Statistical Journal of the IAOS 36 (2020) 767–774
DOI 10.3233/SJI-200640
IOS Press

A Statistical Business Register spine as a new approach to support data integration and firm-level data linking: An ABS perspective

Luisa Ryan[a,*], Chris Thompson[a] and John Jones[b]
[a]*Business Register Unit, Australian Bureau of Statistics, Australia*
[b]*BLADE, Australian Bureau of Statistics, Australia*

Abstract. Statistical Business Registers (SBR) have historically underpinned the compilation of economic statistics by providing consistent unit structures and classifications for survey frame production and business demography data. To meet emerging data needs for both regular statistical production releases and for specific questions asked by policy makers, the SBR can also be used as a data integrating framework. This paper outlines the "spine" approach proposed by the Australian Bureau of Statistics (ABS) to support more flexible integration and linking of firm-level data that will also expand the uses of the SBR. The spine is the minimum set of information required to identify an entity and act as the linking variable(s) to other datasets. Its application involves a new approach to management of input datasets and can be applied across statistical registers.

This paper will provide (1) a description of the ABS spine proposal for statistical registers; (2) benefits of a spine approach for both regular statistical production and new data solutions; and (3) an overview of how the ABS BLADE (Business Longitudinal Analysis Data Environment) is used to integrate firm-level datasets to enable policy evaluation and statistical research by analysts from government and academia.

Keywords: Statistical Business Register, spine, firm level data, BLADE

1. Introduction

Statistical Business Registers (SBR) are foundational statistical infrastructure for compiling high quality economic statistics as they provide economy wide coverage of economic units, using consistent unit structures and classifications. This enables various economic survey frames that are inputs to the National Accounts and business demography to be consistently compiled on the same basis. The traditionally structured SBRs have served this purpose well. Measurement challenges however, have been increasing for national statistical offices (NSOs) as new economic arrangements emerge (e.g. geospatial views, globalisation, digitisation, sharing economy) and these new phenomena need to be captured in the National Accounts. Policy makers are also seeking more evidence based answers to a wide range of policy questions, including balancing short term macroeconomic outcomes (e.g. fiscal and monetary policy) versus longer term sustainability or regional/local impacts. This is against a backdrop of pressures for NSOs to also reduce operating costs and provider burden.

The solutions to modern statistical measurement challenges need to be multi-faceted. This paper outlines a "spine" concept as a SBR data integrating framework to support achievement of both regular statistical production outcomes and tailored data solutions in a more flexible and responsive manner. This includes the potential to leverage a range of source data, including a broader range of administrative data. Section 2 of this paper describes the ABS spine proposal for statistical registers. Section 3 outlines the benefits of a spine approach, and Section 4 provides an overview of how the

[*]Corresponding author: Luisa Ryan, Business Register Unit, Australian Bureau of Statistics, GPO Box 2796Y, Melbourne, VICTORIA, 3001, Australia. Tel.: +61 3 9615 7531; E-mail: Luisa.ryan@abs.gov.au.

ABS Business Longitudinal Analysis Data Environment (BLADE) is used to integrate firm-level datasets to enable policy evaluation and statistical research by analysts from government and academia.

2. Statistical Business Register spine

2.1. Statistical Business Registers current paradigm

The UN/UNECE *Guidelines on Statistical Business Registers (2020 and 2015)* define a SBR as a "regularly updated, structured database of economic units in a territorial area, maintained by an NSI (National Statistical Institute), and used for statistical purposes" [1,2]. This approach is based on all SBR information being stored and integrated within a database/datastore (noting some countries may have multiple SBR databases), and it has worked well to support delivery of standard and stable economic survey frames and business demography programs. Essentially under this approach the number of variables in the SBR grows over time.

However, both local and global data requirements to meet emerging problems are becoming more complex. There is a growing need to be able to link micro or unit level time series data across economic, social and environmental domains, and to analyse the impact of policies within and across countries for use in both GDP and other measures. The strong value in being able to do this linkage work quickly to assist governments' to understand the impact of and rapidly target responses to national crises has been highlighted by the current COVID-19 pandemic.

The current SBR paradigm however, is constrained as the SBR datastore/data model where data integration occurs internally can limit the production of new views of the economy or use of new data sources due to size, structure, or investment timeframes required.

