. An example of such an approach is the GeoRev methodology developed by FactSet, which uses statistics on trade and GDP to develop estimates of the relative weights of countries within a geographical segment. This method is however not ideal, particularly considering that one of the important research questions related to trade and FDI is whether market size (GDP) affects the decision to export or to invest.

Therefore, ADIMA explores the use of alternative data sources and the use of Big Data analytics to develop better estimates of the geographical breakdown of MNEs activities. The first one of these alternative sources of information that is considered is a company's website. Websites tend to be a marketing vehicle of the company and as such inherently reflect their global markets, and may form, therefore, a very useful source of information

on the international spread of sales. However, most information on websites is in an unstructured, textual format. This section discusses how ADIMA uses a corpus of publically available information from company websites and a range of methodologies to develop predictors of the geographic breakdown of the international sales of MNEs, as reported in their geographical segment information.

While using a variety of methodologies, as explained below, the overall approach is to summarise the location information included in websites – either country mentions, the location of subdomains, or external links – with simple frequency counts, which are then used as weights to disaggregate MNEs' reported geographical segment information for sales. As part of this process, detailed, MNE-specific concordance tables between countries and geographical segments are developed to reflect the fact that the definitions regarding geographic segments are not uniform across MNEs.

4.3.1. Methodology

The primary data source of information from MNE webpages is an open source 'copy of the internet' generated via web crawling from the Common Crawl. Box 4.1 describes this project in more detail. Based on the features of the online presence of the MNE and when geographic data exists, three analytical approaches have been developed, each of which is elaborated in greater detail below:

- **Case 1: Text Analytics:** A significant section of the website can be identified which describes the overall global strategy and organisation of the MNE. Advanced text analytics is used to develop frequency tables of country mentions in these webpages, which are used to disaggregate reported segment sales. In general, this methodology tends to be most valid for companies with significant Business-to-Business (B2B) transactions.
- **Case 2: Page Rank:** A company's business is segmented into location specific websites. An example of this is Wal-mart, whose websites include: walmart.com (United States), walmart.ca (Canada), asda.com (United Kingdom) and seiyu.co.jp (Japan). In this case, the popularity, or Page Rank, of each underlying site corresponding to a given location is used to disaggregate the reported segment information.
- **Case 3: Link Analysis:** A company's business is operated primarily via a website with global product-oriented presence. Examples of this include: microsoft.com and ibm.com. In this case, the outward links from the website are used to determine the distribution of sales by country.

Details regarding how the 37 MNEs have been classified in each case can be found in Table 4.5.

Box 4.1. Web crawling using the Common Crawl Project

Web crawling is the process of collecting information from websites on a large and automated basis. Common Crawl, a non-profit organisation "dedicated to providing a copy of the internet to internet researchers, companies and individuals at no cost for the purpose of research and analysis" (Common Crawl, 2018), has since 2014 collected billions of webpages via web crawling on a monthly basis, and has made their data publicly available as an Amazon Web Service Public Dataset.

Common Crawl uses a variety of crawling methodologies, including sampling sitemaps of hosts, following links from home pages (for several 'steps'), revisiting links from previous crawls and following domains donated by other web crawlers. Still, no web crawling project can ever envisage scraping and storing *all* online data. However, given its size and scope, Common Crawl should provide a reflective sample of the available information made accessible to the public by MNEs, in particular as the websites in question are some of the most connected global domains.

Overall, coverage has improved over time. Whilst crawls occur on a monthly basis, not all pages within a website are visited every month. Therefore, groups of 3 months are considered as a reflective sample of the underlying domain. We take the latest 3 month time period (May to July 2017) which intersects with the established 2016 reporting window (i.e. fiscal year-ends between July 2015 to June 2017).

From the website data that is collected, Common Crawl also compiles and makes available a Page Rank database (which measures how connected a website is) and a directed database of links between websites (from which the Page Ranks are derived).

4.3.2.1 Case 1: Text Analytics Method

When MNE websites contain significant sections that describe the global functioning of the underlying company (including for example webpages for investor relations), we assume that the frequency of country mentions reflects the overall importance of that country to the company. In these cases, text analysis is used to develop frequency counts of country mentions, which in turn determine the geographic sales share of each country, benchmarked to the segment information available in Annual Reports, as follows:

$$S_{i,j,k} = \beta_{i,k} \frac{c_{i,j}}{\sum_{m \in \{World\}} c_{i,m}} \qquad where\ \beta_{i,k} = {RS_{i,k}} \Big/ \frac{\sum_{m \in \{k\}} c_m}{\sum_{m \in \{World\}} c_m}$$

Where:

$S_{i,j,k}$ is the estimated sales for company i in country j in geographic segment k

$c_{i,j}$ is the frequency of mentions, within the webpages of company i, of country j

$RS_{i,k}$ is the reported sales for company i of geographic segment k (from Annual Reports)

$\beta_{i,k}$ is the adjustment factor that aligns the reported sales of company *i* for segment *k* with the geographic segment level from the frequency counts

Box 4.2 presents a detailed empirical example of this approach for Apple.

The development of a frequency count of location mentions is less straightforward than it seems, since a simplistic text frequency count may return many false positive results. For example, the Gulf of Mexico, which cannot be assigned to a country, would be falsely classified as Mexico, and a company with a CEO named Peter France would (wrongly) be assumed to do excessive business in the 'Hexagone'.

Natural Language Processing (NLP) is a technique for textual analysis that helps to overcome some of these shortcomings. NLP assigns elements of a sentence to language constructs (adjective, verb, noun etc.). One class of NLP, Named Entity Recognition (NER), aims to extract elements from the sentence which share the characteristics of belonging to a Person, Organisation or Location. This has been automated in Stanford University's NLP engine CoreNLP (see Manning et al., 2014), which was used to intelligently classify location references for ADIMA.

However, it should be noted that NLP is still not perfect. Locations may be missed, or there may be cases where locations may be ambiguous. For example, 'Georgia' can refer to both a State in the United States and to a country. To narrow the scope for these errors, only exact matches on country names were used, and problematic cases such as Georgia were excluded. Still, some overweighting of countries where the capital and the country share the same name (Singapore and Luxembourg) may occur. Among the plans for further development to refine the methodology, is to incorporate information from city locations, although this will also create additional challenges, for example for cases such as Dublin, which exists in both the United States and Ireland.

Box 4.2. Estimating geographic sales using Text Analysis: case study for Apple

Apple's website provides a wealth of information on the company's global presence, with a total of 9,742 pages on the apple.com domain with information on for example company news, jobs suppliers and investor relations. NLP text analytics identified nearly 10,000 location mentions, two thirds of which are countries.

Table 5.2 details how these location references were used to generate more detailed estimates of the international sales of Apple. Comparing the reported and estimated data, it becomes evident that while the estimates align well with the values reported for the United States and Asia & Pacific, improvements are still possible for other segments. Part of the overestimation of Europe and underestimation of China and Hong Kong, may be explained by language. Currently only English has been considered, but this will be elaborated to include other main languages going forward.

Table 4.2. Estimating Apple's global sales

	Reported Sales				Location Mentions			Adjusted Sales Estimate	
Geographic Segment	mln USD	%	Estimated Share (%)	Adjustment Factor	Country	Frequency	Share (%)	%	mln USD
United States	75667	35.1	36.0	1.0	United States	2161	36.0	35.1	75667
Americas (excl. USA)	10946	5.1	7.3	0.7	Canada	196	3.3	2.3	4909
					Mexico	76	1.3	0.9	1904
					Other	165	2.7	1.9	4133
Europe	49952	23.2	37.8	0.6	UK	266	4.4	2.7	5861
					Germany	212	3.5	2.2	4671
					France	172	2.9	1.8	3790
					Switzerland	129	2.1	1.3	2842
					Italy	114	1.9	1.2	2512
					Spain	106	1.8	1.1	2336
					Austria	83	1.4	0.8	1829
					Belgium	72	1.2	0.7	1586
					Ireland	71	1.2	0.7	1564
					Netherlands	65	1.1	0.7	1432
					Sweden	64	1.1	0.7	1410
					Other	913	15.2	9.3	20117
Japan	16928	7.9	4.6	1.7	Japan	276	4.6	7.9	16928
Asia & Pacific (excl. Japan)	13654	6.3	7.2	0.9	Australia	151	2.5	2.2	4784
					Singapore	106	1.8	1.6	3358
					New Zealand	63	1.0	0.9	1996
					Other	111	1.8	1.6	3516
Greater China	48492	22.5	7.1	3.1	China	268	4.5	14.0	30293
					Hong Kong	103	1.7	5.4	11643
					Other	58	1.0	3.0	6556
Total	215639	100	100			6001	100	100	215639

Note: The top 20 referred to countries within the sample are calculated in detail, all other countries are grouped into 'Other' within the appropriate geographic segment.
Source: OECD ADIMA.

