How to construct a global business register?

The OECD approach for ADIMA: Release for FY 2018
The international statistical community has long held the ambition for the creation of a global business register of the largest MNEs. The OECD analytical database of individual MNEs and their affiliates (ADIMA) takes a step towards fulfilling these ambitions by bringing together innovative digital and traditional data sources to analyse the activities of the world’s largest Multinational Enterprises (MNEs). For the world’s largest 500 public MNEs ADIMA provides:

- A comprehensive view of the structure of each MNE and its subsidiaries
- A comprehensive view of the websites of each MNE
- A harmonised set of indicators at the global level for each MNE
- A monitoring tool to identify corporate events such as large company restructurings and headquarter relocations.

Data is available: oe.cd/ADIMA.
Acknowledgements

The ADIMA Database, reference date 31st December 2018 (Financial Year 2018), was compiled by Graham Pilgrim, Henry Dehé, Covadonga Machicado Alvarez and Peio Villanueva from the Statistics and Data Directorate.

The development of the ADIMA methodology benefitted from the valuable feedback of the OECD Working Parties on International Trade in Goods and Services Statistics (WPTGS) and International Investment Statistics (WGIIS).
Table of Contents

Foreword ........................................................................................................................................... 2
Acknowledgements ......................................................................................................................... 3
1. Introduction .................................................................................................................................. 6
2. Challenges to building a global business register ....................................................................... 8
   Confidentiality restrictions prevent data sharing ........................................................................... 8
   Could open source business registries be the solution? ............................................................... 8
   Could Annual Reporting be the solution? ...................................................................................... 9
   Pooling multiple sources can increase coverage ......................................................................... 10
   Pooling data in practice ............................................................................................................... 10
   Constructing ADIMA – summary ............................................................................................ 11
3. Constructing ADIMA – detail ..................................................................................................... 14
   Stage 1. Choose the companies to profile .................................................................................. 14
   Stage 2. Find relationships declared by Annual Reporting ....................................................... 14
   Stage 3. Find relationships declared by big data sources ........................................................... 16
   Stage 4. Combine data into a standardised data structure ........................................................... 17
   Stage 5: Find a network structure for every parent MNE ........................................................... 19
   Stage 6: Generate the Physical Register .................................................................................... 21
   Stage 7: Validate the Physical Register ..................................................................................... 23
   Stage 8: Generate the Digital Register ....................................................................................... 27
   Stage 9: Generate the Indicators ............................................................................................... 28
   Stage 10: Generate the Monitor .................................................................................................. 30
4. Conclusions ..................................................................................................................................... 34
   ADIMA applications ..................................................................................................................... 34
   Improving ADIMA through outreach ......................................................................................... 34
References ......................................................................................................................................... 36

Tables

Table 2.1. Potential sources of data on MNEs: Available Data and Limitations .................................... 10
Table 3.1. Big data sources within ADIMA ....................................................................................... 17
Table 3.2. Data schema for Vertices and Edges within ADIMA ........................................................ 18
Table 3.3. Summary of Relationships (Edges) within ADIMA ............................................................ 19
Table 3.4. Variables within the Physical Register ............................................................................. 23
Table 3.5. Summary of Edges within ADIMA by Validation Type .................................................... 24
Table 3.6. Variables within the Digital Register ............................................................................... 28
Table 3.7. Variables within the Indicators ....................................................................................... 29
Table 3.8. Internationalisation measures for a selection of companies ............................................ 30
Table 3.9. Variables within the Monitor ......................................................................................... 32
Table 3.10. Headlines for AbbVie (26th June 2019) ....................................................................... 33
Figures

Figure 2.1. Example of a network for an MNE ................................................................. 11
Figure 2.2. Step by Step: The ADIMA approach ........................................................... 13
Figure 3.1. Reflecting Joint Ventures in ADIMA ............................................................ 19
Figure 3.2. A subset of the ADIMA network for Apple Inc. ........................................... 21
Figure 3.3. Determining the unique companies belonging to an MNE network .................. 22
Figure 3.4. Determining the scope of validation in ADIMA ........................................... 24

Boxes

Box 3.1. Handling of Atomic Weapons Establishment (AWE) in ADIMA ....................... 26
Box 3.2. Example of internationalisation measures for a selection of companies ............. 29
Box 3.3. Example of analysing a corporate event using the ADIMA monitoring tool .......... 32
1. Introduction

Despite their significant and growing importance, with implications across a range of policy areas, information on Multinational Enterprises (MNEs) remains limited. The international statistical community has long held the ambition for the creation of a global business register of the largest MNEs which provides:

- **Global coverage of affiliates**: MNEs engage in a multitude of activities across a number of economic territories. However, national statistics offices typically work on the basis of residence rather than nationality of production units, and as a result data collections generally focus on the activities of MNEs and their affiliates within their economic territory and rarely provide the view of an entire MNE. In addition, there are significant efforts to introduce greater granularity in the presentation of official statistics (beyond FDI) that highlight the contribution made by foreign owned firms (and the nationality of ownership) but this information is not always readily available in national statistical or administrative registers, creating risks of gaps or indeed inconsistencies (for example in determining the ultimate owning country).

- **Consistent data across jurisdictions**: To gain an understanding of activity of MNEs within their economic territory NSOs often carry out surveys of their foreign activities, for example in collecting FATS statistics. This means that affiliates of the same MNE may be surveyed multiple times, bringing with it the risk that transactions (especially intra-firm transactions) may not always be recorded consistently across countries, leading to imbalances in National Accounting concepts.

The OECD analytical database of individual MNEs and their affiliates (ADIMA) takes a step towards fulfilling these ambitions by compiling statistics on the scale and scope of the international activities of MNEs. At present, the output of ADIMA includes:

- **Physical Register**: A comprehensive view of the structure of each MNE and its subsidiaries

- **Indicators**: A harmonised set of indicators at the global level for each MNE

The innovative methods and techniques which ADIMA applies allows for outputs beyond the conventional scope of traditional business registers, including:

- **Digital Register**: A comprehensive view of the websites of each MNE provides scope to better understand new business dynamics being shaped by digitalisation, whilst also providing potential to inform the estimates of macro-economic statistics and in particular digital trade, as considered in the Handbook on Measuring Digital Trade.

- **Monitor**: A monitoring tool to identify corporate events such as large company restructurings and headquarter relocations that are likely to have an impact on macro-economic statistics, as illustrated by the 26% GDP growth in Ireland in 2015\(^1\), due to a number of large multinational corporations relocating their intellectual property, and indeed headquarters, there.

ADIMA provides a platform for NSOs to gain a greater understanding of affiliates operating within their jurisdiction. However, it is the first of many steps required to fulfil the ambition of creating a global business register. Realising this objective requires the support and engagement of a number of stakeholders and the first steps to engage NSOs, international organisations and MNEs are underway. However, further support and engagement is encouraged to improve future releases of the database.

This paper is structured as follows: section 2 discusses the challenges faced in the construction of a global business register giving an overview of how the OECD overcame these difficulties, how this can be helpful to NSOs and international initiatives, and the next steps to advance the project. Section 3 explains the workflow, detailing sources, methods and techniques. Section 4 concludes.
2. Challenges to building a global business register

The goal of a global business register is to develop the means to describe the geographic and economic footprint of the world’s largest MNEs at a global level (i.e. at the level of the whole MNE) and extend this to the level of each individual affiliate.

Confidentiality restrictions prevent data sharing

In theory official statistics collected by national statistics offices (NSOs) could be pooled together to provide these views. However, in practice, at a global level, this is not possible due to confidentiality restrictions, that typically prohibit data sharing across NSOs and that prohibit the dissemination of confidential information on individual respondents (e.g. individual affiliates).

The EuroGroups Register (EGR) shows the benefits that can be gained by pooling together data from NSOs within the European Statistical System (ESS). The EGR aims to facilitate the coordination of survey frames for producing high quality statistics on global business activities. However, access to this information is strictly limited to those within the ESS and coverage is not global, but, rather, limited to the EU.

Some private sector data vendors do compile this information. However, the cost of data access is often prohibitive for NSOs and International Organisations alike. Licence restrictions often prevent the dissemination of results which reveal individual company level data. In addition, metadata and documentation of data provenance is often missing or incomplete, making appraising data quality difficult.

Could open source business registries be the solution?