Australia commenced using its first SBR in the early 1970s with the 1968 Economic Census as the key input. Since that time the ABS Business Register has continued to evolve to:

- leverage the availability of taxation data as the key maintenance source
- include profiling of the largest and most complex businesses
- introduce a Common Frame as the key quarterly snapshot from which survey frames are derived, and
- support publication of business demography data.

However, to meet emerging data challenges the ABS Business Register needs to further evolve to be able to capture more complex economic transactions as well as support more timely production of alternate views at both the macro and micro levels.

To move to a new paradigm the ABS has developed a spine proposal for statistical registers. This will enable a more flexible approach that will better support current statistical production, as well as provide infrastructure to efficiently create a broader set of economic indicators within the NSO data integration framework.

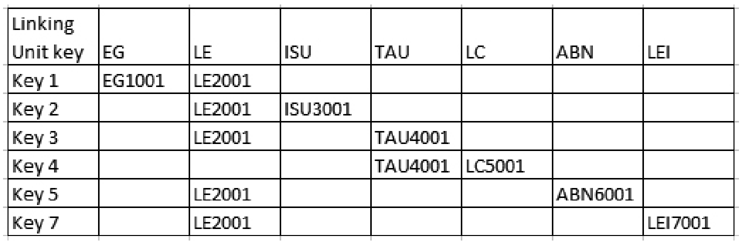
Other NSOs have also begun to adopt or are considering adopting similar approaches to the ABS spine proposal in their statistical production. The United Kingdom Office of National Statistics (ONS) outlined a business spine approach to produce short term indicators in 2017 [3]. It is also noted that an increasing number of countries are now extending the use of the SBR to link micro data and to produce a longitudinal view. This may be achieved using various techniques such as multiple input tables, the live register, snapshots and journal tables. The UNECE *Guidelines on the Use of Statistical Business Registers for Business Demography and Entrepreneurship Statistics* [4] also provides a range of micro data linking case studies. The approach outlined in this paper adds to these options and uses cases within the broader data integration framework.

2.2. Spine concept: A new Statistical Business Register paradigm

The spine concept is a data model to support micro or unit level data linking. The most basic definition of a "spine" is "the minimum set of information required to link two or more datasets". In practice this could include unique identifiers (ideally), or other information such as name and address that can form a matching key or business rule for the linking of one dataset to another. The spine concept maintains a separation between data inputs, the spine and data outputs. The key implication of this approach is that it enables data integration to take place outside of a traditional SBR database.

In Australia, the SBR spine has been developed as a Linkage Table (LT) and the basic representation is provided in Fig. 1. In order to meet the statistical needs of the ABS, the LT includes, but is not limited to:

- relevant Australian Government legal entity identifiers (including the Australian Business Number (ABN)) and international identifiers (including the Legal Entity Identifier (LEI)).



Linking Unit key	EG	LE	ISU	TAU	LC	ABN	LEI
Key 1	EG1001	LE2001					
Key 2		LE2001	ISU3001				
Key 3		LE2001		TAU4001			
Key 4				TAU4001	LC5001		
Key 5		LE2001				ABN6001	
Key 7		LE2001					LEI7001

Fig. 1. Linkage Table illustrative example in the ABS Business Register (for a single Enterprise Group). Note: EG – Enterprise Group, LE – Legal Entity, ISU – Institutional Sector Unit, TAU – Type of Activity Unit (equivalent to the Kind of Activity Unit), LC – Local Unit, ABN – Australian Business Number (Australian government business identifier), LEI – Legal Entity Identifier (international identifier).

