

# Incorporating micro data into macro policy decision-making

Jean-Marc Israël and Bruno Tissot<sup>1</sup>

## Executive summary<sup>2</sup>

In 2019, the Irving Fisher Committee on Central Bank Statistics (IFC) of the Bank for International Settlements (BIS), together with the European Central Bank (ECB) and the Central Bank of Malaysia, decided to analyse the **issues posed by the integration of micro data sets into the framework of official statistics, for the purposes of policymaking**.

This was an opportunity to review the expansion of micro (granular) data sets made available since the Great Financial Crisis (GFC) of 2007–09 and their possible contribution to macro compilation exercises. The review underlined the **importance of broadening official statistical frameworks to benefit from the growing availability of granular data** to assist the assembling of macroeconomic aggregates and/or facilitate the linkage between micro- and macro-level statistics.

This stocktaking exercise also documented that the **“micro data revolution” can bring important analytical benefits and effectively support central bank policies**, with a greater ability to “zoom in” on particular areas of interest and to assess the distribution of macroeconomic aggregates within reporting populations. One significant example is that granular and micro data already assist in assessing how finance can contribute to the greening of economies.<sup>3</sup> The initiative also proved a welcome opportunity to highlight the potential of new types of granular information in “unusual” circumstances. Indeed, and as observed subsequently with the sudden outbreak of the Covid-19 pandemic in early 2020, micro data sources can be a valuable complement to the “traditional” official macro statistics available in times of stress, especially to alleviate compilation disruptions, assess pressure points, and facilitate the implementation of targeted policy measures.

However, there are **also important challenges associated with dealing with these data sets, for instance as regards their quality, confidentiality and manner of accessing them**. Indeed, the task of integrating (granular and) micro financial information in macro frameworks has proved more complex than initially thought. To ensure concrete progress in the future, attention should focus on:

<sup>1</sup> Respectively, former Head of the Monetary and Financial Statistics Division at the ECB ([jeanmarc.il@free.fr](mailto:jeanmarc.il@free.fr)); and Head of Statistics & Research Support, BIS, and Head of the Secretariat of the Irving Fisher Committee on Central Bank Statistics (IFC) ([Bruno.Tissot@bis.org](mailto:Bruno.Tissot@bis.org)).

<sup>2</sup> The views expressed here are those of the authors and do not necessarily reflect those of the BIS, the Central Bank of Malaysia, the ECB, the IFC or any of those institutions represented at the seminar. We thank Barend de Beer, Neil Humphries, Robert Kirchner, Michael Kock, Lim Sheng Ling, Ong Li Ming, Eric Nielsen, Laurent Olislager, Steffi Schuster and Johannes Turner for helpful comments and suggestions.

<sup>3</sup> See eg the OECD’s [Microdata Analysis for Environmental Finance and Investment \(MEFI\)](#) initiative.

- (i) building effective micro data collection frameworks based on a comprehensive data strategy helping to contain reporting burden;
- (ii) accessing and making use of more granular sources of information, with the need to overcome the challenges related to their size and complexity with a view to transforming simple data points into knowledge;
- (iii) promoting the exchange of experience, eg as regards access to micro data sets and external research projects, the development of diversified staff skills, and the combination of different types of data sets;
- (iv) developing new and adequate analytical tools, for instance to enhance data quality assurance processes, extract summary indicators from a wealth of data points, and develop machine learning (ML) / text mining / network analysis approaches to maximise the potential of granular (including micro-level) information; and
- (v) bridging the gap between micro- and macro-level statistical exercises, which can be instrumental in enhancing the understanding of how the financial system functions and interacts with the economy, assessing distributional issues and facilitating sectoral analyses.

In supporting the above tasks, **there is a growing interest in making use of granular information from private sources that are not part of the official statistical offering.** What is unclear, though, is how data producers located outside the national statistical system should feature vis-à-vis the fundamental principles that govern the production of appropriate and reliable official statistics and adhere to certain professional and scientific standards. Helpfully, a number of private firms have already adopted dedicated and transparent mission statements and principles to address these concerns, and such initiatives may deserve to be strongly encouraged by the official (international) statistical community.

Lastly, there are **specific communication challenges, as well as potential legal difficulties, for policymaking institutions** like central banks when (confidential) granular analytical insights are used as the foundation for their decisions.