Open source data can help overcome these limitations, removing the constraint of breaching confidentiality restrictions and providing a common reference point for all. Jurisdictions are increasingly providing open source access to business registry data, including in some cases ownership and financial indicators. Unfortunately, although promising (especially as more countries may move in this direction), currently, coverage is not complete and highly dependent on jurisdiction, and therefore, the information currently available is insufficient to build a global business register alone. For example, in the United Kingdom a bulk download of all companies detailing basic information is made available and a supplementary register of People with Significant Control is also made available, which allows for chains of control to be determined.

---

2 However, some efforts leaded by International Agencies are encouraging and providing tools to overcome these restrictions, like the UNECE Task Force on Exchange and Sharing of Economic Data
3 Providers include Dun & Bradstreet, Bureau van Dijk and Thomson Reuters.
4 Available at: http://download.companieshouse.gov.uk/en_output.html
5 Available at: http://download.companieshouse.gov.uk/en_pscdata.html
Initiatives at the global level also exist. The Legal Entity Identifier (LEI)\(^6\) is a unique global identifier to identify the legal entities participating in financial transactions, and provides additional information to identify the ultimate owner of each legal entity. LEIs are a requirement to participate in a number of financial markets. Unfortunately, their issuance tends to be limited to the subset of firms operating within financial markets, and, so, coverage of legal entities is not complete. Another example of a global initiative includes PermID\(^7\) which provides an identifier for a subset of global companies, however coverage is, again, not universal.

**Could Annual Reporting be the solution?**

One of the best examples of timely open source information is individual Annual Reporting\(^8\). However, this is only a regulatory requirement for publicly traded companies and therefore only provides coverage for a subset of MNEs.

In addition Annual Reporting formats are not standardised, and so the process of harmonising data collection for many MNEs is necessarily a manual and resource intensive process.

Furthermore, the type of information required in Annual Reporting is largely determined by the Exchange, and completeness in a number of areas is not guaranteed. For example, Apple Inc, with a primary listing in the United States, files Annual Reporting with the Securities and Exchange Commission (SEC). SEC *affiliate reporting standards* allow for insignificant subsidiaries to be excluded, and therefore Apple Inc declares only 9 affiliates within 5 jurisdictions for Financial Year 2018\(^9\), whereas evidence of Apple Stores exists in at least 25 jurisdictions\(^10\). Furthermore, details of the economic transactions of each Apple Inc affiliate is not made available – instead aggregations are often made on geographic levels. These geographic levels are non-standardised, with Apple Inc defining reportable segments of Americas, Europe, Greater China, Japan and Rest of Asia Pacific, where Europe consists of European countries, as well as India, the Middle East and Africa\(^11\). Moreover, the example of Apple Inc. is indicative of a wider issue surrounding the depth and consistency of reporting.

\(^6\) More information regarding LEI can be found at: [https://www.gleif.org/en/](https://www.gleif.org/en/)

\(^7\) More information regarding PermID can be found at: [https://permid.org/](https://permid.org/)

\(^8\) Annual Reporting provides a snapshot of the current state of the business, including financial statements describing the company’s performance during the last year. The notes to the financial statements can include further information regarding subsidiaries and geographic breakdowns of performance.


\(^10\) Information regarding Apple Stores available at: [https://en.wikipedia.org/wiki/Apple_Store](https://en.wikipedia.org/wiki/Apple_Store)

Pooling multiple sources can increase coverage

As noted above, Annual Reporting is not the only source of publicly available information on MNEs. Like Annual Reporting however, these additional sources are also, typically, limited to a subset of companies (by function, jurisdiction or significance) and, ownership and financial indicators are also, often, limited, as shown in Table 2.1 (for the key sources used in compiling ADIMA):

<table>
<thead>
<tr>
<th>Source</th>
<th>Coverage</th>
<th>Contact Numbers</th>
<th>Website/Email</th>
<th>Ownership</th>
</tr>
</thead>
<tbody>
<tr>
<td>PermID</td>
<td>permid.org</td>
<td>Subset of global firms</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>LEI</td>
<td>gleif.org</td>
<td>Subset of global firms (typically engaging in financial transactions)</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Business Registers</td>
<td>National Sources</td>
<td>Full coverage of firms within a jurisdiction</td>
<td>Varies</td>
<td>Varies</td>
</tr>
<tr>
<td>Common Crawl (1)</td>
<td>commoncrawl.org</td>
<td>Most important global websites</td>
<td>No</td>
<td>Varies</td>
</tr>
<tr>
<td>X.509 Certificates (2)</td>
<td>opendata.rapid7.com</td>
<td>Websites utilising security certificates</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>WikiData (3)</td>
<td>wikidata.org</td>
<td>Subset of global firms listed on Wikipedia</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Note:
(1) Common Crawl is an open source initiative to provide a monthly snapshot of the internet through web scraping. As part of the project, they provide a quarterly network graph of hyperlinks between domains, which can be used to determine whether two websites are probably linked. For example if example.com links to example.fr, and example.fr links to example.com, then it is likely that they are a member of the same Parent MNE family.
(2) X.509 Certificates are used in a number of internet protocols to ensure secure communications (such as in e-commerce) and have increased in use as internet search engines have begun to prioritise domains with secure communications in their ranking algorithms. X.509 certificates detail the company and the domains that it operates.
(3) WikiData is a community sourced initiative derived from Wikipedia data, it details the website and subsidiaries of a company.

However, although all of these sources come with caveats and gaps, pooling them together provides significant scope to overcome the caveats and fill gaps in a way that provides a more comprehensive view of the whole of the MNE and the location of its affiliates, whilst also providing potential to develop an economic view of each affiliate.

Pooling works because each source typically contains certain identifying information that allow any given unit to be linked to units identified in other sources. As such, a view of all (or at least as many as the sources allow) of the affiliates within a given MNE can be knitted together. These identifiers include “shared” resources such as websites and contact numbers (telephone and fax numbers).

Pooling data in practice

We use a simple example below (see also Figure 2.1) to illustrate how the various sources of data can identify parts of an MNE that may not be identified in separate sources.

Consider a parent MNE, Company A, with three affiliates Company B, Company C and Company D. In the Annual Reporting source Company A reports only its control of company B, but not Company C or Company D (due to the limitations of Annual Reporting
discussed above). However administrative and often statistical business registers provide additional information about Company A, for example its domain website (example.com), and with digital sources (e.g. CommonCrawl) we can establish that this domain name is linked to example.fr, which can be identified as belonging to Company C through X.509 Certificates. Further, and again from administrative registers we can identify that Company B has a shared telephone number with Company D. As such we’re able to identify links between the parent A and two additional affiliated firms, C and D.

Some care is needed in confirming these links and so further validation is needed. For example the use of security certificates to establish links across domains is not fool-proof, as another Company E may also have a right to operate the domain but Company E may not be controlled by Company A, for example Company C may merely have outsourced the provision of digital services on the domain to Company E. The outsourcing of the provision of a contact number is also possible, and validation would also be necessary in the case of Company D.

Validation focuses on determining whether an affiliate is controlled by the parent MNE\(^\text{12}\) (i.e. has a voting power greater than 50%), and is performed by desk research into each individual company.

**Figure 2.1. Example of a network for an MNE**

![Diagram of an MNE network](image)

**Constructing ADIMA – summary**

In theory, with the range of sources described above, significant scope exists to develop a truly global register, although, of course, additional sources will be needed to ensure that coverage of privately owned or government controlled firms, or indeed coverage of MNEs in developing economies (where the sources described above have larger gaps) are adequately captured. However even if all this information were available the necessity to

\(^{12}\) A controlled affiliate (or subsidiary) is an enterprise in which an investor owns more than 50% of its voting power (OECD Benchmark definition of Foreign Direct Investment). The same criterion determines the enterprises covered in the Foreign Affiliates’ Statistics (FATS) statistical framework.
undertake, resource intensive, data capture and validation processes summarily described above, necessarily limits the scope of MNEs that can be covered, which depends on available resources.

The OECD effort in this regard has been to focus on the world’s largest publicly listed MNEs (by market capitalisation), and to develop a register of these firms in the Analytical Database of Individual Multinationals and their Affiliates (ADIMA).

The construction of ADIMA follows ten distinct stages as summarily described below (see also Figure 2.2) and in more detail in Section 3.