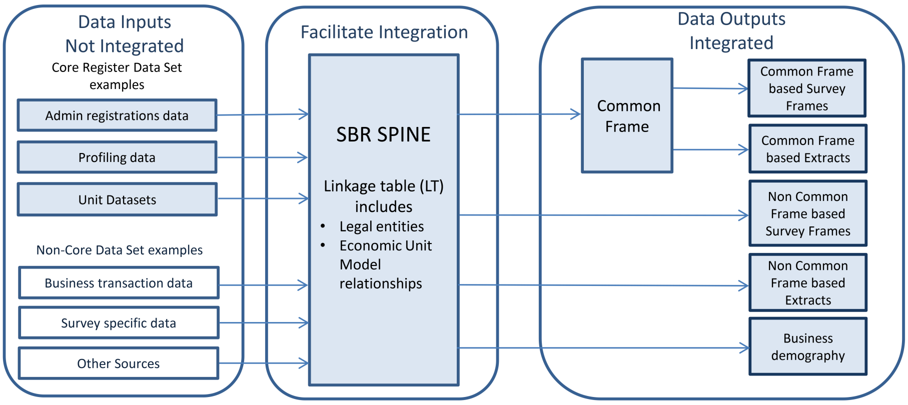

Data Inputs
Not Integrated
Core Register Data Set examples
Admin registrations data
Profiling data
Unit Datasets
Non-Core Data Set examples
Business transaction data
Survey specific data
Other Sources
Facilitate Integration
SBR SPINE
Linkage table (LT) includes
• Legal entities
• Economic Unit Model relationships
Data Outputs
Integrated
Common Frame
Common Frame based Survey Frames
Common Frame based Extracts
Non Common Frame based Survey Frames
Non Common Frame based Extracts
Business demography

Fig. 2. ABS business register as an integrating spine.

– relationships between identifiers and units within the ABS Economic Units Model (includes Enterprise Group, Legal Entity, Type of Activity Unit, Institutional Sector Unit and Location unit).

Under this approach the SBR becomes a register environment rather than a stand-alone register. The register environment includes core register input datasets, the LT (as the spine) and the business rules and views that use this spine to transform data inputs into data outputs. Core register datasets would include administrative and profiling data incorporating standard register data items such as name, address, locations (geocodes), alive status, industry, sector, and size. Non-core datasets would include secondary datasets such as globalisation data, firm age or specific subject matter flags or variables.

Figure 2 presents the view of the proposed ABS Business Register spine model. It highlights that:

1. Input data sets are not integrated.
2. The spine facilitates data integration.
3. Integrated outputs are created using data inputs and the spine.

2.2.1. *Identifiers/linkage variables*

The spine model is a relational data model. If all input datasets in the model were to use the same unique identifier (also sometimes referred to as a primary key), the value in maintaining a spine may be minimal because we can simply merge datasets using the common key to create the desired outputs. Where the spine data model can add value is when different identifiers provide a link between various unit types or data structures. For example, in the SBR context, it is common to have hierarchical business structures. To date, a common approach has been to store the relationships within a registers database or in one (or more) input datasets. Under the spine model, input datasets retain the primary key to the dataset while relationships between units are stored on the LT which forms the proposed ABS Business Register spine.

Where an identifier may have multiple relationships (for example, a business location will relate to both a business and also an address), separate entries should be recorded on the LT. If one of the relationships ceases or changes, you need only update a single relationship (and retire the previous relationship), which provides for a simpler data model, especially for longitudinal analysis where the start and end dates for these relationships are critical.

2.2.2. *Data inputs*

The increase in availability and use of administrative data to maintain traditional registers has been a driver of efficiencies for SBRs internationally, but has also created some challenges. With multiple sources of input data (sometimes conflicting, and of varying quality) and a monolithic register, decisions around the management of input data are made prior to updating register records. Under the proposed ABS Business Register spine data model approach, input data is maintained in a natural state (or with minimal cleansing) and are not used to update a register database or combined with other input datasets. When a new iteration of the source data becomes available, it can either be used to replace the previous iteration or used to create a change (delta) file depending on the business objectives and purpose of the data. New data sources are also able to be used more quickly and easily so long as they have appropriate linkage information and metadata.

Not all data used to update the SBR comes from administrative sources. In Australia, large and complex groups account for less than 1% of administrative units but more than 50% of turnover. Given the importance of these groups to the Australian economy it is important that they are appropriately structured on the ABS Business Register. The structural data is collected through profiling activity which aims to analyse the legal, operational and accounting structures of an Enterprise Group in order to establish the appropriate statistical units to support the collection of economic data. In Australia, the collection of profiling data is predominantly via either survey form or interview, with some administrative data also used to support investigations. The spine principles remain the same regardless of how the data is collected. The profiling dataset(s) in the proposed ABS Business Register spine model will be continuously updated as critical input datasets to the spine model. The profiling structural relationships (e.g. Legal Entity, Type of Activity Unit, Enterprise Group) will be stored in the LT and the profiling characteristics (e.g. address information, industry, employment etc.) will be stored as a core input dataset that links to the spine via the linkage key. Updating of profiling structures and characteristics will be via an interface with the backend linkages seamless to the profiler.