4.3.2.2 Case 2: Page Rank Method

When an MNE hosts a 'family' of location-specific websites that focus on the operations of the business fragments by country, rather than the company as a whole, the text analysis (Case 1) yields inconclusive results. This happens either if a company has specific and (geographically) distinct brands with individual web domains (for example Wal-Mart Stores, whose brands include Walmart (United States, Canada, South America), Asda (United Kingdom), Seiyu (Japan) and Massmart (South Africa)), or when

a company has location specific sales sites. Amazon is prominent example of a company with an online presence structured like this, with websites including amazon.com (United States), amazon.de (Germany), amazon.co.uk (United Kingdom) and amazon.com.mx (Mexico).

For these enterprises, the popularity, or Page Rank, of each underlying location-specific site is used to disaggregate the reported segment information. This involves a 2-stage procedure:

First, all location-specific websites operated by the company are identified by analysing the links present on the primary website (for example amazon.com) to websites that are structured as a combination of the name of the primary website followed by a top level domain (TLD). In the case of Amazon, this means the algorithm searches for links to (i.e. the existence of) sites such as amazon.fr, amazon.be, amazon.org, amazon.net, and so on.

Secondly, the Page Rank of each individual country's website is determined. Page Rank is a measure of how connected or popular a website is, taking into account inward and outward links, and the quality of those links. The shares of each country in the MNE's total is then calculated – a measure of 'popularity' of a given country within a company – to disaggregate the reported segment sales, as follows:

$$S_{i,j,k} = \beta_{i,k} \frac{\sum_{m \in \{Websites\ s.t\ Country=j\}} p_{i,m}}{\sum_{m \in \{Classified\ Websites\}} p_{i,m}} \quad where$$

$$\beta_{i,k} = {RS_{i,k}} \Big/ \frac{\sum_{m \in \{Websites\ s.t\ Geographic\ Segment=k\}} p_{i,m}}{\sum_{m \in \{Classified\ Websites\}} p_{i,m}} \quad (1)$$

Where:

$S_{i,j,k}$ is the sales for company i in country j which is a member of geographic segment k

$p_{i,j}$ is the page rank of website j belonging to company i

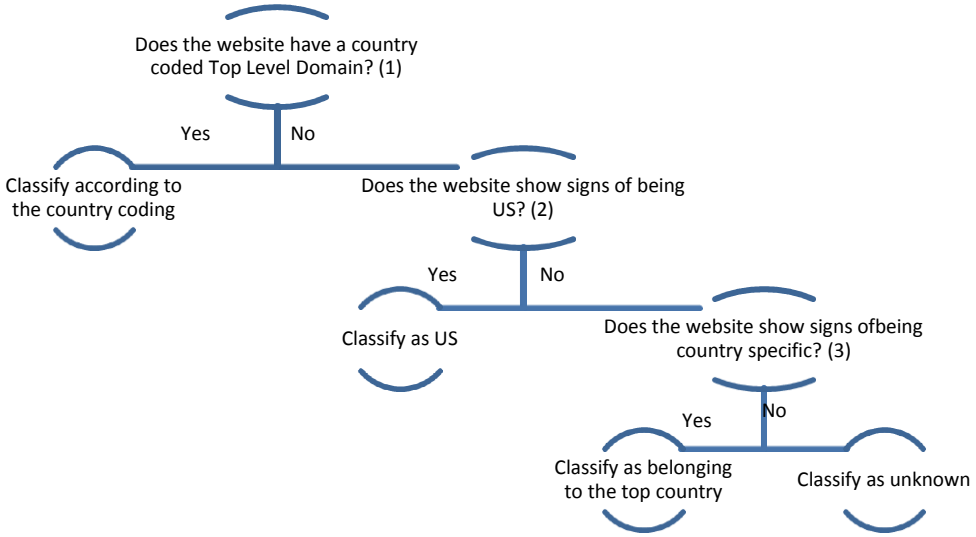
$RS_{i,k}$ is the reported sales for company i of geographic segment k in Annual Reporting

$\beta_{i,k}$ is adjustment factor so reported sales of company i aligns at the geographic segment level with derived sales at the geographic level k ($RS_{i,k} = \sum_{m \in \{k\}} S_{i,m,k}$)

A detailed example of the application of this method can be found in Box 5.3 for Amazon.

The most important challenge in this method is to determine the geographic location of a given website. Whilst this is rather straightforward for websites with a country-coded Top Level Domain (ccTLD) such as *.fr* or *.co.uk*, this is more difficult for sites with a Top Level Domain (TLD) such as .com, .net, .eu or .org, (which may still mean their operations are concentrated on one geographic location). In these cases, websites were allocated to a country using the decision tree displayed in Figure 4.1, using the nature of the outward links from a website, and *their* TLDs.

Figure 4.1. Determining the geographic location of a website



Note: (1) For example: a TLD *.com* is not country coded, whereas *.de* is. There are some notable cases where this fails such as bit.ly, goo.gl and youtu.be.
(2) websites are considered US if the a ratio of geographic outward domains to total outwards domains is less than 0.1. In short, this means that the vast majority of links present on the website are not geographically specified (ie. .com, .net, .org etc.)
(3) websites are considered 'country specific' if the share of geographic outward domains to a particular country is greater than 0.5, i.e., the geographically specified links present on the webpage tend to be point to same country.

This methodology is not complete in all cases. For example, if companies operate through multiple brands (for example Walmart), the assignment of the relevant websites has to be manually mapped. The method is also not suitable when companies operate vastly different business models across their websites. For example, one subsidiary of Amazon is IMDb (Internet Movie Database). Given its popularity, using the Page Rank method for IMDb results accounts for 20% of Amazon's US sales, which given that the platform operates on a freemium business model seems unlikely. In this case, IMDb was excluded from the calculations, but this is not always possible.

Box 4.3. Estimating geographic sales using the Page Rank method: case study for Amazon

Amazon actively encourages links to its websites via its Associates Programme (affiliate marketing). As a result, links to Amazon's websites are likely to be determined by websites looking to promote Amazon products. Using the information on the country coded TLDs of these websites, the Page Rank method is likely to yield results reflective of geographic sales. The results are presented in Table 5.3.

This approach may however overweight small countries when a full range of products (and thus links) are offered to smaller markets. The results indeed seem to indicate this for Rest of the World. Further adjustments to this method may consider therefore the number of outward links relative to the inward links.

Table 4.3. Estimating Amazon's global sales

	Reported Sales				Page Rank Share				Adjusted Sales Estimate	
Geographic Segment	mln USD	%	Estimated Share (%)	Adjustment Factor	Website	Page Rank (%)	Country	Share (%)	%	mln USD
United States	90349	66.4	59.4	1.1	amazon.com	59.4	United States	59.4	66.4	90349
Germany	14148	10.4	9.3	1.1	amazon.de	9.3	Germany	9.3	10.4	14148
Japan	10797	7.9	8.6	0.9	amazon.jp	0.3	Japan	8.6	7.9	10797
					amazon.co.jp	8.3				
United Kingdom	9547	7.0	8.1	0.9	amazon.co.uk	8.1	United Kingdom	8.1	7.0	9547
Rest of World	11146	8.2	14.6	0.6	amazon.com.au	0.7	Australia	0.7	0.4	549
					amazon.com.br	0.7	Brazil	0.7	0.4	526
					amazon.ca	2.6	Canada	2.6	1.5	2004
					amazon.cn	1.1	China	1.1	0.6	855
					amazon.es	1.7	Spain	1.7	0.9	1269
					amazon.fr	3.0	France	3.0	1.7	2309
					amazon.in	1.1	India	1.1	0.6	814
					amazon.it	2.5	Italy	2.5	1.4	1919
					amazon.co.kr	0.0	Korea	0.0	0.0	17
					amazon.com.mx	0.7	Mexico	0.7	0.4	501
					amazon.nl	0.4	Netherlands	0.4	0.2	335
					amazon.com.sg	0.0	Singapore	0.0	0.0	24
					amazon.co.th	0.0	Thailand	0.0	0.0	24
Total	135987	100	100			100			100	135987

Source: OECD ADIMA.