As noted, due to limited resources, ADIMA (currently\(^\text{13}\)) focuses only on the world’s 500 largest listed MNEs, (Stage 1), to capitalise on Annual Reporting (one of the most complete sources for this cohort of firms). The top 500 are selected on the basis of Market Capitalisation, i.e. the total value assigned to a company’s outstanding shares of stock.

Annual Reporting is then used to create the first view (Stage 2) of the structure of each selected MNE (including using manual data capture tools), and other open sources (as described above) are used to complement this information to provide a more comprehensive view of the MNE structure (Stage 3).

The collected data are then standardised into a common data structure (Stage 4). This involves extracting metadata for each unit, so for example data regarding the location and address. In addition, connections are also extracted, so, for example, in the case of a website, data regarding the company that has a certificate to communicate securely on the domain and linked websites is collected\(^\text{14}\).

For each MNE group structure, ADIMA can build a physical register of affiliates (Stage 6). However, as the physical register relies on the accuracy of a number of data sources, validation is required to ensure accuracy (Stage 7). Validation involves checking the identified and matched link of each affiliate that has been derived by non-standard links such as via websites. This process is continued iteratively until the physical register has been fully validated.

As the data sources used by ADIMA go beyond those utilised by a traditional business register, it is possible to build extensions, including a digital register of the websites operated by the MNE (Stage 8, see also Section 3), a harmonised set of indicators (Stage 9) and a monitoring tool (Stage 10) for large corporate events.

---

\(^{13}\) In theory, MNEs outside of the top 500 could be considered, in line with increases in resources and data availability, including through partnerships with national statistics offices and other international organisations.

\(^{14}\) This information is then brought together and linked. Because of the multitude and complexity of connections across parents and affiliates ADIMA uses a network database approach, which is better suited to describe the connections between parents and affiliates. Network databases provide scope to perform complex queries which under a traditional database would be computationally complex and expensive. An example of this is the search performed in Stage 5 (below) to determine all identifiers which are linked to a Parent MNE. In a network graph architecture, relationships are followed from each identifier until all routes are exhausted. However, within a traditional database this would involve iteratively querying the whole database until no additional information was discovered.
Figure 2.2. Step by Step: The ADIMA approach

- **Building ADIMA**
  - Stage 1: Choose the companies to profile
  - Stage 2: Find relationships declared by Annual Reporting
  - Stage 3: Find relationships declared by big data sources
  - Stage 4: Combine data into a standardised data structure
  - Stage 5: Find a network structure for every parent MNE
  - Stage 6: Generate the Physical Register
  - Stage 7: Validate the Physical Register

- **Building ADIMA Extensions**
  - Stage 8: Generate the Digital Register
  - Stage 9: Generate the Indicators
  - Stage 10: Generate the Monitor

*Source: OECD ADIMA.*
3. Constructing ADIMA – detail

Stage 1. Choose the companies to profile

ADIMA currently covers 500 MNEs, however the methods and techniques used to compile this information are scalable and could be expanded to cover a wider sample. The current size was selected to achieve a good balance between relevance (for example, the top 500 MNEs account for a significant share - around 80% - of Global FDI) and feasibility (in terms of OECD internal resources and the OECD’s program of work). To give a further notion of their scale, the revenue of the ADIMA 500 is higher than the total output in the 2nd largest OECD economy, Japan.

Multinationals are selected for inclusion according to the following criteria:

- **An entity within the group is publically traded:** Annual Reporting is the primary data source for Stage 2, and, so, a pre-requisite for inclusion is a listing on a stock exchange;
  - Relevant Example: Samsung Electronics is a publically traded company and part of the wider Samsung Group (a private company). As a result (see also below), the entire Samsung Group was included in the ADIMA universe.

- **Market Capitalisation:** Entities were ranked by market capitalisation as of 31/12/2018 prior to sample selection of top 500. In cases where firms have dual listings, only the primary listing was considered.

- **Group structure not previously covered:** Complex chains of control mean that some firms that could be included within the universe under the previous two filtering conditions would be controlled by a firm that is also a member of the sample. These cases were excluded in order to avoid double counting.
  - Relevant Example: Hindustan Unilever Limited and Unilever are publically traded and both are within the world’s top 500 market capitalisations. However, Unilever is a majority shareholder and has economic control over Hindustan Unilever. As Unilever was already a member of the ADIMA universe due to its larger market capitalisation, Hindustan Unilever Limited was not included as a separate corporate entity, but rather as a part of Unilever.

Stage 2. Find relationships declared by Annual Reporting

To mirror the traditional business register approach, the controlled affiliates of each company are collected from Annual Reports (see also below). To best align with the FY2018 reference period the Annual Reporting for the financial year ending closest to 31/12/2018 is selected. For each controlled affiliate the following information was

---

15 This size also aligns well with those currently chosen by large case units (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).

16 Rugman (2005) for example showed that a relatively small set of MNEs (~500) was responsible for the overall majority (80 percent) of global FDI flows.
extracted: parent company, affiliate name, country and state of incorporation, country and state of operation, voting share and ownership share of the parent, type of consolidation in parent’s financial statements, source of information and date.

There exists no standardised format for reporting subsidiary information within Annual Reporting, and therefore data sources were prioritised with regards to the simplicity of data extraction. Sources were consulted in the following order:

- **SEC filings:** Filings with the US Security and Exchanges Commission (SEC) via their web portal Edgar\(^\text{17}\) are prioritised, as, in many cases, the format is harmonised, and in other cases the data are provided in an easily extractible text format. The majority of US companies file an Annual 10-K report and foreign companies with a US listing file an Annual 20-F report.
  - Formatted 10-K subsidiaries are obtained from the CorpWatch API\(^\text{18}\), which uses automated parsers to extract the subsidiary information from Exhibit 21 of the 10-K report.
  - In cases where this is unsuccessful (or for 20-F filings) text is extracted and formatted manually.
- **German Federal Business Register:** German Federal Law obliges companies to publish yearly consolidated financial statements with a full list of their subsidiaries in a legally defined format registered with the German Federal Business Register (Bundesanzeiger)\(^\text{19}\). These data are extracted and formatted manually.
- **Annual Reporting:** In absence of an alternative source, data are manually extracted from Annual Reporting made available by companies (usually in the Investor Relations sections of their websites). The majority of these reports are made available in PDF format, and a variety of tools were used in order to streamline this process (for example, automated table extraction and Optical Character Recognition (OCR) for images).

Data collection involves overcoming a number of hurdles including:

- **Language:** Data collectors in the ADIMA team had a relatively broad language coverage (English, French, German and Spanish) meaning the vast majority of firms could be researched and information found. However, a lack of knowledge of other languages (Chinese and Japanese in particular) were notable issues. Some companies within the sample only provided Annual Reports in these languages, meaning data collectors were unable to use Annual Reporting and thus limiting coverage. In other cases translations were provided, however not at the same level of timeliness as the original Annual Reporting. Further efforts to broaden the scope of languages covered will be undertaken in subsequent releases.
- **Inconsistent depth of reporting:** A number of companies only provide information on their most significant subsidiaries, whilst others report all subsidiaries. However, there is no uniform definition for “significant” (see the example in Section 2), and therefore it is often difficult to determine differences in

\(^{17}\) Accessible via: [https://www.sec.gov/edgar/searchedgar/companysearch.html](https://www.sec.gov/edgar/searchedgar/companysearch.html)

\(^{18}\) Accessible via: [http://api.corpwatch.org/](http://api.corpwatch.org/)

\(^{19}\) Accessible via: [https://www.bundesanzeiger.de/](https://www.bundesanzeiger.de/)
the depth of reporting between companies. This means that some companies within
ADIMA may have a greater coverage ratio than others. There exists no mechanism
to determine the actual coverage ratio but the wide range of sources used by
ADIMA does mean that ADIMA has higher coverage than any of the individual
sources.

- **Interpretation issues**: Data often proved difficult to interpret and extensive
  metadata was needed to validate data. In many cases at least one of the following
cases occurred:
  - Unknown consolidation basis;
  - Unknown whether location refers to incorporation or physical operations (this
    was especially true for special purpose entities);
  - Unknown whether ‘percentage of ownership’ refers to the percentage of a
    subsidiary/affiliate owned by the parent or the percentage controlled.

Due to these hurdles, extensive metadata was attached to each data point, enabling the
decisions made during data collection to be challenged during validation (Stage 7).