## Introduction

**Increasingly vast and complex amounts of granular data are becoming available for public authorities to use.** This is particularly the case for central banks in their role as statistics compilers, a function that has clearly expanded in recent decades, reflecting the growing importance of finance in the economy (see Greenwood and Scharfstein (2013) for the US situation). Another trigger was the launch of various and important micro data collections, especially in the context of the G20-endorsed Data Gaps Initiative (DGI) in response to GFC. In particular, the first phase of the DGI (2009–15) stressed the importance of complementing macro-level aggregates with distributional information (see FSB and IMF (2009), Recommendation no 15). The second phase (2016–21) called for collecting more granular data to “help straddle the divide between micro and macro analysis” (FSB and IMF (2015)). Such an emphasis on the importance of micro data was relatively new in the area of official statistics. For instance, the last version of the Manual for the System of National Accounts (SNA) devotes only a two-page section (out of almost 700 pages) to “The use of microdata for macroeconomic accounting” – stating that “there would be considerable analytical

advantages in having micro databases that are fully compatible with the corresponding macroeconomic accounts” but that this task “may be difficult, if not impossible, to achieve” (2008 SNA; see European Commission et al (2009), pp 9–10).

Today, **the importance of micro and granular information is now fully recognised**, not just by those market supervisors tasked with monitoring individual institutions (“suptech”; see Broeders and Prenio (2018)), but also more broadly by all the public authorities in charge of macro policies in the economic and financial sphere (Box 1). A key reason is that financial stress experienced at the level of individual entities, transactions or instruments can quickly reverberate into the entire financial system and beyond. Macro-level analyses should therefore be complemented with a micro-level, or in cases granular, approach – “that is, we need to see the forest as well as the trees within it” (Borio (2013)). More precisely, the high granularity of certain data sets can help: (i) “zoom in” and get idiosyncratic information that is important from a system-wide perspective; (ii) give a better sense of the distribution of economic aggregates; (iii) assemble detailed information bricks to compile more precise “macro” estimates; (iv) facilitate the design and evaluation of evidence-based policies, especially as regards financial sector reforms; and (v) trigger new, innovative ways of looking at economic and financial phenomena (see Tissot (2018)).

#### Box 1

### Granular and micro data

Micro data and granular information relate to two close but different concepts – an important distinction as the kinds of analyses will differ depending on the type of layer (micro or granular) they are run on.

As regards **micro data**, the traditional approach is to consider that it is “an observation data collected on an individual object - statistical unit” (OECD Glossary of Statistical Terms).<sup>1</sup> In the area of economics and finance, this will comprise “data on individual reporting units or specific transactions/instruments, which in most cases allow the identification of individual entities and are therefore considered confidential. In addition, publicly available data on individual reporting units are considered non-confidential although they can still be subject to data sharing limitations due to commercial property rights” (IAG (2017)).

Financial micro data will, for instance, typically refer to (confidential or not) institution-level information such as balance sheets or P&L accounts which reflect the activity reported by individual players like banks or non-financial corporations – covering, for instance, information from central balance sheet data offices (IFC (2017b)).

**Granular data** can be defined as comprising “disaggregated data and micro data” (IAG (2017)). **Disaggregated data** are “data below the level of aggregated data<sup>2</sup> and with a higher likelihood of identifying individual reporting units than in the aggregated data”. So granular data cover less aggregated data than traditional statistics plus micro data, including when these are anonymised for dissemination.

In the financial sphere, the concept of granularity will, for instance, typically refer to data on individual transactions, eg on loans or securities, that are an input to the production of aggregated information by compiling institutions, and can also be an area for analyses and research on their own.

<sup>1</sup> **Macro data** being “an observation data gained by a purposeful aggregation of statistical micro data conforming to statistical methodology” according to this glossary. <sup>2</sup> With **aggregated data** defined as “data aggregates that have a low likelihood of identification of individual reporting units, such as those found in traditional datasets” (IAG (2017)).