2.2.3. *Data integration*

The spine data model as described thus far has a set (or federation) of input datasets that are not integrated, and these coalesce around a LT that forms the spine. The model resembles a typical hub and spoke design. Data integration is facilitated through the development of a set of business rules that govern the interaction between the input data and the LT in order to produce data outputs. These business rules are the algorithms that are used to transform the input data sets through the spine. They include both rules hierarchies to prioritise the use of input data sets and derivations to create or amend data items. As more input data sources are identified and output requirements expand, the catalogue of business rules also expand. The management and maintenance of business rules becomes a key activity under the proposed ABS Business Register spine model, rather than the management and maintenance of data that is the predominant and costly function of the traditional SBR data model.

2.2.4. *Data outputs*

The spine model enables the production of traditional data outputs. By linking input data through the spine, survey frames are created as they would be under the traditional register model. However, the benefit of the spine approach is that we can readily introduce new input datasets to enhance existing outputs or produce new integrated outputs quickly and efficiently without having to modify and/or redefine the register. This increases the capacity to produce customised solutions. When creating outputs, we can enact business rules that specify the provenance (source) of the input data, which is important for quality assurance purposes. By applying version control to the input data sources, longitudinal outputs are also supported by the spine model.

2.2.5. *Application to other statistical registers and linkages between spines*

Where business structures are hierarchical, and administrative data sources can be linked through multiple identifiers, the spine model has broad utility for most statistical business registers. Furthermore, given that some businesses operate across international jurisdictions using multiple identifiers, there is the potential to link SBRs globally through the implementation of a common spine model. The ABS proposed spine model was presented to the United Nations Committee of Experts on Business and Trade Statistics (UNCEBTS), and it was noted as having potential to “facilitate horizontal (across countries) and vertical (national to global) integrations of SBRs” [5]. Following on from this, the strategic view on SBRs prepared by the UNCEBTS for the United Nations Statistical Commission in 2020 acknowledged that “innovative approaches can be developed to fully exploit SBRs to enhance data integration, using a spine model consisting of a core set of business characteristics" [6]. The UNCEBTS has suggested that the integration of business registers according to the spine model could be explored to support the Global Groups Register of the largest multi-national enterprises (MNEs) presently being developed by the UN Statistical Division, building on the experience of the EuroGroups Register operated by Eurostat [6].

The spine model is equally applicable for use by other types of statistical registers. To date, the ABS has developed a "person" spine and an "Address Register" spine model. It is envisaged that this approach could be extended to other register classes with potential candidates including land, assets or products.

An extension of the spine model is the linkage that can occur between spines. For example, a person spine could be linked to a business spine (e.g. through employment and directorship relationships). The ABS is developing a "business location" spine to link the ABS Business Register spine and Address Register spine. It is likely that as the number of spines grow, the linkages across spines will also expand, providing additional linkage infrastructure across data domains.

2.3. Considerations

In pursuing a spine approach there are a number of considerations that need to be addressed, including the nature of the identifier(s) to be used to enable linking, technology, data custodianship and confidentiality.

2.3.1. Identifiers

Whilst a common identifier across input datasets is ideal for data linking, in practice this is often unobtainable. The LT can cater for all types of relationships between various unit types, and is readily adaptable and extensible. In Australia there is no persistent unique identifier for persons in the way there is for businesses. This means that person data linkage is done with deterministic linkage methods using combinations of identifying variables such as anonymised name (to prevent the name from being identified), geocoded address, date of birth, and sex or gender. The ABS "person" linkage spine is created from a three-way linkage between Medicare Consumer Directory, Social Security and Related Information, and Personal Income Tax datasets, and the spine table contains the identifier concordances that result from this linkage.

Descriptive data, including metadata, is also required to help describe spine relationships. The coverage of the spine would ideally be the identifiers of all units in scope of the population.

2.3.2. Technology

The UN/UNECE *Guidelines on Statistical Business Registers* [1,2] note that the most common approach and also the recommended approach to SBR data management by NSOs is a relational database management system, though other systems such as a key-value store, hierarchical database system, flat files or spreadsheets may also be used. In contrast, the spine model is a data management framework. As such, it is largely technology agnostic. Standard tools for data linking/merging, data access controls, storage capability (preferably an enterprise-wide data repository to enable broader use of data outputs) and the ability to create and apply business rules to data are essential requirements. As the model may have a number of input datasets of various size and structure, expertise may be required to set up a data structure for optimal performance. The way business rules are compiled may also impact on performance.