**Stage 3. Find relationships declared by big data sources**

To enhance the traditional business register approach (Stage 2) a number of additional data
sources are also considered (see Table 3.1 and Section 2).

All data sources are collected via a bulk data download or an Application Program Interface
(API). With bulk downloads, information is downloaded and scripts executed in order to
filter and harmonise the output structure. With APIs, scripts are developed to submit the
required queries, and results are returned and harmonised. The output of harmonised data
is therefore largely automated within Stage 3.

One example of data harmonisation is simplifications at the data collection stage. Firstly,
Telephone and Fax Numbers are categorised as a Contact Number. Secondly, websites and
e-mail addresses are harmonised to extract only the underlying domain. For example, the
and e-mail address [graham.pilgrim@oecd.org](mailto:graham.pilgrim@oecd.org), would both be harmonised to extract only
the domain [oecd.org](http://www.oecd.org/sdd/its/measuring-multinational-enterprises.htm).

All data sources have a shorter (more timely) update frequency than Annual Reporting, and
whilst ADIMA focusses on FY 2018, this does allow for the potential to produce data on a
close to real-time basis in the future. Furthermore, the collection scope of all sources listed
exceeds the 500 MNEs and thus provides scope to scale up beyond the top 500 MNEs.
Table 3.1. Big data sources within ADIMA

<table>
<thead>
<tr>
<th>Source</th>
<th>Collection Method</th>
<th>Average update frequency</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>PermID permid.org</td>
<td>Bulk Download</td>
<td>2 weeks</td>
<td>PermID provides a database detailing individual firms by the identifier PermID. Coverage is determined by companies deemed of interest by the database issuer. Data for each PermID includes Business Name, Address, Contact Numbers and website.</td>
</tr>
<tr>
<td>LEI gleif.org</td>
<td>Bulk Download</td>
<td>8 hours</td>
<td>LEI provides a database detailing individual firms by the identifier LEI. Coverage is determined by those companies requesting an LEI and legislation exists in a number of jurisdictions requiring an LEI for engaging in certain transactions. Data for each LEI includes Business Name, Address and Parent LEI.</td>
</tr>
<tr>
<td>Business Registers</td>
<td>National Sources</td>
<td>Bulk Download Depends on source</td>
<td>Business Registers detail each firm by the national Business ID. When available coverage tends to include all firms within the jurisdiction. Data availability depends on the source but can include Business Name, Address, Contact Numbers, Website, Email Address and Parent Business ID.</td>
</tr>
<tr>
<td>Common Crawl</td>
<td>commoncrawl.org</td>
<td>Bulk Download 3 months</td>
<td>The CommonCrawl compiles a network graph of hyperlinks existing between websites. ADIMA processes this data to find linked domains. A linked domain is defined when two domains have a matching second level domain (SLD) and both link to one another. In the case in Figure 2.1, if example.com links to example.fr and example.fr links to example.com as they have the same SLD (example) they are considered to be linked domains.</td>
</tr>
<tr>
<td>X.509 Certificates</td>
<td>opendata.rapid7.com</td>
<td>Bulk Download 2 weeks</td>
<td>X.509 certificates are heavily used in the security of online communications. They also determine the identity of the party being communicated, which is verified by a Certificate Issuing Authority (CA). Various levels of verification exist, but ADIMA utilizes only Organisation Level (OV) and Enterprises Level (EV) certificates as for issuance the CA has to verify the existence of the company and that they have the right to operate the underlying website. ADIMA processes this data to find Business Names, associated Business IDs and the operated domains.</td>
</tr>
<tr>
<td>WikiData</td>
<td>wikidata.org</td>
<td>API Real time</td>
<td>WikiData provides structured data sourced from underlying Wikipedia projects. In particular, for every Wikipedia concept (WikiID) which is identified as a business data can be downloaded for the WikiID of affiliates and associated websites.</td>
</tr>
<tr>
<td>Name Matching permid.org</td>
<td>API Real time</td>
<td></td>
<td>Business Names are declared in the Annual Reporting (Stage 2) and X.509 Certificates, but often company names are not harmonized (for example capitalization, acronyms and abbreviations) and no link exists to another identifier within the database. The process of name matching looks to map the numerous ways of referring to one identifier. In this case we use the API provided by the PermID database to map organization names onto the associated PermID.</td>
</tr>
</tbody>
</table>

Source: OECD ADIMA.

Stage 4. Combine data into a standardised data structure

In order to combine the data collected during Stage 2 and Stage 3 a network database is utilised. This involves extracting metadata and connections for each identifier collected. An identifier for a company may have metadata regarding its location and address, and connections for the affiliates that it declares within Annual Reporting. However in the case of an identifier for a website, no metadata exists, but connections to linked domains (via CommonCrawl) and companies which are authorised to operate on the domain (via X.509 certificates) do exist. Formally the identifiers are known as “vertices” and the connections as “edges”, and each has a standardised data structure within the ADIMA framework (and described in more detail below).

Each edge has a starting point (“From Vertex”), ending point (“To Vertex”) and an explanatory variable for the reason for the connection (“Relationship Type”). A further metadata variable to reflect the direction in which the points are linked (“Direction”) allows for Joint Ventures to be determined (discussed later). The aggregate set of edges defines the vertices, and metadata is appended to these vertices to provide detail on Name, Location and Addresses for the Vertex. This schema is summarised in Table 3.2.
Table 3.2. Data schema for Vertices and Edges within ADIMA

<table>
<thead>
<tr>
<th>Variable</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Vertices Data Schema</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Vertex</strong></td>
<td></td>
</tr>
<tr>
<td>Type</td>
<td>Options: PermID, LEI, Business Name, Business ID, Domain, Contact Number, WikiID</td>
</tr>
<tr>
<td>ID</td>
<td>Text</td>
</tr>
<tr>
<td><strong>Metadata</strong></td>
<td></td>
</tr>
<tr>
<td>Name</td>
<td>Text</td>
</tr>
<tr>
<td>Location</td>
<td>Text</td>
</tr>
<tr>
<td>Addresses</td>
<td>Text</td>
</tr>
<tr>
<td><strong>Edges Data Schema</strong></td>
<td></td>
</tr>
<tr>
<td><strong>From Vertex</strong></td>
<td></td>
</tr>
<tr>
<td>Type</td>
<td>Options: PermID, LEI, Business Name, Business ID, Domain, Contact Number, WikiID</td>
</tr>
<tr>
<td>ID</td>
<td>Text</td>
</tr>
<tr>
<td><strong>To Vertex</strong></td>
<td></td>
</tr>
<tr>
<td>Type</td>
<td>Options: PermID, LEI, Business Name, Business ID, Domain, Contact Number, WikiID</td>
</tr>
<tr>
<td>ID</td>
<td>Text</td>
</tr>
<tr>
<td><strong>Metadata</strong></td>
<td></td>
</tr>
<tr>
<td>Relationship</td>
<td>Options: HasAffiliate, HasLEI, HasContactNumber, HasDomain, HasBusinessID, HasParent, HasLinkedDomain, HasPermID</td>
</tr>
<tr>
<td>Direction</td>
<td>Options: ←, →, ↔</td>
</tr>
</tbody>
</table>

Source: OECD ADIMA.

The “Direction” variable defined within the data schema allows ADIMA to implement a directed network structure, which is necessary to be able to handle Joint Ventures and controls the flow of how searches are made within the ADIMA network. In general, where control exists the relationship of the edge is defined in both directions (“↔”), however in the case of a joint venture (two parties with joint control of the same company) the direction of the relationship needs to be defined from parent to child (“←”), or child to parent (“→”), to prevent two MNEs from being falsely connected.

Figure 3.1 gives a representation of this methodology in practice. In this case MNE Group 1 has declared that the Parent Company A has child Company B, whereas MNE Group 2 has declared that Parent Company C also has child Company B. With an undirected graph we would determine that Company A and Company C are part of the same MNE Group, however in this case as we only allow a path from Company C to Company B, from the perspective of MNE Group 1 Company C is unrelated to Company A, and likewise from the perspective of MNE Group 2 Company A is unrelated to Company C.
Table 3.3 provides a summary of the type of edges determined from the data found in Stage 2 and Stage 3, and gives a count of the number of edges within ADIMA. In total, ADIMA contains almost 10 million corporate relationships, of which 99% are found through big data sources (Stage 3).