4.3.2.3 Case 3: Link Analysis Method

Neither Method 1 or Method 2 can deal with MNEs who operate their websites on a global scale with a product-oriented approach, such as IBM and Microsoft. In this case, the outward links from the website are used to determine the distribution of sales by country, the logic being that MNEs link to websites that are important to their

operations[20]. A greater frequency of links to a particular country is an indication of an increased presence in that particular country (whereby the geography of a website is again determined using the methodology in Figure 4.1). A prerequisite for this analysis is a sufficient number of outward links. Using this approach, country-level sales are derived as follows:

$$S_{i,j,k} = \beta_{i,k} \frac{\sum_{m\in\{Websites\ s.t\ Country=j\}} 1}{\sum_{m\in\{Classified\ Websites\}} p_{i,m}}\ where$$

$$\beta_{i,k} = RS_{i,k} \Big/ \frac{\sum_{m\in\{Websites\ s.t\ Geographic\ Segment=k\}} 1}{\sum_{m\in\{Classified\ Websites\}} 1} \quad (1)$$

Where:

$S_{i,j,k}$ is the sales for company i in country j which is a member of geographic segment k

$RS_{i,k}$ is the reported sales for company i of geographic segment k in Annual Reporting

$B_{i,k}$ is the adjustment factor so reported sales of company i aligns at the geographic segment level with derived sales at the geographic level k ($RS_{i,k} = \sum_{m\in\{k\}} S_{i,m,k}$)

Box 4.4. presents an example of this calculation for IBM.

[20] While the use of inward links was considered as well in the development of this methodology, it was discarded because many smaller websites will link to highly connected websites in order to improve their ranking within search engines.

Box 4.4. Estimating geographic sales using link analysis: Case study for IBM

IBM's website is designed as a global hub to support the sales of their underlying business and products. The total IBM website comprises of around one million individual pages. On these pages, nearly 5000 well connected external links could be identified. Table 5.4 presents the results.

Table 4.4. Estimating IBM's global sales

Geographic Segment	Reported Sales		Estimated Share (%)	Adjustment Factor	Geographic Links			Adjusted Sales Estimate	
	mln USD	%			Country	Links	Share (%)	%	mln USD
United States	30194	37.8	47.0	0.8	United States	2285	47.0	37.8	30194
Japan	8339	10.4	5.5	1.9	Japan	269	5.5	10.4	8339
America (excl. USA)	7643	9.6	7.5	1.3	Brazil	176	3.6	4.6	3685
					Canada	139	2.9	3.6	2911
					Other	50	1.0	1.3	1047
Europe, Middle East and Africa	24769	31.0	29.9	1.0	United Kingom	362	7.4	7.7	6167
					Germany	297	6.1	6.3	5059
					France	125	2.6	2.7	2129
					Italy	75	1.5	1.6	1278
					Russia	69	1.4	1.5	1175
					Switzerland	54	1.1	1.2	920
					Spain	53	1.1	1.1	903
					Netherlands	40	0.8	0.9	681
					Sweden	36	0.7	0.8	613
					Ireland	33	0.7	0.7	562
					Israel	30	0.6	0.6	511
					Other	280	5.8	6.0	4770
Asia & Pacific (excl. Japan)	8974	11.2	10.0	1.1	Australia	112	2.3	2.6	2060
					China	90	1.9	2.1	1655
					India	79	1.6	1.8	1453
					Korea	34	0.7	0.8	625
					New Zealand	25	0.5	0.6	460
					Other	148	3.0	3.4	2722
Total	79919	100	100			4861	100	100.0	79919

Note: The top 20 countries linked to by IBM are included. All other countries are aggregated into an 'Other' component within the appropriate geographic segment.
Source: OECD ADIMA.

4.4. Data validation

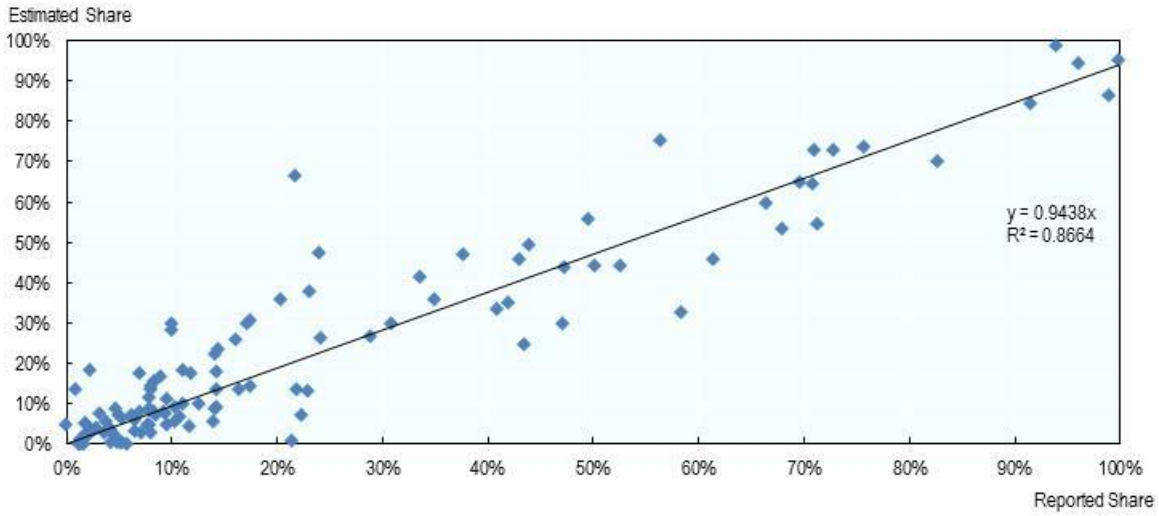
4.4.1. Comparing estimated and reported figures

In order to validate our results, the shares of the geographical segment data reported by MNEs are compared with the (unadjusted aggregated country level) estimations obtained via the methodologies developed above. For this purpose, a detailed MNE-specific concordance table was developed to align countries with the geographical segments as reported.

Figure 4.2 presents the results of this comparison between the estimated and reported geographic segments. Overall, the estimated values align well with the reported geographic segments, as evident from the high R-squared (0.87), supporting the use of these estimated values at the country level, benchmarked to the total geographical segment data as reported. However, it is also clear that in a few cases, the estimation

could be further improved. The concluding section lists a variety of techniques that will be explored to further improve these results.

Figure 4.2. Comparison of reported geographic segments and estimated geographic segments



Source: OECD ADIMA.

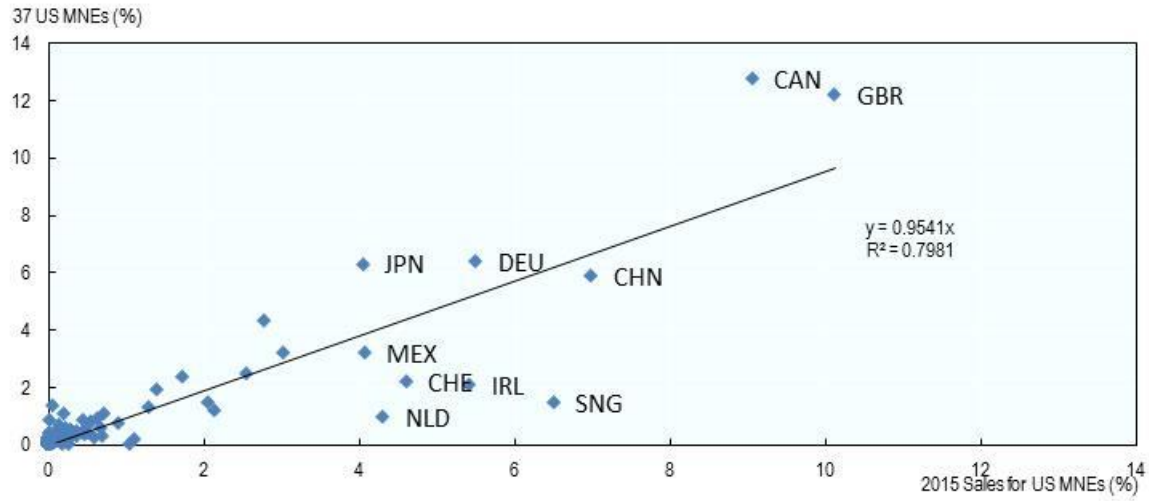
4.4.2. Comparison with US outward FATS statistics

Using the pilot sample of 37 of the largest US MNEs, the sales data by country that were generated using the above procedures and benchmarked to the reported segment data in MNE reports, were compared with the official outwards FATS statistics.