2.3.3. Data custodianship

Under the proposed spine model, input data sets may be core to the SBR or non-core. Register staff would maintain custodianship of core datasets, particularly if they involve register derivations. However, one of the benefits of the spine model is that non-core data custodians maintain control over their data. Rather than ingesting data into a register database where control and governance over data is conferred to register staff, under the spine model, data is stored and managed (including controlling who can access the data) by the data custodians. The rules and obligations put in place by data custodians no longer need to be replicated in the register environment, as the access to these (non-core) datasets is managed by the data custodian outside the register's environment.

2.3.4. Confidentiality

By enabling data custodians to maintain control over who has access to their data and for what purposes, confidentiality of input data can be securely maintained. Access to the SBR LT can be more liberal because the relationships between units are meaningless without access to the constituent dataset. Appropriate confidentiality rules for aggregate and unit record data would need to continue to be applied to any data outputs from the application of the spine model.

2.3.5. Time taken to implement

A number of factors will determine the length of time it takes to implement or transition to the spine model approach, such as the complexity of the current data model and the need to maintain business continuity. In the ABS context, the implementation of the spine model is a project expected to take between 3 and 5 years. This timing will be refined after early trial results.

3. Benefits of a spine approach

3.1. Statistical production benefits

Regular statistical production includes outputs such as the National Accounts, industry or thematic outputs, and prices statistics. These usually have a heavy reliance on survey or administrative data that is acquired and processed for a specific purpose. When new economic phenomena emerge including in response to crises (e.g. the COVID-19 pandemic), these need to be captured in regular statistical outputs. Often historical processes such as classifications can be slow to adapt, and the costs of collecting new information can be prohibitive. The implementation of the spine model has the potential to benefit regular statistical production in the following ways:

- Survey frame production – the spine conceptual framework supports a neat data linking solution to produce standard business register products such as survey frames.
- Data replacement – where administrative data are available these can be applied via the spine as an alternate data source to direct collect, including at the data item level. Such micro data linking will enable a reduction in provider burden as both core or non-core register datasets can be used for data substitution. For instance, employment data sourced from administrative sources could replace the collection of employment in economic surveys. Direct collect data could also be re-purposed through application of the spine model for use in other products.
- New views – the spine model enables linking of an expanded range of administrative data sources (subject to appropriate linking variables). This means that new or alternate views using register information can be created, and it will also expand analytical opportunities. For instance, for MNEs more detailed views of foreign ownership and/or activity could be created by linking globalisation data via the SBR spine. A broader range of data could also be presented at regional levels (e.g. sub-state levels), using the spine model to support linking.
- Alternate industry views or indicators – satellite or alternate views of the economy at a more micro level (that still link back to core measures) could be created through the use of alternate industry views where this is appropriate. For instance, if sharing economy units could be flagged via the spine, the economic contribution of these could be calculated. Alternate views could be published or used as an analytical tool to better understand emerging activities and industries in an increasingly global economy or in response to national and international crises.
- Improved coherence between the production and sector view of the economy – the base unit (legal entity) in the spine will be used by the ABS Business Register to build both production and sector views enabling a more coherent approach when delineating Enterprise and Kind of Activity Units.
- International linkages – the spine could be used to integrate international identifiers (e.g. the *Legal Entity Identifier (LEI)*), thereby potentially enabling the linking of international datasets (subject to appropriate confidentialisation). This would assist in improving understanding of MNE activity and associated production and capital flows.

3.2. Data solution benefits

Data solutions are tailored outputs or studies aimed at answering specific questions, including for policy development. Data solutions often have a longitudinal element and where the SBR are used are focused on firm behavior or characteristics. The implementation of the spine model has the potential to benefit data solutions in the following ways:

- Timely and responsive – the spine model provides a rigorous approach to combining different datasets across time and geography to answer a specific question in a timely manner. Often the limiting factor is the lack of availability of an appropriate linking variable. Examples could include analysis of the impact of and response by businesses to an emergency event.
- More detailed analysis – the spine can also be used to conduct more in depth analysis. This might involve linking multiple registers or datasets, and potentially across different domains (e.g. economic, environmental and social), to understand the longitudinal impact of a specific policy. An example might be investigating the characteristics of busi-

nesses (e.g. by size) that use particular government services and the performance outcomes of these businesses.

- The COVID-19 pandemic provides a good case study, where the business and location spines could be used to support analysis via the linking of various datasets on business activity, taxation, employment, and use of government support packages. Big data sources such as electricity or mobile phone usage could also potentially be linked (subject to appropriate confidentialisation) to draw new insights into changing behavioral patterns.