After more than a decade since the GFC, **what have been the results of the various and ambitious initiatives undertaken to collect granular information on the financial system?** Central banks and financial supervisors have been at the forefront of this journey, for instance to ensure greater consistency between new micro level data sets and traditional macro aggregates, adapt statistical frameworks to the rapidly evolving financial system, and exploit micro, firm-level data sets for financial stability work.<sup>4</sup> For instance, many jurisdictions around the globe now have at their disposal very large and granular loan-by-loan databases, representing the bulk of what is considered “big data” by central banks (Schubert (2016); IFC (2017a)). Information on derivatives transactions reported by trade repositories (TRs) is another example of the detailed data sets that are increasingly of interest to central banks (IFC (2018)).

**Yet the task of integrating (granular and) micro financial information in macro frameworks has proved more complex than initially thought.** It remains hindered by the limited availability of reliable and timely reports at the level of individual institutions, especially in the least regulated corners of the financial system. Moreover, the lack of international harmonisation continues to challenge the collection of comparable, entity-by-entity data among corporations, not least across jurisdictions. Furthermore, the underlying financial statistical infrastructure is still incomplete, reflecting the slow development of global identifiers, standards for exchanging information, and data sharing arrangements – despite notable progress since the GFC as regards the Legal Entity Identifier (LEI; see LEIROC (2016)), the Statistical Data and Metadata eXchange standard (SDMX; see IFC (2016a)), and the actual international sharing of granular information on global financial institutions (see Bese Goksu and Tissot (2018)).

**The difficulties above were particularly evident when Covid-19 struck,** for instance during the related turmoil observed in financial markets at the beginning of 2020 (FSB (2020)). The pandemic also highlighted that a wealth of available micro data might not be fully exploited. Yet the response of official statisticians in various countries showed that “quick wins” could be achieved by making better use of the information already collected (Ducharme et al (2020)); for instance, to complement conventional survey statistics that became disrupted, to provide detailed statistics on target populations with richer insights on their distribution, or to shed light on the unexpected consequences of the pandemic (eg assessment of mobility patterns and consumer habits; De Beer and Tissot (2020)).

Moreover, **the growing interest in using granular information from alternative sources has to be balanced with the fundamental principles that govern the production of appropriate and reliable official statistics** and adhere to certain professional and scientific standards (United Nations (2013)). For instance, a key objective should be that the micro data involved are unbiased and produced in a transparent way with adequate metadata information. Another is to ensure that the provision of this kind of information is actually made independent of the data

<sup>4</sup> See IAG (2017) for a description of relevant experience by countries as well as regional and international organisations that illustrates the efforts undertaken in the various related areas covering the use of common statistical identifiers; the exchange of experience on statistical work with granular data and improvements in transparency; the balancing of confidentiality and user needs; the linking of different data sets; the provision of data at the international level; the ways for improving data-sharing of granular data; and the collection of data only once.

providers' business goals. A third aspect is to favour benchmarking exercises, especially against other types of data, to ensure accuracy and appropriate context. A final point is to certify that the integrity and confidentiality of the data is respected and that the information is used honestly.<sup>5</sup>

**The above considerations clearly underscore the importance of improving the integration of micro data into macro frameworks.** A relevant initiative from this perspective was the joint organisation by the Central Bank of Malaysia, the ECB and the IFC of a satellite seminar on "Post-crisis data landscape: micro data for the macro world", held on the occasion of the 62nd biennial World Statistics Congress (WSC) of the International Statistical Institute (ISI) in Malaysia in 2019. Participants from about 50 organisations, including central banks, national statistical institutes/offices, international organisations, the private sector and academia, convened to jointly analyse this topic. The event proved to be another important milestone in the IFC's ongoing work on micro-level information – especially in finding appropriate sources, developing new methodological concepts and techniques, compiling policy-relevant indicators and making use of them, and taking advantage of rapid improvements in technology (the "big data revolution") (IFC (2016a,b,c)). It also provided another opportunity to highlight existing best practices and potential opportunities, especially to support policymaking, as well as to take stock of the challenges to be addressed as a priority.

In particular, and as emphasised by the senior panel invited to conclude the seminar, **there are clear limitations in analysing economies based solely on the supply of macro statistics.** In contrast, a growing and lasting demand for more granular insights has emerged from a wide range of users – for instance to monitor risks in the financial and non-financial sectors, analyse interconnectedness and assess international spillovers.