Stage 5: Find a network structure for every parent MNE

This stage aims to determine all of the identifiers which are linked to a parent MNE, whether physical or digital. In simple terms, this involves finding all identifiers which can be linked to the parent MNE via the paths declared within ADIMA. From a technical perspective, ADIMA performs an iterative search for connected edges beginning from the vertex of the parent MNE to find vertices of type: Business Name, Business ID, PermID, LEI, Domain, Contact Number and WikiID. The search takes into account the directional variables, meaning that joint ventures are handled. The average network for an MNE contains 650 vertices, which is comparatively large to other initiatives such as the EGR.
which have an average of roughly 7 vertices per MNE group, suggesting that ADIMA is focusing on some of the most complex cases.  

A partial implementation of the iterative search is shown for Apple Inc. in Figure 3.2. The process begins from the vertex of the PermID with the value 4295905573 (Apple Inc). PermID 4295905573 is linked to the domain name apple.com, and also has an associated LEI HWUPKR0MPOU8FGXBT394. In addition, LEI 549300QKDHYRRQH2MB86 defines its parent as LEI HWUPKR0MPOU8FGXBT394. Therefore with only three edges we have managed to determine a number of identifiers for companies within the MNE Group and find a controlled domain.

---

20 The maximum number of vertices in ADIMA is 7,142 (Goldman Sachs), whilst the minimum is 3 (China Tower Corp).
Stage 6: Generate the Physical Register

The network determined in Stage 5 for each MNE is used to generate the physical register. As no unique vertex type exists for every affiliate, and ADIMA has to be able to map between multiple vertex types and determine equivalent concepts, we define two concepts “Physical Vertices” and “Equivalent Edges”.

A vertex is defined as physical if it relates to a physical company. Therefore vertex types of PermID, LEI, Business Name and Business ID are defined as physical as they relate to physical companies.
An edge is defined as equivalent if it provides a one-to-one mapping between two identifiers for the same physical company. Therefore, edge relationship types of *Has LEI*, *Has Business ID* and *Has PermID* are defined as equivalent.

To identify distinct companies in the absence of non-unique identifiers, we find clusters of physical nodes which are connected by equivalence relationships for a given MNE network. Figure 3.3 gives a visual representation of this process, where three affiliate clusters are identified for Company A, Company B and Company C. The level of detail available for each cluster depends on the vertices which form the cluster, with the most detailed cluster containing information for Business Names, PermIDs, LEIs and Business IDs for the given company.

**Figure 3.3. Determining the unique companies belonging to an MNE network**

Each entry within the physical register relates to an affiliate cluster. Table 3.4 details the presented variables and calculation method used for the Physical Register. The level of detail available for each affiliate cluster depends on the vertices which form the affiliate cluster, and therefore a coverage ratio is also presented.
### Table 3.4. Variables within the Physical Register

<table>
<thead>
<tr>
<th>Parent MNE</th>
<th>Description</th>
<th>Calculation Method</th>
<th>Coverage Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Affiliate Name</td>
<td>Name of the parent MNE</td>
<td>The name of the source vertex used for determining the Group Network</td>
<td>100%</td>
</tr>
<tr>
<td>Affiliate Name</td>
<td>Name of the affiliate</td>
<td>The name of the affiliate identified by the most trusted source within the affiliate cluster. The most trusted source is obtained from the first source with available data in the following order: Country Business Registers, LEI, Annual Reporting, X.509 Certificates and PermID.</td>
<td>100%</td>
</tr>
<tr>
<td>Jurisdiction</td>
<td>Jurisdiction of the affiliate</td>
<td>The jurisdiction of the affiliate identified by the most trusted source within the affiliate cluster</td>
<td>99%</td>
</tr>
<tr>
<td>Joint Venture</td>
<td>Is the affiliate a joint venture?</td>
<td>An affiliate cluster is a joint venture if there does not exist a path from the source vertex to the affiliate cluster when one directional edges are removed</td>
<td>100%</td>
</tr>
<tr>
<td>Business ID</td>
<td>Business ID for the affiliate</td>
<td>The Business IDs for vertices of type &quot;Business ID&quot; within the affiliate cluster</td>
<td>24%</td>
</tr>
<tr>
<td>LEI</td>
<td>LEI for the affiliate</td>
<td>The LEIs for vertices of type &quot;LEI&quot; within the affiliate cluster</td>
<td>27%</td>
</tr>
<tr>
<td>PermID</td>
<td>PermID for the affiliate</td>
<td>The PermIDs for vertices of type &quot;PermID&quot; within the affiliate cluster</td>
<td>65%</td>
</tr>
<tr>
<td>Alternative Names</td>
<td>Alternative names which have been used within ADIMA to refer to the affiliate</td>
<td>The names of the affiliate identified by sources which were not presented in the &quot;Affiliate Name&quot; variable</td>
<td>41%</td>
</tr>
<tr>
<td>Address</td>
<td>Address of the affiliate</td>
<td>The address of the affiliate identified by the most trusted source within the affiliate cluster</td>
<td>64%</td>
</tr>
<tr>
<td>Affiliate Distance</td>
<td>The number of steps required in the Network Graph to discover the affiliate</td>
<td>The minimum distance from the source node within the affiliate cluster</td>
<td>100%</td>
</tr>
</tbody>
</table>

**Note:** Some coverage ratios are significantly less than 100%, and this reflects available data, for example as the LEI covers only a subset of firms, generally limited to those engaging in financial transactions, it cannot be expected to give complete coverage.

**Source:** OECD ADIMA.

### Stage 7: Validate the Physical Register

As previously mentioned, the average network for each MNE within ADIMA contains 650 vertices and a large number of these vertices are only discovered after a large number of edges have been searched. It is therefore necessary to consider how to validate this information to ensure data quality.

The primary objective of the validation approach in ADIMA is to find a path from the source vertex to all affiliate clusters which is of reasonable quality. A judgement is made about the quality (or reliability) of each edge depending on the source and the relationship type to determine a list of sources and relationship types which are considered of high enough quality to be automatically validated. Table 3.5 provides an extension to Table 3.3 to determine which relationships are automatically validated. The general rule is that ‘Equivalent Edges” and ownership relationship defined by authoritative sources are validated; in total 33% of edges are determined to be of high enough quality to automatically validate. The remaining 66% require further validation checking however could be validated through a combination of already automatically validated links which verify a path (ie. A to B to C validates the A to C relationship).
Table 3.5. Summary of Edges within ADIMA by Validation Type

<table>
<thead>
<tr>
<th>Stage</th>
<th>Source</th>
<th>From Vertex Type</th>
<th>Relationship Type</th>
<th>To Vertex Type</th>
<th>Count in ADIMA</th>
<th>Automatic Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stage 1</td>
<td>Annual Reporting</td>
<td>PermID</td>
<td>Has Affiliate</td>
<td>BusinessID</td>
<td>105,666</td>
<td>Yes</td>
</tr>
<tr>
<td>Stage 2</td>
<td>PermID</td>
<td>PermID</td>
<td>Has LEI</td>
<td>LEI</td>
<td>1,338,349</td>
<td>Yes</td>
</tr>
<tr>
<td>Stage 2</td>
<td>PermID</td>
<td>PermID</td>
<td>Has Contact Number</td>
<td>Contact Number</td>
<td>1,656,684</td>
<td>No</td>
</tr>
<tr>
<td>Stage 2</td>
<td>PermID</td>
<td>PermID</td>
<td>Has Domain</td>
<td>Domain</td>
<td>888,025</td>
<td>No</td>
</tr>
<tr>
<td>Stage 2</td>
<td>LEI</td>
<td>LEI</td>
<td>Has Business ID</td>
<td>Business ID</td>
<td>1,197,855</td>
<td>Yes</td>
</tr>
<tr>
<td>Stage 2</td>
<td>LEI</td>
<td>LEI</td>
<td>Has Parent</td>
<td>LEI</td>
<td>142,335</td>
<td>Yes</td>
</tr>
<tr>
<td>Stage 2</td>
<td>Country Business Registers</td>
<td>Business ID</td>
<td>Has Contact Number</td>
<td>Contact Number</td>
<td>732,775</td>
<td>No</td>
</tr>
<tr>
<td>Stage 2</td>
<td>Country Business Registers</td>
<td>Business ID</td>
<td>Has Domain</td>
<td>Domain</td>
<td>507,471</td>
<td>No</td>
</tr>
<tr>
<td>Stage 2</td>
<td>Country Business Registers</td>
<td>Business ID</td>
<td>Has Parent</td>
<td>Business ID</td>
<td>26,827</td>
<td>Yes</td>
</tr>
<tr>
<td>Stage 2</td>
<td>Common Crawl</td>
<td>Domain</td>
<td>Has Linked Domain</td>
<td>Domain</td>
<td>1,486,604</td>
<td>No</td>
</tr>
<tr>
<td>Stage 2</td>
<td>X.509 Certificates</td>
<td>Business Name</td>
<td>Has Business ID</td>
<td>Business ID</td>
<td>105,131</td>
<td>Yes</td>
</tr>
<tr>
<td>Stage 2</td>
<td>X.509 Certificates</td>
<td>Business Name</td>
<td>Has Domain</td>
<td>Domain</td>
<td>1,020,061</td>
<td>No</td>
</tr>
<tr>
<td>Stage 2</td>
<td>WikiData</td>
<td>WikiID</td>
<td>Has Affiliate</td>
<td>WikiID</td>
<td>9,606</td>
<td>No</td>
</tr>
<tr>
<td>Stage 2</td>
<td>WikiData</td>
<td>WikiID</td>
<td>Has Domain</td>
<td>Domain</td>
<td>110,584</td>
<td>No</td>
</tr>
<tr>
<td>Stage 2</td>
<td>Name Matching</td>
<td>Business Name</td>
<td>Has PermID</td>
<td>PermID</td>
<td>260,692</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Source: OECD ADIMA.