The next section discusses how the ABS BLADE has applied a longitudinal spine to support delivery of data solutions.

4. ABS BLADE application

4.1. Background

The Business Longitudinal Analysis Data Environment (BLADE) utilizes information from the ABS Business Register to create a longitudinal version of the business register spine. This spine facilitates the combination of a time series of Australian Taxation Office (ATO), ABS Survey and other administrative data to provide a better understanding of Australian businesses and the economy. Authorized researchers working on approved projects can use BLADE data to study how businesses fare over time and the factors that drive performance, innovation, job creation, competitiveness, trade, and productivity.

4.2. BLADE data asset – what is it and what does it do?

The ABS Business Register spine forms an integral part of the BLADE. Various point in time snapshots of the ABS Business Register are taken and treated to retrospectively create a longitudinal spine. This longitudinal spine is then used to integrate ATO Business Activity Statement (BAS), Business Income Tax (BIT) and business level Pay As You Go (PAYG) employment payment summaries. The spine is also used to integrate a variety of ABS survey data, as well as Intellectual Property data obtained from the IP Australia government agency. All of these data are treated so that they align with the ABS Economic Units Model.

Not all data integrated in BLADE is easily reconcilable with the ABS Economic Units Model. For example, some data in the BLADE asset, such as Merchandise Imports and Exports, is available at the transaction level. Similarly, other administrative data integrated for specific research projects is often provided at an ABN level. A method has been developed that reconciles ABN and industry information from both the business spine and ABN/transaction level datasets to facilitate integration. This integration produces a concordance file that allows the ABN/transaction on the administrative dataset to be matched with the statistical units in the ABS Economic Units Model in the BLADE. Integrated BLADE data is accessed by authorised researchers in a secure technical environment, with outputs being subject to a confidentiality assessment.

4.3. Case study: Innovation business performance

BLADE microdata has been used to inform a wide variety of analysis including: impact of government services and funding; climate; trade; jobs growth; productivity and trade. A summary of some of the research is accessible on the ABS website [7].

"The impact of persistent innovation on business growth" published by Hendricksen et al. [8] is an example of the type of research BLADE can facilitate. The use of BLADE microdata allowed Hendricksen et al. to control for various other, potentially confounding business characteristics. The researchers used Propensity Score Matching (PSM) to establish a positive relationship between innovating businesses on growth of sales, value added, employment, profit and other performance indices. The effect was particularly strong in small and medium sized businesses. This conclusion led to the authors advocating for targeted policy focusing on small and medium businesses.

5. Conclusion

This paper has outlined a business spine proposal to support a more flexible approach to the integration and linking of firm-level data within a registers environment, in order to better support both regular statistical production and delivery of new data solutions. ABS have developed a future roadmap to implement a business spine into the ABS Business Register environment, and implementation is planned by taking a modular approach over subsequent years subject to available resourcing. The overview of the ABS BLADE in this paper provides an insight into how the longitudinal application of the spine concept can be operationalized using the SBR as the foundation.

Acknowledgments

The authors would like to acknowledge the contribution of Willem Erasmus and Michael Smedes in the development of the ABS business spine concept, and Carrie Samuels, AJ Lanyon and Matthew Mullaly for their feedback on the paper.

References

[1] UN, *Guidelines of Statistical Business Registers*, 2020.
[2] UNECE, *Guidelines on Statistical Business Registers*, 2015.
[3] Office for National Statistics (United Kingdom), *Transforming Short-Term Turnover Statistics: October 2017*, 2017.
[4] UNECE, *Guidelines on the Use of Statistical Business Registers for Business Demography and Entrepreneurship Statistics*, 2018.
[5] UN, *Summary Report of the 2nd Meeting of the UN Committee on Business and Trade Statistics*, 2019.
[6] Bureau of UN Committee of Experts on Business and Trade Statistics, *Strategic View on Business Statistics*, presented to the United Nations Statistical Commission meeting in 2020.
[7] ABS, *Blade Research Projects*, 2018: https://www.abs.gov.au/websitedbs/D3310114.nsf/home/Statistical+Data+Integration+-+BLADE+Research+Projects.
[8] Hendricksen, Taylor, Ang, Cao, Nguyen & Soriano, *The impact of persistent innovation on business growth*, Office of the Chief Economist: Department of Industry, Innovation and Science, 2018: https://www.industry.gov.au/sites/default/files/2018-12/oce-impact-of-persistent-innovation-on-business-growth.pdf.