The total foreign turnover of the 37 US MNEs equalled $1.3 trillion USD which alone account for one-fifth of US Outward FATS.[21] Figure 4.3 compares the sales of the 37 US MNEs with official FATS statistics at the country level, showing a strong correlation between the two (r = 0.8). However, for several countries, important differences are visible. For countries often used for fiscal optimisation purposes, like the Netherlands, Ireland and Singapore, official FATS data record (much) higher sales than in the database of OECD MNEs. An important reason for this difference is that the outward FATS data by the United States record sales based on the location of the foreign affiliate - even if some ambiguity remains as this is not always fully identifiable in complex hierarchical layers of MNE structures involving these countries – whereas the sales data in ADIMA, reflecting the annual reports by companies and their websites, are rather a depiction of the sales structures by location of the final client.

[21] As a share of most recent US Outward FATS estimates (2015).

.Figure 4.3. Sales of Foreign US Affiliates compared to estimated foreign sales of the 37 US MNEs



Note: When 2015 Sales data has been supressed in provisional 2015 data, 2014 values are utilised.
Source: OECD ADIMA, US BEA: https://www.bea.gov/international/usdia2015p.htm

4.5. Data analysis: Internationalisation of selected MNEs

As discussed in Section 3.4, a variety of measurements of the scale and scope of international activities of MNEs have been developed in the academic literature on MNEs and International Business. Partly driven by the limitations in data, one of the most prominent measures used remains the *degree of internationalisation*, expressed as the share of foreign activities in an MNEs total activities. This approach for example also underpins UNCTAD's Trans Nationality Index (TNI) which is calculated as the average of the share of foreign assets, sales and employment.

An important drawback of this measure is that it provides limited details on the spread of these activities. For example, two US MNEs may have the same degree of internationalisation, but one may achieve all its international sales in neighbouring Canada, while the other sells widely around the globe. The latter is arguably 'more international'. In addition, in cross-country comparisons, the ratio of foreign activities to total activities is hampered by differences in the size of the domestic economy: MNEs from smaller countries will quickly seem more international.

The data developed by ADIMA includes Country-by-Country Sales data which mitigates these issues and overcomes the limitation imposed by non-uniform company geographic segment reporting. Table 4.5 presents harmonised sales per geographic segment for the US MNEs in the pilot. One notable example from the data is in the Manufacture of refined petroleum products (ISIC Rev. 4: 1920) where Exxon Mobil focuses on international sales in Europe (30.7%) with limited sales in Asia (12.2%), however Chevron has the opposite picture focussing on international sales in Asia (25.8%) and limited sales in Europe (5.6%).

Table 4.5. Geographic Sales by company

Harmonised Geographic Sales Share by region for 37 US MNEs, with method used to obtain degree of internationalisation

	Method	United States	North America (excl. USA)	South America	Europe	Asia	Africa	Oceania
Wal-Mart Stores	Case 2	75.7	12.3	5.0	4.8	2.1	0.1	0.0
Exxon Mobil	Case 1	33.6	10.3	2.3	30.7	12.2	8.0	2.8
Apple	Case 1	35.1	4.2	0.8	20.9	35.2	0.5	3.2
McKesson	Case 1	82.8	7.3	0.1	4.6	2.4	1.5	1.3
UnitedHealth Group	Case 2	91.1	0.6	3.7	2.8	1.8	0.0	0.0
CVS Health	Case 2	95.5	0.0	4.5	0.0	0.0	0.0	0.0
General Motors	Case 2	71.1	2.8	2.9	14.3	5.3	0.0	3.6
AT&T	Case 2	98.4	1.6	0.0	0.0	0.0	0.0	0.0
Ford Motor	Case 2	61.6	6.9	1.0	24.3	4.4	0.2	1.7
AmerisourceBergen	Case 2	95.9	4.1	0.0	0.0	0.0	0.0	0.0
Amazon.com	Case 2	66.4	1.8	0.4	21.7	9.2	0.0	0.4
Cardinal Health	Case 2	96.2	2.4	0.0	0.0	1.5	0.0	0.0
Verizon Communications	Case 3	90.2	1.2	0.5	4.9	1.6	0.3	1.4
General Electric	Case 1	43.1	5.7	2.8	17.5	27.3	3.6	0.0
Costco Wholesale	Case 2	72.9	17.6	0.0	1.3	6.1	0.0	2.1
Walgreens Boots Alliance	Case 1	61.6	11.1	0.3	26.3	0.6	0.1	0.1
Kroger	Case 2	100.0	0.0	0.0	0.0	0.0	0.0	0.0
Chevron	Case 1	43.5	3.0	7.9	5.6	25.8	10.4	3.7
Express Scripts	Case 2	99.9	0.1	0.0	0.0	0.0	0.0	0.0
Home Depot	Case 2	91.5	8.5	0.0	0.0	0.0	0.0	0.0
Boeing	Case 1	41.0	2.8	1.5	14.6	35.7	2.5	1.9
Alphabet	Case 3	47.4	3.1	3.1	35.1	8.5	0.4	2.4
Microsoft	Case 3	50.3	3.1	1.9	30.8	9.5	0.6	3.8
Anthem	Case 2	97.2	0.0	0.0	2.8	0.0	0.0	0.0
Phillips 66	Case 1	70.9	2.3	0.3	23.9	2.6	0.0	0.0
Comcast	Case 3	91.8	0.8	0.8	4.0	0.8	0.6	1.3
IBM	Case 3	37.8	4.2	5.4	29.2	18.8	0.9	3.8
Valero Energy	Case 1	68.0	13.8	0.0	16.2	1.9	0.0	0.0
Johnson & Johnson	Case 1	52.6	3.0	5.0	21.9	11.8	4.9	0.8
Target	Case 2	99.4	0.6	0.0	0.0	0.0	0.0	0.0
Procter & Gamble	Case 1	42.0	6.1	5.0	23.0	20.9	2.9	0.2
Lowe's	Case 2	94.2	5.8	0.0	0.0	0.0	0.0	0.0
MetLife	Case 2	69.6	0.5	1.6	8.5	18.9	0.3	0.6
Marathon Petroleum	Case 2	100.0	0.0	0.0	0.0	0.0	0.0	0.0
Aetna	Case 1	99.0	0.0	0.0	0.4	0.3	0.2	0.1
PepsiCo	Case 1	58.5	10.6	4.5	13.1	9.5	3.2	0.6
Archer Daniels Midland	Case 1	50.1	2.8	5.1	31.4	7.6	2.0	1.0
37 US MNES		**72.1**	**5.1**	**2.0**	**11.1**	**7.4**	**1.2**	**1.0**
GDP as a share of world		**24.7**	**4.1**	**4.7**	**25.2**	**36.5**	**2.9**	**2.0**

Source: OECD ADIMA.

5. Conclusions and next steps

Consistently and comparatively measuring the international activities of MNEs has been a longstanding and increasingly pertinent challenge in economic statistics. Given that national statistical institutes are typically limited in their (legal) ability to capture activities outside their jurisdiction, an international and 'whole of the MNE' approach is required to better understand the global scale and scope of MNEs, but also to support the consistent treatment of MNEs in national statistics. ADIMA, for which this paper presented a proof of concept, aims to provide such an approach. Thus far the database contains data for 37 US-based MNEs, and the innovative approaches used to estimate their international activities match well with comparable official statistics.

Whilst already very promising, work will continue to further refine both the register of parent-affiliate relationships, for example by improving the fuzzy matching algorithm, as well as the estimated geographical breakdown of sales, for example by mapping the web domains related to each MNE through their server security certification. As new data sources, in particular the LEI, expand and mature, their information is incorporated on a nearly real-time basis. The application of the techniques to non-US MNEs (towards a set of 100 MNEs by the end of 2018 and 500 MNEs by 2020) will also in all likelihood require further refinements of existing approaches. Additional verification of the data with official sources, in particular in collaboration with Eurostat, should further validate the results and point to areas for improvement.

Collaboration with Eurostat has also started with respect to the Monitoring tool, to ensure alignment and complementarity with the Early Warning System. The innovative approaches explored in the Register and Indicators provide ample fuel for developing an MNE monitoring tool that provides very timely information on changes in Multinational activities, including for example investments and divestments, mergers and acquisitions, or changes in leadership and headquarters. The information can be classified by MNE and the countries and industries affected. Pilot tests with interested countries are envisaged to ensure that the specifications meet compilers' needs as much as possible.

ADIMA can further support the development of internationally consistent statistics on MNEs at the national level by publishing – subject to country feedback – a proposed country classification for the Ultimate Controlling Institutional Unit (or Global Decision Centre) for each MNE, as well as information on its consolidated industry classification.

