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

**Keywords:** OECD, Big Data, Multinational Enterprise, MNE, affiliates, open data, web analytics, register, indicators, early warning system, API, digital, risk management.

# Introduction

Despite their significant and growing importance, with implications across a range of policy areas, information on Multinational Enterprises (MNEs) remains at best patchy. 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 the global activities of firms headquartered in the economy (and even in these cases it is not clear that the coverage of the MNE’s 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, at least for now, possible.[[1]](#footnote-1)

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.

ADIMA attempts to meet these goals through the development of three distinct (but related) outputs for 100 of the largest MNEs (by sales), comprising MNEs headquartered in 16 countries, by the end of 2018, with the aim to increase coverage to around 500 MNEs by 2020:

* a Register of MNE affiliates and presence;
* 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.

ADIMA is part of a growing international response in the area of MNE data, which ADIMA builds on through partnerships and, in turn, complements through its focus on publishing open, harmonised, and traceable data, including 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 corporate entities worldwide. This paper provides an overview of the methodology employed in compiling each individual MNE ‘Knowledge Graph’, the core tool underlying **ADIMA outputs**.

# Methods and Results (forthcoming)

An important innovation of ADIMA is the leveraging of publicly available Big Data within a structured framework. The database combines traditional data sources, such as **annual company reports**, with newly emerging sources, such as **Thomas Reuters PermID**, the **Legal Entity Identifier (LEI)** and data contained within **company websites**, using innovative data collection methods (XBRL, web-scraping and text analytics) and Big Data analytics (Spark). The use of a wide number of sources with innovative analytics means ADIMA forms a significant step towards operationalising the vision set out in the OECD Smart Data Framework.[[2]](#footnote-2)

The methodology underlying ADIMA is to link the above data sources and determine which pieces of information belong to the same MNE family. In order to obtain this information the ADIMA database implements a graph database, similar to databases underlying modern social media platforms, these databases have a structure that allows for the relationships between information sets to be just as important as the underlying information.

ADIMA outputs are determined via calculations over the graph database. **Register** components refer to the family of interconnected data within the graph database. **Indicator** components refer to economic variables assigned to the Parent MNE (or ‘root node’). **Monitor** components refer to changes in structure of the graph database that have occurred over time.

### **A Register of parent and affiliate relationships**

The Register components are determined firstly by obtaining a set of data points that belong to the same MNE family. In order to obtain this information the Depth First Algorithm explores from the Parent MNE onwards. Each underlying data point has indicators of geography, and, therefore, a register of presence by country can quickly be determined for each MNE. To explain how the MNE family is determined the following presents a simplified (and more linear than in reality, and one directional) example for the **French MNE Total SA.** In practice, as shown in Figure 1 and described in more detail below, the connections between data points can be verified by numerous paths and results in a complex web of information.

**Figure 1. Data sources within the graph database and their relationships**



Source: OECD ADIMA

The **PermID Organisation Database** is the first input into the knowledge graph, it is the largest open source freely downloadable database of companies known, and provides the most complete starting point for analysis. The database provides information on Legal Name, Jurisdiction, Sector of Activity, LEI and Websites. The entry for Total SA details a number of identifiers we can use to link Total SA to other sources (Table 1).

**Table 1. Links from the Total SA entry in the PermID Organisation Database**

|  |  |  |
| --- | --- | --- |
| **Identifier Type** | **Value** | **Reference in Figure 1** |
| LEI | 529900S21EQ1BO4ESM68 | Label (1) |
| Website | total.com | Label (2) |

Source: <https://permid.org/1-5001170594>

Websites are generally owned and operated uniquely by a single company, and therefore the Website declared is used in order to build more links to the initial company. Firstly, the underlying website of other companies within the PermID Organisation database is used to find links to other companies (Figure 1, Label (3)). In the case of Total SA this yields another 20 companies including: Total Petrochemicals & Refining SA[[3]](#footnote-3), Total Olefins Antwerp NV and Total Lesotho Pty Ltd.

Secondly, information from **MNE webpages** is used from an open source ‘copy of the internet’ generated via web crawling from the Common Crawl[[4]](#footnote-4). This process develops a graph of the links between companies, from which cases are identified where there exists a link both outwards from the root website, backwards from the child website, and the website excluding the top level domain is identical (Figure 1, Label (4)). In the case of total.com, this relationship is satisfied for 61 websites including total.fr, total.co.uk and total.co.in.

Websites can also have an associated security certificate (SSL) in order to verify that identity of the company operating the website and that the data communicated between parties using that website is secure. The use of these certificates has increased rapidly since the announcement that SSL security is a determining factor for rankings on search engines. The **SSL certification data** is sourced from Rapid7’s Open Data Sets[[5]](#footnote-5). Each SSL certificate can contain information for Legal Name of Company, Jurisdiction of Company, Business Register Identifier and Other websites operated. Table 2 details the linking information available for total.com from one issued certificate.