Addressing this supply/demand information gap puts a premium on finding effective **ways to drill down from aggregates to more micro information.** This can offer a wide range of analytical opportunities, by: (i) reducing the risk of overseeing phenomena that may not be noticed at the macro level; (ii) better supporting the design of macro policies, especially in the areas of monetary and financial stability; and (iii) allowing for more accurate measurement of the impact of authorities' actions. In turn, such monitoring would also enable some fine-tuning where and when appropriate. Moreover, making better use of granular (including micro) data collections should be favourable for reporting agents, by lowering their reporting burden with more stable and resilient requirements and reducing the need for them to compile aggregated data. However, harmonising and streamlining data reporting is essential to reap those benefits, and corresponds in fact to a strong demand expressed by the financial industry.

**Looking ahead, to make concrete progress attention should focus on four main tasks:** building up new micro data collection frameworks (Section 1); accessing and making use of more granular sources of information (Section 2); developing new and adequate analytical tools (Section 3); and bridging the gap between micro- and macro-level statistics (Section 4).

<sup>5</sup> See eg the [Research Mission Statement and related Principles](#) published by the US online real estate company Zillow, whose data are now being used by the Board of Governors of the US Federal Reserve System for official statistical purposes; see Hume McIntosh (2021).

## 1. New frameworks for collecting micro data

**A first, obvious issue for producers of official statistics is how to organise themselves to collect the increasing amounts of granular information which are potentially available.** This is particularly important for central banks and supervisory authorities, which are confronted, on the one hand, with increased reporting fatigue among respondents and, on the other hand, with a steady rise in data needs to conduct evidence-based macro policies and monitor the financial sector.

Various countries' experiences suggest that this trade-off can be best addressed by **integrating reporting requirements and standardising data in generic “cubes”**, ie structures that allow for the storing, reporting and analysing of data points with multiple dimensions, so as to cover the wide range of reporting exercises and analytical exercises. Such an approach can also allow new information needs to be addressed with more agility and less effort as they come in.

According to [Eurofiling](#), a forum initiative started in 2005 to improve collaboration on and awareness of European regulatory reporting among regulators, supervisors and entities from both the public and private sectors, **a key lesson of the qualitative case studies for recent international projects is that there are three main factors necessary for success** (Chart 1). First is the need to have effective collaboration between the multiple parties involved. To ensure high-quality inputs in data collections, central bank statisticians should in particular understand how the various reporters are organised, how they can cooperate, what should be done so that they remain engaged over time, and how data platforms should be set up to support the provision of continuous feedback. The second main issue is the demand for comprehensive data methodologies and models. Since statistical compilation chains involve many stakeholders, it is important that methodological definitions are well documented, compatible and applied consistently for the variety of purposes considered and with sufficient agility (for instance to reconcile the legal and business approaches supporting data collections). A third issue is to develop adequate platforms and standards for data compilation. This calls for the development of comprehensive technical interfaces (for instance between firms' internal business data warehouses and their submissions for official statistics reporting), shared standards (ideally at the international level since the financial system is global) and common identifiers to correctly identify financial agents as well as their transactions and products.

Certainly, **a number of important initiatives are already being developed in several central banks to address these issues.** In terms of methodologies, the Bank of Spain has built up a Credit Registry Dictionary and has worked on using a data point model, based on the XBRL (eXtensible Business Reporting Language) standard, to streamline its data collection. Singapore has developed an initiative to foster a common understanding of supervisory and statistical reporting requirements in the financial industry. Turning to Europe, the ECB has designed a single data dictionary, the Banks' Integrated Reporting Dictionary (BIRD), to foster cooperation with the banking industry in the field of regulatory reporting, alleviate the reporting burden and improve data quality. This has been complemented with the provision of a large set of precise definitions that are operational and can even help beyond regulatory reporting – for instance, to support the compilation of the Analytical Credit Database (AnaCredit)) (ECB (2019)) – and the use of a single multidimensional data model (SMCube) based on standard identifiers such as the LEI and ISIN (International