Figure 3.4 gives a representation of the validation scope. The source vertex is assumed to be the PermID for Company A, as all “equivalent edges” are automatically validated, this means that the contents of the cluster for Company A are deemed to be valid. The path from the PermID for Company A to the Business Name Company B LLC is of sufficient quality as it is sourced from Annual Reporting, therefore the cluster for Company B is deemed to be valid as well. However, the path from Company A to Company C can only be reached via digital sources which are not of sufficient quality to validate automatically, as a result the relationship to Company C needs further investigation before it can be included in ADIMA.

Figure 3.4. Determining the scope of validation in ADIMA

Source: OECD ADIMA.

For each company not currently validated two options are possible:
• **Valid:** Company is manually added into the validation scope from the perspective of the parent company. In the example above, if Company C was determined to be a child of Company A we would validate this relationship.

• **Invalid:** Reasons for how the link has been determined are investigated, and the cause of the issue is resolved. This normally consists of deleting an incorrect edge, or changing the direction of the edge to reflect a Joint Venture which was not declared in the raw data sources. In the example above, if Company C was determined not to be a child of Company A we would investigate the paths for determining this relationship, a common case is that Company C may provide services to the domain example.fr, and therefore this relationship should be removed.

Validation is completed when there does not exist a physical company which has not been validated, Step 5 and Step 6 are subsequently repeated and the process continues until no further validations are needed. In total for the ADIMA FY 2018 release, over 25,000 companies were validated, and 12,000 relationships were edited. Given the fact that ADIMA FY 2018 consists of over 115,000 subsidiaries, this means that validation was necessary for around 20% of affiliates.

The quantity of data validation necessary for ADIMA requires a team of data validators, and in order to ensure consistency and reproducibility, each data validator rigorously followed the same process using a purpose built interface.

In order to validate a company, each validator performed desk research attempting to locate two independent and up to date sources confirming the association between legal entities. Due to the extremely large workload, this rule was relaxed in cases where the connection was obvious (e.g. self-explanatory name, however there are weaknesses to this approach, as the data review from individual MNEs displayed). The validation team had to employ personal judgement when performing desk research regarding the reliability of each source. However, in most cases validations were achieved using sources which the validators deemed of high quality, such as company websites or online business directories and registers.

Certainty was unattainable in a limited number of cases. In these cases the data point was allowed if it was deemed highly probable that it was part of the MNE group, and metadata associated with the validation explaining the decision process that had been applied was documented. In general, limited confidence in the relationship was attributable to one of the following reasons:

• **Language:** A lack of knowledge of other languages (Chinese and Japanese). In some cases, despite the use of a translation service sufficient information could not be found to validate relationships.

• **Non-existence of Information:** In some cases only the data derived by ADIMA could verify the relationship between the parent and associated affiliate. This was particularly applicable to special purpose entities located in jurisdictions with less stringent reporting frameworks.

• **Issues following corporate events:** Following corporate events it was often difficult to determine how the action had been executed. The boundary between assets and the corporate entity was often difficult to determine, and furthermore resulting liquidations and renaming actions were complex to track.
Where invalid relationships were discovered, the source of the error was researched and the invalid relationship (edge) deleted. Excluding generic data errors, frequently occurring errors included:

- **Non-distinct business register numbers**: For a number of jurisdictions a harmonised country level business register does not exist. In these cases there was the possibility that two businesses held the same identification number, therefore business numbers from these jurisdictions were excluded.

- **Service providers**: Due to the exploratory way in which ADIMA derives relationships between companies when services were provided by external service providers this could lead to false connections (such as Company E in Figure 2.1). Examples include shared telephone numbers, the provision of a social media profile, use of an external company for email services and provision of website hosting or management services.

Company specific errors were also present and these were generally attributable to one of the following:

- **Corporate events**: Corporate events were often not immediately reflected by all sources. Examples of corporate events include mergers, spin-offs, divestments and name changes. A common source of error was outdated contact numbers following a corporate event.

- **Joint Ventures**: Joint Ventures were often not declared and handled as regular corporate relationships. In these cases connections between two MNEs were often present and “Direction” variables needed to be applied to edges to reflect the directional nature of the joint venture.

Given each case and decision was unique, comprehensive metadata was recorded to justify the decision process. Complex cases were referred to the most senior member of the validation team for in-depth research. Box 3.1 discusses one such case and its handling within ADIMA.

### Box 3.1. Handling of Atomic Weapons Establishment (AWE) in ADIMA

AWE Plc has a shared shareholder ownership structure as follows:

- Lockheed Martin: 51%
- Jacobs Engineering Group: 24.5%
- Serco Group: 24.5%

In addition to this structure the UK Government owns a “Golden Share”, which at present has no voting rights and no entitlement to profits, but does allow the UK government to take control of the company if deemed necessary.

Whilst the UK Government owns AWE sites and facilities, they are operated under a long term lease. The existence of the long term lease and over 50% ownership share of AWE Plc means that ADIMA considers AWE Plc to be part of only the Lockheed Martin Group.

As an additional validation step, the ADIMA team is engaging with individual MNEs to validate and correct their information where errors are found. This process of shared validation and collective ‘ownership’ of the data available in ADIMA by individual MNEs
is important, as this helps to further build trust within the dataset. The process involves making the Physical and Digital Register (Stage 8) available to the MNE, and implementing alterations to the network graph where necessary to respond to individual feedback.

Stage 8: Generate the Digital Register

The network determined in Stage 5 for each MNE is also used to generate the Digital Register. The digital register consists of vertices from the network which are of type Domain.

Each entry in the digital register relates to an individual domain belonging to a MNE and Table 3.6 details the presented variables. The data for each domain are extended by joining two measures of global importance at the domain level from the Common Crawl Project\(^{21}\). The first measure, Page Rank looks to reflect the percentage chance that a random web user (randomly clicking links) has of being on a given domain. The second measure, Harmonic Centrality, looks to reflect a measure of the distance that other domains are from the given domain.