Importantly however, the data sources and analytical techniques presented in this paper are merely scratching the surface of what is possible. Digitalisation may pose measurement problems for statistics, but also presents new solutions. Social media, Open Street Map, new services such as GDELT (global news), job vacancy websites, and many more are currently being explored to develop more insights on MNEs, their international activities, jobs and business functions, and (supplier) relationships with other MNEs. One of the first priorities will be to use these data sources to develop breakdowns by country of MNEs' total assets and employment.

As a final remark, it is important to note that other applications of the web-analytics techniques currently used in ADIMA may also be explored. For example, particularly promising insights can be derived for measuring digital trade and the digital economy, by analysing for example if (and where or how) MNEs and their affiliates are engaged in ecommerce. A recently published study (Statistics Netherlands, 2016) provides a very good example of such an approach. While such analyses are not the main focus of this paper nor of the overall MNE database, it is hoped that the work presented here provides contributes to the wider debate on the smart use of big data in the development of official statistics.

References

BAE Systems (2018), Website detailing job linkages, https://www.baesystems.com/en-uk/download-en-uk/20171128153134/1434607493611.pdf (accessed 28 February 2018).

BEA (2014), "U.S. International Economic Accounts: Concepts and Methods : Statistics on the Activities of Multinational Enterprises", Bureau of Economic Analysis, Washington D.C., https://www.bea.gov/international/concepts_methods.htm.

CocaCola (2018), Website detailing employment profile, http://www.coca-colacompany.com/careers/who-we-are-infographic (accessed 28 February 2018).

Common Crawl Project (2018), Website detailing the project, commoncrawl.org (accessed 28 February 2018).

EY (2013), "US GAAP versus IFRS", http://www.ey.com/Publication/vwLUAssets/EY-US-GAAP-vs-IFRS-the-basics-2013/$FILE/EY-US-GAAP-vs-IFRS-the-basics-2013.pdf.

EY (2017), "Financial Reporting Developments: Segment reporting: Accounting Standards Codification 280: Rev. May 2017", pp. 62.

Fortanier, F., and R. V. Tulder (2007), "Internalization Trajectories Of Multinational Enterprises: 1990-2004", CIBS Conference Paper, Reading, 16-17 April 2007.

Fortanier, F., and R. V. Tulder (2009), "Internationalization trajectories—a cross-country comparison: Are large Chinese and Indian companies different?", Industrial and Corporate Change, Vol. 18 No. 2, pp. 223-247, https://doi.org/10.1093/icc/dtp003.

Gal, P. (2013), "Measuring Total Factor Productivity at the Firm Level using OECD-ORBIS", *OECD Economics Department Working Papers*, No. 1049, OECD Publishing, Paris, http://dx.doi.org/10.1787/5k46dsb25ls6-en.

Gramlich, J. and J. Whiteaker-Poe (2013), "Disappearing subsidiaries: The cases of Google and Oracle", http://dx.doi.org/10.2139/ssrn.2229576.

Hitt, M., R. Hoskisson and H. Kim (1997), "International Diversification: Effects on Innovation and Firm Performance in Product-diversified Firms", The Academy of Management Journal, Vol. 40 No. 4, pp. 767-798, http://www.jstor.org/stable/256948.

Ietto-Gillies, G. (1998), "Different Conceptual Frameworks in the Assessment of the Degree of Internationalization: Empirical Analysis of Various Indices for the Top 100 TNCs", Transnational Corporations, Vol. 7 No. 1, United Nations Conference on Trade and Development, Geneva, pp. 17-40.

ILO (2017), "World Employment and Social Outlook", International Labour Organization, Geneva.

Jacquemin, A. P. and C. H. Berry (1979), "Entropy measure of diversification and corporate growth", The Journal of Industrial Economics, Vol. 27 No. 4, pp. 359-369.

Kalemli-Ozcan, S. et al. (2015), "How to Construct Nationally Representative Firm Level data from the ORBIS Global Database", National Bureau of Economic Research Working Papers, No. 21558.

LEI ROC (2016), "LEI ROC Report on collecting data on direct and ultimate parents of legal entities in the Global LEI System", https://www.leiroc.org/publications/gls/lou_20161003-1.pdf.

Manning, C. et al. (2014), "The Stanford CoreNLP Natural Language Processing Toolkit", Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 55–60.

Marano, V. et al. (2016), "Home country institutions and the internationalization-performance relationship: A meta-analytic review", Journal of Management, Vol. 42 No. 5, pp. 1075-1110.

Ramaswamy, K., K. Kroeck and W. Renforth (1996), "Measuring the Degree of internationalization of a Firm: A Comment", Journal of International Business Studies, Vol. 27 No. 1, pp. 167-177.

Pinto Ribeiro, S., S. Menghinello and K. De Backer (2010), "The OECD ORBIS Database: Responding to the Need for Firm-Level Micro-Data in the OECD", OECD Statistics Working Papers, No. 2010/01, OECD Publishing, Paris. http://dx.doi.org/10.1787/5kmhds8mzj8w-en.

Rugman, A. (2005), "The Regional Multinationals", Cambridge University Press, Cambridge.

SEC (2016), "Report on Modernization and Simplification of Regulation S-K", Securities and Exchange Commision, Washington D.C.

Statistics Netherlands (2016), "Measuring the internet economy in The Netherlands: a big data analysis", Statistics Netherlands Discussion Papers, No. 14.

Sullivan, D. (1994), "Measuring the Degree of Internationalization of a Firm", Journal of International Business Studies, Vol. 25 No. 2, pp. 325-342.

Annex A. Measuring MNEs activity: policy questions and statistical challenges

The importance of better measuring the role of MNEs in the global economy has long been recognised in academia, among policy makers and certainly also among statisticians. The paucity of data on their international activities is therefore not a result of a lack of understanding of their relevance or of the policy demand for such information, but rather due to the important statistical challenges involved, at the conceptual, institutional, and practical/empirical level. This section reviews some of the main current policy themes regarding the role of MNEs. It aligns these challenges with the data that would be needed to support the discussions, and refers to the statistical challenges and emerging solutions involved in collecting these data.

It's important to note up-front that not all answers can be provided by ADIMA on its own, but in these cases ADIMA is able to provide part of the solution, in combination with data collected by national statistical offices.

Understanding the MNE itself

Policy questions

Although only limited information is available on the exact scale, scope and nature of the international activities on MNEs, the data that are available paint a clear picture regarding their role and importance. A prerequisite for any policy analyses related to competitiveness or trade is therefore to better understand the MNE itself – both as it is currently structured as well as on (planned) changes over time. This implies a need for information about for example the nature of the activities and business functions that are taking place in the compiling economy, how much value added is generated, how many jobs – and what kind – are provided, and how the domestic activities are linked to an MNE's operations abroad via international trade, investment and knowledge creation and use. But also information on strategic changes such as outsourcing and offshoring – or reshoring – will contribute to providing such basic insights.

While current data collections on MNEs, including FATS statistics and of course FDI data, provide important insights, they are only able to provide limited answers to these questions and they cannot provide a whole picture for individual MNEs.

National approaches

At the national level, data linking is among the most obvious and promising solutions to developing more statistics on MNEs – see for example the recently developed TEC and STEC statistics that include a foreign ownership dimension – and can be used to incorporate other statistical areas as well, including employment (see also below). The integration of such linked primary statistics within the larger accounting frameworks of National Accounts (e.g. via Extended Supply and Use tables with a foreign ownership

breakdown, as currently developed in several countries under the coordination of the CSSP Expert Group on Extended Supply and Use Tables) or Balance of Payments (e.g. the creation of Current Account statistics by ownership, as currently explored by an IMF BOPCOM Working Group led by OECD) represents an important next step.

This process is however not without important, well-known challenges. It requires both a correct *identification* of MNEs and their affiliates (e.g. in the business register) and an *integrated profiling* of their data across statistical domains to ensure consistency.

The identification of enterprises that are foreign controlled, as well as (the country of) the Ultimate Controlling Institutional unit, often requires the integration of multiple data sources or even manual checks, which is complicated further by the fact that a responding company may not have the required information. It is even more challenging to correctly identify the affiliates of domestic MNEs *outside* the compiling economy (e.g. for outward FATS) as these are outside the jurisdiction of the statistical office.

The treatment of MNEs across statistical domains – such as business statistics, trade, and FDI – invariably results in discrepancies due to e.g. distinct survey frames or classifications, informed by the different purposes of these statistics. The integrated profiling of MNEs involving the confrontation and harmonisation of data sources, is therefore an important but also resource-intensive task.[22]

International approaches

Yet even if these challenges are tackled, two important questions remain that can only be addressed internationally: first, how to ensure a consistent treatment of the same MNE *across* countries, and second, how to obtain information on the role and relevance of the activities in the compiler's country within the overall geographical profile of MNEs.