**Table 2. Data assigned to total.com from SSL certification**

|  |  |  |
| --- | --- | --- |
| **Identifier Type** | **Value** | **Reference in Figure 1** |
| Website | total.com | (5) |
| Jurisdiction of Company | France | (6) |
| Legal Name | TOTAL SA |
| Business Register Number | 542051180 | (7) |

Source: [Censys](https://censys.io/certificates/22350e07bbabf28bf0fc91a40d061f0b63dd342e8b905f505121d5d64c192a0c).

The originally discovered PermID for Total SA also has a LEI associated with it. The Legal Entity Identifier (LEI) is a 20-character reference code to 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’). Level 1 information on Business Register Information is used to provide a link between the LEI and information from other sources such as SSL certificates. In the case of Total SA the Business Register Information is once again confirmed as 542051180[[6]](#footnote-6) (Figure 1, Label (8)). Whereas Level 2 information is used to provide a link between other LEIs (Figure 1, Label (9)), in all there exist 91 entities that declare Total SA to be their Ultimate Parent Entity.

Further information is obtained by manually collecting the affiliate names for each MNE from Annual Reports (Figure 1, Label (10)). For Total SA there are 836 affiliates listed in the annual company reports, alongside the company names from SSL certificates (Figure 1, Label (6)). These legal names (with legal jurisdiction used to filter and improve matching opportunities) are applied to a fuzzy matching routine to determine the most likely PermID within the Organisation PermID database corresponds with this name. The linking of Legal Name to an underlying PermID uses the PermID Record Matching service[[7]](#footnote-7). This service has an API allowing for automated download scripts to be created.

Through the exploration of these wide and different data sources, a number of identifiers have been identified for Total SA. These include 836 unique companies in Annual Reports, 533 PermIDs, 187 websites, 40 different SSL certificate names, 112 Business Register Serial Numbers and 117 LEI’s. The right hand side of Figure 1 depicts each one of these 1825 identifiers and how they relate to one another, the positioning of this information shows that a large quantity of information exists around the centre component (reflecting that Total SA reports the majority of its affiliates in Annual Reporting). In cases where companies disclose as many affiliates as in Annual Reporting, we would expect to see a greater level of clustering in the diagram.

### **Economic indicators**

Once all identifiers are determined, the geography associated with the node can be determined (either directly from the data source, or indirectly such as total.be being attributed to Belgium through a geographically specified webdomain). The importance of a website can then be used to build further breakdowns of estimates for geographic exposures. The first suite of OECD ADIMA Indicators therefore focuses on the Geography or Country-level dimension derived from the knowledge graph. For example, for each country, a summary of which MNEs identified affiliates in that country, either through annual reports or through the LEI registration and when MNEs demonstrate presence in that country through the operations of websites attributed to that geography, an indicator of digital presence is derived. This allows indicators on MNE presence, split by physical and digital presence, to be calculated by geography. For example, 113 countries identified as having Total SA presence in the knowledge graph, 91 of which reported physical affiliates. There were 36 OECD countries which showed a presence for Total SA, 31 of which reported physical affiliates.

The second suite of OECD ADIMA Indicators describes the economic structure and performance of MNEs via a selected set of variables reported in Balance Sheets, Income Statements and Cash Flow statements. All these indicators are collected at the C**onsolidated MNE level** and are captured with traceability to the original financial statements of the Annual Reports of the respective MNEs.

### **Monitoring tool**

As the register and geography and country-level indicators described above reflects sources that are updated on a real time basis, the ADIMA database can help describe major corporate events, and changes identified in *the graph of information* will provide a key input into the ADIMA monitor. While Wikipedia and GDELT news services are being tested to highlight MNE peak mentions in media as part of the ADIMA monitor, it is important to note that these ‘mentions’ are often speculative in nature, so structural changes may not occur. Moreover if they do occur, it may not be until some months after the initial speculative peak. The knowledge graph on the other hand is designed to capture and highlight active and actual examples of structural changes in the firm, for example the changing of a security certificate for a website which has transferred ownership. Collaboration with Eurostat has started with respect to the Monitoring tool, to ensure alignment and complementarity.

In addition the tool provides significant scope to improve the quality of current AMNE/FATS statistics. By design, AMNE/FATS statistics only capture the view of activity of a given affiliate and not the MNE global view. In practice therefore, for affiliates where the ultimate owner may not be obvious, it is possible that the same affiliate may be recorded in outward FATS statistics of more than one country, creating double counting at a global level, and of course asymmetries in FATS flows. One example is an affiliate in the United States that is consolidated under both Total SA (France) and BASF (Germany), meaning that it is possibly captured by France’s outward FATS and Germany’s outward FATS. OECD ADIMA includes a concept called ‘node overlap’ that captures these instances and that can be made available for review by the affected national compilers, who in turn can work to reduce asymmetries. A confrontation with initial ADIMA results (March 2017) for a selection of 37 US MNEs with official US Outward AMNE/FATS showcased considerable alignment; however, it also highlighted that for several countries, important differences were visible, notably for countries often used for fiscal optimisation purposes, like the Netherlands, Ireland and Singapore. In these countries, official FATS data record (much) higher sales than in ADIMA. These results will be explored further with the universe of 100 MNEs.