\(^{21}\) Accessible via: https://commoncrawl.org/2019/05/host-and-domain-level-web-graphs-febmarapr-2019/
### Table 3.6. Variables within the Digital Register

<table>
<thead>
<tr>
<th>Description</th>
<th>Calculation Method</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Parent MNE</strong></td>
<td>Name of the Parent MNE</td>
<td>100%</td>
</tr>
<tr>
<td>Domain</td>
<td>Domain controlled by the Parent MNE</td>
<td>100%</td>
</tr>
<tr>
<td><strong>TLD</strong></td>
<td>The Top Level Domain (TLD) of the Domain</td>
<td>100%</td>
</tr>
<tr>
<td>Jurisdiction</td>
<td>Jurisdiction of the TLD</td>
<td>54%</td>
</tr>
<tr>
<td><strong>Page Rank Value</strong></td>
<td>The Page Rank of the domain</td>
<td>77%</td>
</tr>
<tr>
<td>Iteratively calculated where:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$PR(x) = \sum_{y \in B_x} PR(y) / L(y)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Where $PR(x)$ represents the Page Rank of domain $x$, $B_x$ represents the set of all domains linking into Domain $x$ and $L(y)$ is the number of links from domain $y$</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Page Rank Position</strong></td>
<td>The rank of the domains Page Rank relative to all domains covered by the Common Crawl</td>
<td>77%</td>
</tr>
<tr>
<td></td>
<td>Global Ranking of Page Rank Value</td>
<td></td>
</tr>
<tr>
<td>Harmonic Centrality Value</td>
<td>The Harmonic Centrality of the Domain</td>
<td>77%</td>
</tr>
<tr>
<td>Calculated by:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$H(x) = \sum_{y \in C_x} 1 / D(x, y)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Where $H(x)$ represents the Harmonic Centrality of domain $x$, $C_x$ represents the set of domains where there exists a path from $x$ and $D(x, y)$ represents the minimum distance path between domain $x$ and $y$.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Harmonic Centrality Position</td>
<td>The rank of the domains Harmonic Centrality relative to all domains covered by the Common Crawl</td>
<td>77%</td>
</tr>
<tr>
<td></td>
<td>Global Ranking of Harmonic Centrality Value</td>
<td></td>
</tr>
<tr>
<td>Domain Distance</td>
<td>The number of steps required in the Network Graph to discover the domain</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>The minimum distance from the source node to the domain</td>
<td></td>
</tr>
</tbody>
</table>

*Source: OECD ADIMA*

### Stage 9: Generate the Indicators

The network determined for each MNE (Stage 5), Physical Register (Stage 6) and Digital Register (Stage 8) are compiled to produce a set of harmonised indicators including measures on the level of internationalisation.

A number of indicators are included in ADIMA (see also Table 3.7 and Box 3.2). The *International share* indicator reflects the share of affiliates which are not within the parent company’s jurisdiction, the closer the measure is to zero the less international the MNE. Indicators referred to as C3 and C5 give measures of the share of the largest 3 and 5 countries respectively. The *Herfindahl–Hirschman Index* (HHI) gives a measure of the concentration of affiliates belonging to particular jurisdictions, the closer the measure to zero the less concentrated the MNE.
Table 3.7. Variables within the Indicators

<table>
<thead>
<tr>
<th>Description</th>
<th>Calculation Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parent MNE</td>
<td>Name of the Parent MNE</td>
</tr>
<tr>
<td>Headquarters of Parent MNE</td>
<td>Headquarters of the Parent MNE</td>
</tr>
<tr>
<td>Number of Affiliates</td>
<td>Number of discovered affiliates for each MNE</td>
</tr>
<tr>
<td>Number of Domains</td>
<td>Number of discovered domains for each MNE</td>
</tr>
<tr>
<td>Number of Jurisdictions declared in Annual Reporting</td>
<td>Count of the number of jurisdictions declared within Annual Reporting for each MNE</td>
</tr>
<tr>
<td>Number of Jurisdictions with a Physical Presence</td>
<td>Count of the number of jurisdictions discovered with presence in the Physical Register for each MNE</td>
</tr>
<tr>
<td>Number of Jurisdictions with Presence</td>
<td>Count of the number of jurisdictions discovered with presence in either the Physical or Digital Register for each MNE</td>
</tr>
<tr>
<td>International Share</td>
<td>Share of discovered affiliates in the Physical Register which are not within the Jurisdiction of the Parent MNE for each MNE</td>
</tr>
<tr>
<td>C3</td>
<td>Share of discovered affiliates in the Physical Register which are within the 3 jurisdictions with the most affiliates for each MNE</td>
</tr>
<tr>
<td>C5</td>
<td>Share of discovered affiliates in the Physical Register which are within the 5 jurisdictions with the most affiliates for each MNE</td>
</tr>
<tr>
<td>HHI</td>
<td>The Herfindahl–Hirschman Index</td>
</tr>
<tr>
<td>Discovery Method By Jurisdiction</td>
<td>Discovery method of existence for each Parent MNE within each jurisdiction</td>
</tr>
</tbody>
</table>

Source: OECD ADIMA.

Box 3.2. Example of internationalisation measures for a selection of companies

Companies that are **primarily domestically focused** tend to have low internationalisation shares, C3 or C5 that are equal to one and HHI measures very close to one for example.

As a specific example, ADIMA shows that only 0.4% of affiliates for Dominion Energy Inc are outside of the Headquarter jurisdiction (USA), and with C3 and C5 ratios equal to one as Dominion Energy is present in 2 jurisdictions namely the United States and Canada,
according to ADIMA. These companies tend to be those where a sale activity is carried out in one jurisdiction, but a support function is carried out from a different jurisdiction.

Companies that show **concentrated markets** tend to have international shares that are relatively high, C3 and C5 ratios are close to one, and HHI measures are relatively low. For example, ADIMA shows that 72.4% of Orsted A/S affiliates (another energy company) are outside of the Headquarter jurisdiction (Denmark), however 96% are present within only 5 jurisdictions (United States, Denmark, the United Kingdom, Germany and the Netherlands) – indicating that its activity is highly concentrated to a small number of jurisdictions.

Companies with a **global focus** tend to have high international shares, C3 and C5 ratios are low, and low HHI measures. For example. ADIMA discovers that 97.2% of Mondelez International’s (a food company) affiliates are outside of the Headquarter jurisdiction (USA), and only 34% are present within the top 5 jurisdictions – furthermore a low HHI score (0.038) indicates notable presence in a high number of jurisdictions. In fact, ADIMA finds physical affiliates within 85 jurisdictions.

<table>
<thead>
<tr>
<th>Classification</th>
<th>International Share</th>
<th>C3</th>
<th>C5</th>
<th>HHI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dominion Energy Inc</td>
<td>Primarily domestically focused</td>
<td>0.004</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Anthem Inc</td>
<td>Primarily domestically focused</td>
<td>0.022</td>
<td>0.99</td>
<td>1.00</td>
</tr>
<tr>
<td>Welltower Inc</td>
<td>Concentrated Markets</td>
<td>0.409</td>
<td>0.95</td>
<td>1.00</td>
</tr>
<tr>
<td>Orsted A/S</td>
<td>Concentrated Markets</td>
<td>0.724</td>
<td>0.81</td>
<td>0.96</td>
</tr>
<tr>
<td>Mondelez International Inc</td>
<td>Global Focus</td>
<td>0.972</td>
<td>0.27</td>
<td>0.34</td>
</tr>
<tr>
<td>Nestle SA</td>
<td>Global Focus</td>
<td>0.932</td>
<td>0.26</td>
<td>0.34</td>
</tr>
</tbody>
</table>

**Source**: OECD ADIMA.

**Stage 10: Generate the Monitor**

The network determined in Stage 5 for each MNE is also used to generate the monitoring tool. The tool is based on analysing media coverage on a daily basis in order to identify events of significant interest relating to each MNE within ADIMA.

To identify search terms for media coverage for each MNE the vertices which are of type WikiData are extracted from the network graph. For each WikiData identifier the daily page views and names are extracted for the corresponding English Wikipedia article. Due to the large number of WikiData identifiers those that are not deemed significant for the given MNE are excluded. The significance threshold is determined by 20% of the yearly page views for all English Wikipedia pages belonging to the MNE. Media coverage is then searched using the given WikiData name as a search term. Information on media coverage is extracted from the Global Database of Events, Language

---

22 More details available at: https://wikitech.wikimedia.org/wiki/Analytics/AQS/Pageviews
and Tone (GDELT\textsuperscript{23}), which provides information on the strength of media coverage and the headlines for the Top 10 matching articles.

The headlines extracted are then processed using Natural Language Processing (NLP). NLP is a technique for textual analysis that 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, Location or Monetary Values. This has been automated in Stanford University’s NLP engine CoreNLP (see Manning et al., 2014\textsuperscript{24}), which was used to intelligently classify headlines within ADIMA.

Each entry in the Monitor corresponds to a daily measure of each MNE’s most important WikiData identifiers and a number of indicators are presented in order to filter media coverage and determine a set of “events” (Table 3.9). These include Wikipedia page views and media coverage intensity, including indicators as to whether these are outside of normal boundaries. NLP results are also presented, allowing filtering on verbs as well as Locations, Organisations, People and Monetary Values. An example for a particular corporate event is given in Box 3.3.