This is however not without its own challenges, related to the absence of microdata, the high cost and unknown quality of commercial databases, the difficulties related to data sharing and information exchange across countries, and the requirements of timeliness and cross-country consistency.

ADIMA aims to make a contribution towards tackling these challenges, complementing and collaborating with existing efforts including EuroStat's EuroGroups Register and the Global Legal Entity Identifier Foundation (GLEIF) (see also section 2.4). By providing consistent information on the activities of MNEs in the countries in which they operate, the ADIMA database aims not only to improve our understanding of MNEs, but also to support national statistical institutions with their profiling activities, which form the foundation for all further statistical work on better measuring MNEs. In many instances, the same MNEs (and the same acquisitions, mergers, relocations…) are profiled across multiple countries. An argument can be made that centralisation of at least part of this profiling work will not only further enhance cross-country consistency in national statistics but also alleviate the compilation burden of statistical offices for these statistics.

[22] See for example the results of the 2018 WPTGS Stocktaking Survey (question on Large Case Units).

MNEs and international trade: the trade-investment nexus and the consequences of MNEs for global, regional and national value chains

Policy questions

Closely following the policy-relevant question of 'what MNEs actually do' in countries is the question of 'what is their impact', typically phrased in economic terms. Answering this question often requires the consideration of a multitude of dimensions: how much do MNEs contribute to the productive base of a host country, e.g. by building up new enterprises, enlarging existing ones, or recapitalising and reorganising inefficient firms? What are the consequences of the entry or presence of MNEs for competitors within the same industry: do MNEs contribute to improved competition and resource allocation? How much, if any, of MNEs' technological and managerial know-how is brought into the country and spills over to local enterprises, e.g. via the turnover of employees?

Particularly relevant within the context of analysing global (and regional, and national) value chains, is the question of linkages between MNEs with buyers and suppliers – both nationally (forward and backward linkages, respectively) or internationally (resulting in trade transactions). Linkages with MNEs do not only create potential new business for local enterprises, including enhanced (indirect) access to overseas markets, but may also be important channels for knowledge spillovers. At the same time, MNEs may choose to use non-resident suppliers or distributors, including their other affiliates, augmenting trade. Finally, important questions remain as well with respect to the consequences of MNEs for taxes, and the practices MNEs use – e.g. via the strategic shifting of intellectual property across countries, intra-firm trade and profit repatriation – to minimise their overall tax burden.

Despite decades of academic research, increasingly sophisticated econometric models, and improved microdata (often from national statistical offices), many of these issues remain empirical questions, and the exact answer typically differs across countries and industries because it is dependent not only on the type and purpose of an MNE investment but also a function of the absorptive capacity of domestic firms and advantageous institutional and policy environments.

National and international statistics

Improved statistics on MNEs, particularly when integrated within accounting frameworks, can help provide higher quality, more definitive answers to many of these questions. In addition, whereas academic studies tend to be 'one-off' and aimed at finding relations between variables, official statistics are better suited to track key developments over time in a timely and comparable fashion, and therefore are arguably better at informing policy makers on the role of MNEs in their countries.

A particular improvement in our understanding of the role of MNEs in global, regional and national value chains (i.e. the combined overview of domestic backward and forward linkages, and the relationship between trade and investment) will be provided by the Extended SUTs with foreign ownership breakdowns that are currently being developed in several countries. Whereas even the most sophisticated academic studies on linkages have had to confine themselves with industry-wide technical coefficients to be used in their regression models, Extended SUTs can directly provide insights into how much additional value added is generated, both upstream and downstream, due to the activities of MNEs.

Likewise, when such extended tables are subsequently integrated within global SUTs (requiring, as a minimum, additional information on the differences in geographical and product breakdown of trade by MNEs), not only the effect of MNEs on national but also on international linkages (trade) can be demonstrated, including the ultimate sources of traded value added and their final consumers. The report published by OECD and the Nordic Statistical offices (2017) provides an illustration of such an approach.

Extending the national accounts framework to also include FDI income flows, can help provide insights into how much of the value added generated by foreign investors actually 'sticks' in the economy. As illustrated by the paper by Mehigan *et al* discussed at the joint WPTGS-WGIIS meeting in 2016, while wages and taxes typically remain in the economy, the operating surplus or profits accrues to the foreign parent, who decides whether these profits are reinvested in the affiliate or are repatriated to the home country. This is not a negligible part: FATS statistics indicate that around 45% of value added produced by foreign owned firms in OECD countries consists of operating surplus and hence can (potentially) be repatriated. Likewise, when *imports* by MNEs are sourced from its affiliates, some of the profits generated by these affiliates from this sale will in the end accrue to the parent.

Importantly, other statistics that are currently developed provide insights into the relationship between investment and trade. Trade by Enterprise Characteristics (TEC) and Services Trade by Enterprise Characteristics (STEC) form a first step towards Current Account statistics with a foreign ownership breakdown, which can provide indications for example on how much domestic MNEs and foreign affiliates contribute to current account surpluses and deficits.

It is clear that none of these statistical innovations is without important challenges, in terms of both data availability and coherence across statistical domains. In many ways, these reflect those presented above. However, additional difficulties include ensuring that the accounting adjustments – some of which, in certain countries, are made only at the total-economy level – are made separately for different types of enterprises.Again, it is hoped that collaborative international work, on measuring MNEs but also on the creation of the building blocks for high quality global SUTs, will provide both incentives and efficiency gains in advancing this work.

MNEs, employment and inclusive globalisation

Policy questions

The policy debate in recent years has increasingly focused on what has become referred to as 'inclusive globalisation' – the growing realisation that the benefits of globalisation have not accrued equally to all members of society, and that policies should focus on better supporting those that have been, even if only temporarily, left behind. While such inclusivity is multidimensional, an important aspect clearly revolves around employment and wages – how many jobs have been created due to globalisation, where, what type and quality, with which skills and for what wage, and which were lost.

Many of these policy questions on the employment benefits (or costs) of globalisation can directly be translated to similar questions on the employment consequences of MNE investment. According to the latest UNCTAD World Investment Report (2017) MNEs directly employ 82 million workers (or 2.5 percent of the global workforce (ILO, 2017).

But this only tells part of the story, as MNEs can generate significant multipliers employment effects. For example, BAE Systems calculated that for every 100 of its jobs, it supported an additional 380 in the UK economy as a whole[23]. And Coca-Cola proudly claims that the 'Coca-Cola system' ranks among the world's top 10 private employers with more than 700,000 system employees[24].

Moreover, in developed economies in particular, MNEs often create 'higher quality' jobs, are typically found to pay higher wages than domestic enterprises, although the 'MNE-premium' could vary between high and low skilled labour so as to exacerbate the relative wage of skilled employees. At the same time MNEs often offshore low skilled jobs to low-wage countries, which partly explains the globalisation backlash seen in some developed economies.

National and international statistics

Providing statistics to support the very complex discussions on inclusive globalisation requires the integration of detailed information on employment and wages, typically available in administrative data (tax records) or in Labour Force Surveys, with economic statistics. While data linking has by now become a regular practice within the compilation of business and economic statistics in many statistical offices, the bridge to social statistics has not yet been crossed frequently.

One of the important exceptions is the work by Statistics Netherlands, regularly published in their *Internationalisation Monitor*, which has linked employers and employees for a number of years and has recently also integrated this analysis within an input-output framework, indicating for example that foreign controlled affiliates in the Netherlands accounted for 700 thousand jobs (full-time equivalents) directly, and a near-similar number of jobs indirectly (i.e. at upstream suppliers) (Statistics Netherlands, 2017). These results are in line with a similar study developed by the OECD secretariat in collaboration with the Nordic Statistics offices (2017).

The integration of data on employees and enterprises in the Netherlands and the Nordics has been facilitated by administrative records on their interrelationships (e.g. social security records), but still required overcoming important challenges, for example related to re-weighting the data to avoid potential biases in the results, or the alignment of the linked business statistics with national accounts concepts (which in both studies was achieved *post-hoc* by applying industry-level breakdowns to the official National Accounts estimates).

While these examples may be useful for other countries to follow – for example, the WPTGS stocktaking survey showed that Ireland and the US have developed projects to link enterprise and employee data – the required data sources may not be available in all countries. In these cases, additional data on employment by MNEs generated at the international level (including e.g. on their business functions), which is one of the outputs planned by the ADIMA may help in bridging some of the gaps.