# Conclusions

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.

The information contained within the knowledge graph is a contribution to support the consistent treatment of MNEs in national statistics, to analysts wishing to profile MNEs both individually and on aggregate. The graph can provide lists of known entity names, allowing use in fuzzy matching with an administrative data set, for example, such as customs (merchandise trade). Furthermore it can provide indicators of digital presence, detecting national activity for those MNEs in information industries and that function on a web platform as opposed to the more ‘traditional’ physical affiliate. These examples of ADIMA applications are merely scratching the surface of what is possible. **A selection of these Indicators will be presented at the NTTS conference, including those published in late 2018 as well as several in-development indicators which showcase new directions for ADIMA.**

|  |
| --- |
| **Annex I. 100 MNEs, OECD ADIMA** |
|   |  | Wal-Mart Stores IncToyota Motor CorpRoyal Dutch Shell PLCVolkswagen AGExxon Mobil CorpApple IncMcKesson CorpUnitedHealth Group IncBP PLCCVS Health CorpSamsung Electronics Co LtdGeneral Motors CoAT&T IncDaimler AGGlencore PLCFord Motor CoAmerisourceBergen CorpAmazon.com IncHon Hai Precision Industry Co LtdCardinal Health IncTotal SAVerizon Communications IncHonda Motor Co LtdGeneral Electric CoCostco Wholesale Corp | Walgreens Boots Alliance IncFiat Chrysler Automobiles NVKroger CoChevron CorpNissan Motor Co LtdNippon Telegraph and Telephone CorpExpress Scripts Holding CoGazprom PAOBayerische Motoren Werke AGLegal & General Group PLCHome Depot IncBoeing CoAlphabet IncMicrosoft CorpSiemens AGPrudential PLCNestle SANK Lukoil PAOAnthem IncPhillips 66Carrefour SAHitachi LtdComcast CorpInternational Business Machines CorpSoftBank Group Corp | Hyundai Motor CoDeutsche Telekom AGValero Energy CorpDowDuPont IncAeon Co Ltd Enel SpAJohnson & JohnsonUniper SEEngie SAAirbus SETesco PLCTarget CorpSony CorpPanasonic CorpCeconomy AGProcter & Gamble CoLowe's Companies IncMarubeni CorpMarathon Petroleum CorpAetna IncJXTG Holdings IncPepsiCo IncArcher Daniels Midland CoDell Technologies IncUnited Parcel Service Inc | BASF SEFedEx CorpDeutsche Post AGIntel CorpMitsubishi CorpUnited Technologies CorpPeugeot SAArcelorMittal SAWalt Disney CoUnilever NVSysco CorpTelefonica SAHumana IncRenault SAPfizer IncWesfarmers LtdJbs SAKoninklijke Ahold Delhaize NVReliance Industries LtdVodafone Group PLCHewlett Packard Enterprise CoRoche Holding AGNovartis AGBayer AGHP Inc |

1. BEPS Country by Country (CbC) data is published in the form of aggregated and anonymised tabulations that will ensure that the confidentiality of taxpayer information is preserved, while providing governments with a more complete view of the global activity of the largest MNEs. These statistical analyses would not disclose individual taxpayer specific information. The underlying CbC microdata is restricted within national governments under strict confidentiality rules. If provisions are made for data access collaboration between taxation and statistical offices in each country, this provides an opportunity for statistical analyses that can significantly improve the data on multinational enterprises available for the future analysis of BEPS. [↑](#footnote-ref-1)
2. OECD Smart Data Framework, ["Which strategies for NSOs in the Digital Era? Towards ‘Smart Data’Strategies"](http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=SDD/CSSP(2018)7&docLanguage=En), *15th meeting of the Committee on Statistics and Statistical Policy (CSSP),* 20-21 June 2018. [↑](#footnote-ref-2)
3. PermIDs: <https://permid.org/1-4295859409>, <https://permid.org/1-4296643925>, <https://permid.org/1-4296674086> [↑](#footnote-ref-3)
4. The Common Crawl can be accessed at www.commoncrawl.org. [↑](#footnote-ref-4)
5. Sonar can be accessed at https://opendata.rapid7.com/sonar.ssl/. [↑](#footnote-ref-5)
6. https://www.gleif.org/lei/529900S21EQ1BO4ESM68 [↑](#footnote-ref-6)
7. https://permid.org/match [↑](#footnote-ref-7)