\textsuperscript{23} More details available at: \url{https://www.gdeltproject.org/}

\textsuperscript{24} More details available at: \url{https://stanfordnlp.github.io/CoreNLP/}
Table 3.9. Variables within the Monitor

<table>
<thead>
<tr>
<th>Description</th>
<th>Calculation Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parent MNE</td>
<td>Name of the Parent MNE</td>
</tr>
<tr>
<td>WikiID</td>
<td>The identifier code for the WikiID</td>
</tr>
<tr>
<td>WikiTopic</td>
<td>The English Wikipedia topic for the WikiID</td>
</tr>
<tr>
<td>Date</td>
<td>Date</td>
</tr>
<tr>
<td>Annual Wikipedia views (MNE Level)</td>
<td>A summation of all Wikipedia page views for the year 2019 for all Wikipedia topics belonging to the MNE</td>
</tr>
<tr>
<td>Annual Wikipedia views (WikiID level)</td>
<td>A summation of all Wikipedia page views for the year 2019 for the Wikipedia topic belonging to the WikiID</td>
</tr>
<tr>
<td>Annual Significance</td>
<td>Calculated by:</td>
</tr>
<tr>
<td>Daily Wikipedia views</td>
<td>Extracted from the Wikipedia page views API</td>
</tr>
<tr>
<td>Excess Wikipedia views</td>
<td>Excess daily page views from a monthly model with a trend component and handling for weekday effects</td>
</tr>
<tr>
<td>Wikipedia Anomaly?</td>
<td>Positive Excess page views which is greater than six times the distribution of the 25% to 75% inner quartile range of Excess page views are considered to be an anomaly</td>
</tr>
<tr>
<td>Daily Media Interest</td>
<td>Extracted from GDEL API for search terms defined by the WikiID</td>
</tr>
<tr>
<td>Excess Media Coverage</td>
<td>Excess media coverage from a monthly model with a trend component and handling for weekday effects</td>
</tr>
<tr>
<td>Media Anomaly?</td>
<td>Positive Excess Media Coverage which is greater than six times the distribution of the 25% to 75% inner quartile range of Excess Media Coverage are considered to be an anomaly</td>
</tr>
<tr>
<td>Extracted Locations</td>
<td>Extracted via NLP with count of frequency</td>
</tr>
<tr>
<td>Extracted Organisations</td>
<td>Extracted via NLP with count of frequency</td>
</tr>
<tr>
<td>Extracted People</td>
<td>Extracted via NLP with count of frequency</td>
</tr>
<tr>
<td>Greatest Monetary Value</td>
<td>Extracted via NLP with the maximum value taken</td>
</tr>
<tr>
<td>Greatest Monetary Value with an Organisation</td>
<td>Extracted via NLP with the maximum value taken when an Organisation exists in each of the top 10 headlines.</td>
</tr>
<tr>
<td>Extracted Verbs</td>
<td>Extracted via NLP with count of frequency</td>
</tr>
<tr>
<td>Headlines</td>
<td>Extracted from GDEL API for search terms defined by the WikiID</td>
</tr>
</tbody>
</table>

Source: OECD ADIMA.

Box 3.3. Example of analysing a corporate event using the ADIMA monitoring tool

AbbVie to buy Allergan

On 25th June 2019 the ADIMA Monitor detected two unusual activity signals for the company AbbVie Inc (a US based pharmaceutical company). Firstly, Wikipedia page views for the day totalled 8,146 compared to an expected value of 1,404 - almost 5 times above normal expectations. Secondly, 0.12% of media articles for the day mentioned AbbVie – almost 15 times larger than normal expectations.
Given this unusual level of activity, the ADIMA monitor extracted the top 10 headlines (Table 3.10) for the given day to determine whether the event was significant (i.e. provided an indication of a significant corporate restructuring, which would have an impact on official macro-economic statistics). Using Natural Language Processing (NLP), the monitor was able to determine that the organisations AbbVie and Allergan are referenced, the verbs include “buy” and “acquire” are utilised, and the largest numerical value is 63 billion – a clear indicator of ‘significance’.

Table 3.10. Headlines for AbbVie (26th June 2019)


Source: OECD ADIMA

At the time of writing the transaction is yet to complete (expected Q2 2020), but this is an indication of a large corporate event that was will likely impact on Irish statistics in particular. Allergan is headquartered in Ireland, whilst AbbVie is headquartered in the United States, and hence this represents a significant foreign investment by a US firm within Ireland.
4. Conclusions

ADIMA is ready to act as a single reference point for the statistical community on data regarding the activities of MNEs, and this paper shows it is possible to develop insights into these MNEs in greater detail using only publically available data.

ADIMA applications

ADIMA provides a flexible framework for developing a number of outputs that allow for the profiling of MNEs in greater detail, and further outputs will be developed as the project matures and expands. A number of key use cases can already be identified however:

- **Supporting research:** Obviously, and by design, the granular nature of the data provides a means to analyse individual MNEs, and in particular the scale and nature of their cross-border activities.

- **Aiding work of NSOs:** As illustrated in SDD/CSSP(2018)6 the measurement of flows related to MNEs and determining where they are registered is presenting significant challenges for the measurement of GDP. This has led to growing demands for improved statistics that provide insights on the scale and complexity of international MNE activity as well as timely information on any restructurings they undertake. ADIMA provides a single reference point, free from confidentiality restrictions that may impede discussions between NSOs, and features a monitoring tool to help identify significant corporate restructurings.

- **Building on and contributing to other initiatives with a focus on MNEs:** A number of important international efforts are being developed to better understand the structure and importance of MNEs. ADIMA aims to complement these and the OECD are working in close collaboration with GLEIF (for the Legal Entity Identifier (LEI)), and Eurostat’s teams developing the Euro Groups Register (EGR) and Early Warning System (EWS).

- **Building on and contributing to other initiatives with a focus on Big Data analysis:** Smart data infrastructure forms the foundation of the ADIMA project. The team hopes to share this experience and knowledge, becoming a key use case for the OECD Smart Data Strategy and providing support and assistance to NSOs in the implementation of web-scraping and text analysis techniques. The recent investigation to the gender composition of boards25 within ADIMA companies provides further evidence of the usefulness of this data beyond the traditional statistical community.

Improving ADIMA through outreach

Although the tools currently developed provide good coverage of MNE activities they cannot guarantee exhaustiveness. Strategic outreach through a number of channels is necessary to further improve the coverage of ADIMA.

---

Outreach with National Statistical Offices

Adopting ADIMA’s framework as a data source for information on MNEs is the first step a number of countries can take to help in the profiling of firms (including for surveys) whilst also potentially, directly improving the quality of foreign affiliate trade statistics, balance of payments statistics and modes of supply statistics in countries.

The OECD supports more active engagement with NSOs to formalise the use of ADIMA in national statistical infrastructures but also as a global resource. Countries are encouraged to:

- Compare the scope of MNE data (i.e. coverage of units) they currently have and use with ADIMA;
- Consider the scope for integrating business register data (which is often not confidential) within the ADIMA framework;
- Work with the OECD to identify non-disclosive mechanisms to quality-assure and validate ADIMA data.

Outreach with other initiatives

A number of international initiatives can benefit from ADIMA.

- The Global Legal Entity Identifier Foundation (GLEIF) are seeking to create a harmonised identification number (LEI) of all entities worldwide and ADIMA data are providing a framework for assessing current coverage rates and guiding strategies for improving data accuracy and coverage.
- ADIMA’s monitoring tool complements Eurostat’s Early Warning System (EWS) for large corporate events by providing additional indicators and the ability to simulate the impacts of a company restructuring, it also provides a framework to shift the Euro Groups Register (EGR) from a European to a Global perspective.

Outreach with Multinationals

MNEs can help to improve the accuracy of the initiative by engaging to validate and enhance current data. Engagement can be beneficial for both the OECD and responding firms. Increasing the quality and coverage of data improves our collective understanding of MNEs and their global reach, which is fundamental for many of today’s global challenges but it is also a tool and resource that provides valuable information about the firms themselves, and the benefits they can provide in the economies where they operate.

Whilst the tools that we use to estimate firms’ activities are as robust as possible, direct responses from the firms to a dedicated (and voluntary) questionnaire, will, potentially significantly, enhance quality, and provide a direct mechanism for firms to validate our data.
References