[23] https://www.baesystems.com/en-uk/download-en-uk/20171128153134/1434607493611.pdf

[24] http://www.coca-colacompany.com/careers/who-we-are-infographic

Annex B. 100 MNEs in initial sample

Table B.1. Largest 100 MNEs ordered by revenue

#	Company Name	#	Company Name	#	Company Name	#	Company Name
1	Wal-Mart Stores	26	Cardinal Health	51	Alphabet	76	Johnson & Johnson
2	China Petroleum & Chemical	27	Total	52	Microsoft	77	Uniper
3	Toyota Motor	28	Verizon Communications	53	Siemens	78	Engie
4	Royal Dutch Shell	29	Honda Motor	54	Assicurazioni Generali	79	Airbus
5	PetroChina	30	General Electric	55	Prudential	80	Indian Oil
6	Volkswagen	31	Japan Post Holdings	56	Nestle	81	Tesco
7	Berkshire Hathaway	32	Costco Wholesale	57	Petrobras	82	Target
8	Exxon Mobil	33	Walgreens Boots Alliance	58	NK Lukoil	83	SK Holdings
9	Apple	34	Fiat Chrysler	59	Anthem	84	Sony
10	McKesson	35	Kroger	60	Phillips 66	85	Aviva
11	UnitedHealth Group	36	Allianz	61	Carrefour	86	Panasonic
12	BP	37	Chevron	62	Hitachi	87	Ceconomy
13	CVS Health	38	SAIC Motor	63	NK Rosneft	88	Procter & Gamble
14	Samsung Electronics	39	Nissan Motor	64	Comcast	89	Lowe's
15	General Motors	40	Ping An Insurance Group Co of China	65	IBM	90	Muenchener Rueckversicherungs
16	AT&T	41	Nippon Telegraph and Telephone	66	SoftBank	91	Zurich Insurance Group
17	Daimler	42	China Mobile	67	China Life Insurance	92	Marubeni
18	Glencore	43	Express Scripts	68	Hyundai Motor	93	People's Insurance Group of China
19	Ford Motor	44	Gazprom	69	Deutsche Telekom	94	MetLife
20	Exor	45	BMW	70	Japan Post Insurance	95	Marathon Petroleum
21	AmerisourceBergen	46	Legal & General Group	71	Valero Energy	96	Aetna
22	China State Construction Engineering	47	Home Depot	72	Electricité de France	97	JXTG
23	Amazon.com	48	Boeing	73	DowDuPont	98	PepsiCo
24	AXA	49	China Railway Group	74	Aeon	99	Audi
25	Hon Hai Precision Industry	50	China Railway Construction	75	Enel	100	Archer Daniels Midland

Note: Of the 39 US companies in the Top 100 by revenues, 37 of 39 were selected for the Pilot. DowDuPont and Berkshire Hathaway were the two MNEs excluded. DowDuPont was excluded because of its recent reorganisation (merger of DuPont and Dow Chemical) which has not yet been reflected in commercial, annual reports or website data sources. Berkshire Hathaway was excluded from the pilot sample because its economic variables do not relate directly to the economic performance of MNEs under control of Berkshire but rather the performance of Berkshire Hathaway investment services (ie sales of Berkshire not consolidated sales of Dairy Queen, Kraft, etc).
Source: OECD ADIMA.

Annex C. Register construction procedure

The register of parent-affiliate data is compiled from three sources: Orbis, annual reports significant subsidiaries (as filed in the 10-K to the SEC, and downloaded from Edgar), and LEI. This annex describes the compilation procedure in more detail for each source, as well as the matching procedure used to combine them.

Orbis

Orbis is a commercial database (owned by Bureau van Dijk) with financial and other information for more than 250 million companies worldwide. While the data sources and methodology used to create Orbis remain opaque (purposefully, in light of competitors), it is one of the very few sources that provides an overview of parent-subsidiary relationships, including ownership shares.

A list of majority-owned affiliates was extracted for the pilot enterprises of 37 MNEs, for the financial year 2016[25], by capturing all enterprises which listed the Orbis identifier (BvD identifier) of the MNE as Global Ultimate Owner (GUO).

The MNE BvD identifier was identified via a match to MNEs' ISIN codes, which was straightforward for 36 out of 37 US MNEs in the pilot study. Only for Wal-Mart, the ultimate owner was listed as the US-resident Walton Family, hence not generating any challenges with respect to the geographical attribution of the MNE.[26]

The result of the preliminary extraction was a flat file linking all majority-owned subsidiaries to the Ultimate Parent MNE. Establishment level data ('branches' in Orbis) were excluded from the dataset, and only active subsidiaries, with links to the parent active in 2017, were retained. Demographic information such as country, city, and internationalised names were subsequently integrated into the dataset.

Annual reports data (SEC filings)

For all 37 MNEs, the Exhibit 21 forms were extracted from Edgar, the data repository of the US SEC, which where company annual reports are filed using form 10-K. all 'significant' subsidiaries are listed in this Exhibit, including their names and country of incorporation. In a few cases, more details on the percentages of ownership were also available.

[25] Annual reports usually are published in the first quarter of the subsequent year

[26] More complex cases would involve dual-listed firms, or firms ultimately owned by holdings. Under these scenarios MNEs could be subsidiaries of other trees. For example, Coty Inc (a US cosmetics producer) and Reckitt Benckiser Group Plc (a UK consumer goods company), are both at least partially owned by JAB Holdings and appear as 'children' within this tree in Orbis. (Complicating matters further, JAB Holdings is owned by Agnaten (from Austria), which in turn is controlled by private individuals).

Legal Entity Identifier

The compilation procedure for LEI relationships was similar to the one adopted for Orbis.

First, LEIs of the MNE ultimate parent were identified via desk research. After extracting LEI relationship data, all LEIs related to the MNEs LEIs were selected, including both direct and ultimate consolidation linkages, with potentially different time stamps and different reporting sources. Subsequently, all linkages denominated as 'inactive' were removed, and when a duplicate of the same linkage was found, with different periods, only the latest (and active, i.e. with no end date) was retained. No selection was made as to whether the information was classified as "entity supplied only", "fully corroborated", or "partially corroborated".

To ensure a full coverage of MNEs' hierarchies, when MNEs were shown as ultimate but not direct owner, a secondary extraction was performed for this direct owner to identify additional links.

Table C.1. LEI ultimate relationships

	Number of affiliates with LEI	Number of foreign affiliates with LEI	Number of foreign countries	Foreign country with most LEIs (#LEIs)
AT&T	15	9	4	Mexico (3)
Aetna	85	5	3	United Kingdom (1)
Alphabet	2	1	1	Ireland (1)
Amazon.com	3	0	0	
Anthem	22	0	0	
Apple	15	14	9	Ireland (4)
Archer Daniels Midland	64	62	18	Germany (9)
Boeing	1	0	0	
Comcast	20	11	4	United Kingdom (5)
Costco Wholesale	12	10	7	Canada (2)
Exxon Mobil	7	4	2	United Kingdom (2)
Ford Motor	4	3	3	
General Electric	124	118	32	France (14)
General Motors	1	1	1	Germany (1)
Home Depot	1	0	0	
IBM	35	26	15	Netherlands (8)
Johnson & Johnson	124	83	19	Ireland (17)
Kroger	1	0	0	
Marathon Petroleum	1	0	0	
McKesson	3	3	3	
MetLife	90	19	12	Argentina (3)
Microsoft	4	2	2	Bermuda (2)
PepsiCo	24	17	11	Netherlands (3)
Procter & Gamble	5	4	3	United Kingdom (2)
Target	2	0	0	
UnitedHealth Group	14	3	3	
Valero Energy	8	5	3	United Kingdom (3)
Verizon Communications	2	0	0	
Wal-Mart Stores	2	1	1	United Kingdom (1)
Walgreens Boots Alliance	9	8	4	Luxembourg (4)

Note: Data displayed was last extracted in early February 2018.
Source: LEI

Table C.1. shows the geographic distribution of affiliates covered by the LEI relationship data for the US MNEs in the pilot study, showing large variation in both the share of foreign entities in the total affiliates with an LEI, as well as in the most connected countries. For example, 4 out of 9 of AT&T's foreign entities are in Mexico, and 19 of

the 83 foreign entities that declare to be owned by Johnson and Johnson are registered in Ireland. Both these findings reflect the density of affiliates identified by Orbis.

Fuzzy matching algorithm

The subsidiaries identified in Orbis, Annual reports and the LEI were combined using fuzzy matching, which is commonly used to match records from different sources in the absence of a common identifier. The algorithm calculates a measure of similarity based on textual information such as names and addresses. The matching routine consisted of four main steps:

1. **Name 'pre-processing'** of all three databases, which included the conversion of all names into upper case, removal of extra spaces, and removal of punctuation and other signs such as brackets

2. **Matching Annual Reports and Orbis,** using information on company names and countries. The calculation of similarity is based on bigrams: i.e. company names are broken into two-letter pieces, and the similarity score is based on the ratio of the number of common bigrams of the two names and the average name length. (Note that for country names, only perfect matches were allowed).

   To identify a 'match', a minimum similarity score was required of 0.96 (after exploring various other alternatives, carefully balancing the number of false positives (if the similarity score is set too low) and false negative (if the similarity score is set to high). Since the matching was done within-MNE, and one-to-one[27], the complexity of the comparisons was low, and avoided the long computation times that are often a problem in fuzzy matching exercises.

   Finally, the two data sources were combined (with flags), deduplicating the matched records and keeping the unmatched records from both sources.

3. **Integrating LEI records.** The newly formed dataset was subsequently matched with information from LEI records using a similar procedure. In addition to enterprise names and countries, city information used as well in the matching algorithm. The minimum threshold imposed here was 0.99, as manual checks indicated that a lower threshold would result in too many false matches.

The final dataset was further checked manually, in particular for matches where the similarity was relatively low (i.e. not perfect). The results of the pooled data for the 37 US MNEs are provided in table C.2.

Table C.2. Affiliate counts in the United States and top 10 partner countries

	USA	CAN	GBR	NLD	IRE	DEU	AUS	MEX	FRA	CHN	LUX
AT&T	674	8	<5	8	<5	<5		**15**			
Aetna	384		**7**		<5					<5	
Alphabet	149	8	9	<5	**14**	8	<5	<5	<5		
Amazon.com	140	<5	**18**	<5	<5	15	<5	<5	<5	<5	10
AmerisourceBergen	100	5	**8**	<5		<5	<5		<5		<5
Anthem	363				<5					<5	

[27] This is a crucial decision based on the objective of the exercise, i.e. deduplication. There can only be one matching entity between each source and any other imperfect match is a different entity that gets added to the data.

Apple	98	**8**	6	<5	**8**	<5	<5	<5	<5		
Archer Daniels Midland	264	26	17	19	6	**33**	<5	<5	<5	<5	<5
Boeing	337	**15**	10	9	<5	<5	14	<5		<5	
CVS Health	577										
Cardinal Health	285	**14**	5	<5	<5	<5	<5	5	<5	<5	<5
Chevron	274	11	9	11	<5		**23**		<5		
Comcast	1074	21	**73**	5		8	23	7	5	<5	
Costco Wholesale	148	**31**	6					<5	<5		
Express Scripts	147	<5	<5	<5		<5					
Exxon Mobil	338	73	22	**107**		18	27		10	6	10
Ford Motor	169	**22**	14	5	<5	12	<5	<5	6	<5	<5
General Electric	1055	98	**216**	63	144	44	53	13	73	8	12
General Motors	366	25	**44**	12	<5	35	10	8	6	14	
Home Depot	136	**17**	<5							<5	
IBM	197	**28**	26	42	14	**28**	23	<5	8	10	<5
Johnson & Johnson	338	16	21	20	**48**	22	6	9	12	12	5
Kroger	923	<5									
Lowe's	389	**55**						<5			
Marathon Petroleum	183	<5									
McKesson	184	14	39	<5	**69**	17			14	<5	<5
MetLife	530	5	13		18		8	**34**	<5	<5	
Microsoft	115	8	**21**	8	12	14	5	<5	8	<5	10
PepsiCo	607	**86**	20	44	12	5	21	63	<5	21	26
Phillips 66	131	<5	**14**		<5	5					
Procter & Gamble	237	10	**31**	18	<5	25	8	7	17	24	7
Target	120	**10**				<5				7	<5
UnitedHealth Group	941	**21**	9	5	<5		5	<5		<5	10
Valero Energy	222	**23**	12	<5	<5			<5			<5
Verizon Communications	359	<5	9	6	9	7	6	<5	9		**12**
Wal-Mart Stores	382	**178**	27	<5	<5			26		27	21
Walgreens Boots Alliance	583	<5	**125**	15	5	13	<5	9	15	25	32

Note: CVS Health appears to own a number of subsidiaries in Brazil according to the Exhibit 21. This relationship is detected in Orbis but deemed inactive.
Source: OECD ADIMA

Annex D. Activity classifications

MNE Parent-level data

When submitting regulatory reports to the SEC, US listed enterprises are also required to list their industry code that best describes their activities, using the SIC classifications, which is still maintained by the SEC for continuity even if it has been superseded by NAICS in 1997. These SIC codes were converted to ISIC Rev 4 codes using the official concordances from the US Census Bureau between SIC, NAICS 2002, 2007, 2012 and ISIC4[28].

For 4 MNEs (Target, Costco, Wal-Mart and Alphabet), the SIC code corresponded to multiple ISIC codes. In these cases annual reports were used to manually identify the closest matching candidate. The resulting concordance is shown in Table D.1.

Table D.1. SIC and ISIC Rev 4. Classification of MNEs

	SIC Code	ISIC Rev.4	ISIC label
Archer Daniels Midland	2070	1040	Manufacture of vegetable and animal oils and fats
Boeing	3721	3030	Manufacture of air and spacecraft and related machinery
Target	5331	4719	Other retail sale in non-specialized stores except auto and home stores
Exxon Mobil	2911	1920	Manufacture of refined petroleum products
Ford Motor	3711	2910	Manufacture of motor vehicles
General Electric	3600	2710	Manufacture of electric motors, generators, transformers & electricity distribution/control app.
IBM	3570	6202	Computer consultancy and computer facilities management services
Kroger	5411	4711	Retail sale in non-specialized stores with food, beverages or tobacco predominating
Lowe's	5211	4752	Retail sale of hardware, paints and glass in specialized stores
Cvs Health	5912	4772	Retail sale of pharmaceutical & medical goods, cosmetic & toilet articles in specialized stores
Pepsico	2080	1104	Manufacture of soft drinks; production of mineral waters and other bottled waters
Procter & Gamble	2840	2023	Manufacture of soap/detergents, cleaning/polishing preparations, perfumes, toilet prep.
Chevron	2911	1920	Manufacture of refined petroleum products
Wal Mart Stores	5331	4719	Other retail sale in non-specialized stores except auto and home stores
Johnson & Johnson	2834	2100	Manufacture of pharmaceuticals, medicinal chemical and botanical products
Apple	3571	2620	Manufacture of computers and peripheral equipment
Home Depot	5211	4752	Retail sale of hardware, paints and glass in specialized stores
Cardinal Health	5122	4649	Wholesale of other household goods medical instruments and orthopedic devices
Unitedhealth Group	6324	6512	Non-life insurance
Verizon Communications	4813	6110	Wired telecommunications activities except direct-to-home satellite television service
At&T	4813	6110	Wired telecommunications activities except direct-to-home satellite television service
Microsoft	7372	1820	Reproduction of recorded media
Costco Wholesale	5331	4719	Other retail sale in non-specialized stores except auto and home stores
Mckesson	5122	4649	Wholesale of other household goods medical instruments and orthopedic devices
Amazon Com	5961	4791	Retail sale via mail order houses or via Internet
Valero Energy	2911	1920	Manufacture of refined petroleum products
Metlife	6311	6511	Life insurance

[28] https://www.census.gov/eos/www/naics/concordances/concordances.html

Aetna	6324	6512	Non-life insurance
Amerisourcebergen	5122	4649	Wholesale of other household goods medical instruments and orthopedic devices
Anthem	6324	6512	Non-life insurance
Comcast	4841	6020	Television programming and broadcasting activities except transmission only
General Motors	3711	2910	Manufacture of motor vehicles
Marathon Petroleum	2911	1920	Manufacture of refined petroleum products
Express Scripts	5912	4772	Retail sale of pharmaceutical & medical goods, cosmetic & toilet articles in specialized stores
Phillips 66	2911	1920	Manufacture of refined petroleum products
Walgreens Boots Alliance	5912	4772	Retail sale of pharmaceutical & medical goods, cosmetic & toilet articles in specialized stores
Alphabet	7370	6312	Web portals web search portals

Source: OECD ADIMA

Affiliate-level data

Industry characteristics of the affiliates were drawn from Orbis. The data are reported using NACE Rev.2 4-digit codes, but NAICS and US SIC codes are also available. All data were mapped to ISIC Rev.4 using the official UN correspondence table. At the 4-digit level, the concordance between NACE Rev.2 and ISIC Rev. 4 is (by design) simple, with only a few cases where multiple NACE codes map into one ISIC code.