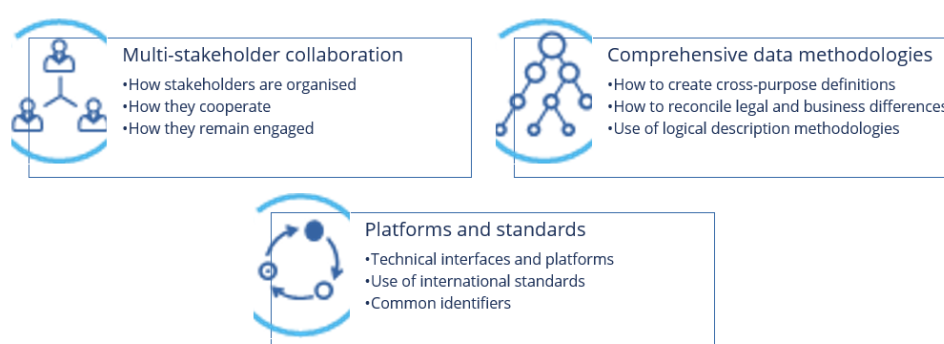
Securities Identification Number), in turn allowing the same micro data to be used for multiple purposes. Similar efforts have been conducted by the European Securities and Markets Authority (ESMA), which is the rule setter for securities and derivatives markets in Europe, for instance to design a reporting framework for derivatives transactions through TRs combining methodological definitions, the description of the related reporting obligations and the use of an ISO standard. On its side, the European Commission has also been working on enhancing financial data standardisation and has also developed a fully fledged “European strategy for data”, especially as regards the financial data space (see European Commission (2020), pp 30–1).

---

### Three main paradigms for modern statistical data frameworks: a view from Eurofiling Foundation

---

Chart 1



Source: M Piechocki, “Paradigms for modern statistical data frameworks and systems”, *IFC Bulletin*, no 53, April 2001.

---

In addition to those strategies aiming at streamlining data reporting processes, **it may be useful to define a comprehensive data strategy for the institution in charge of data collection.** This should encompass the related business cases, the projects envisaged (eg building up a data platform), as well as the related governance issues faced, including data sharing/access and collaboration exercises. The experience of the [Bank of Thailand](#) is that such a strategy can bring many benefits. The reason is that information is a strategic asset and that organisations could make more active use of it. The central bank has therefore developed a multi-year plan with the goal of improving the stocktaking, analysis and sharing of the various economic and financial micro data potentially available in the country – especially through greater use of high-frequency data from various sources including unstructured ones, eg text (Chart 2). The primary focus has been on micro-level statistics, since their granularity can provide various insights for policy use, for instance to analyse monetary policy mechanisms or developments in property markets, as well as to improve the efficiency of the market operations conducted by the central bank. Certainly, this project has faced many challenges, especially regarding staff skills (to be developed through external recruitment or internal training), the selection of technology to handle vast and various types of data sets (eg structured as well as unstructured), the development of adequate legal and data governance frameworks, and the need to comply with data protection rules.



## Taking stock of micro data sources: the experience of the Bank of Thailand

Chart 2



Source: P Chansang, "Data management in the data evolution era at Bank of Thailand", *IFC Bulletin*, no 53, April 2021.

**Turning to the other side of data collection processes, reporting institutions are also making active progress.** This is particularly the case for those banks and other financial institutions that have faced increasing reporting requirements, especially in the wake of the GFC. The resulting burden has been challenging their traditional way of collecting data from internal reports, which were typically organised along separated business processes – with the risk of redundancies, inflexibilities and inconsistencies between different reporting exercises – and were rarely standardised (implying that any new collection exercise would lead to extra reporting costs).

The experience in Austria is that one way to overcome these issues and make data collection exercises more efficient on the reporting side is to better integrate the reporting process (Chart 3). To do so, **the central bank has developed a "regtech approach"** (which basically refers to the provision of methodology, technology and processes to financial institutions to support regulatory monitoring, reporting and compliance). An important factor in this approach was the decision by the largest Austrian banks – in coordination with the central bank – to establish a central reporting platform (AUREP). The goal was to ensure that all the data to be reported by commercial banks are sourced from a general ledger – that is, a general "basic cube" (input layer) with ad hoc validation rules to support data quality management. This allows storing and organising all the information in "smart cubes" made of specific data selected/aggregated/transformed so that these are easily available for external reporting exercises with different types of aggregation level (eg consolidated, sectoral). This setup can also be used effectively to support the banks' internal reporting.

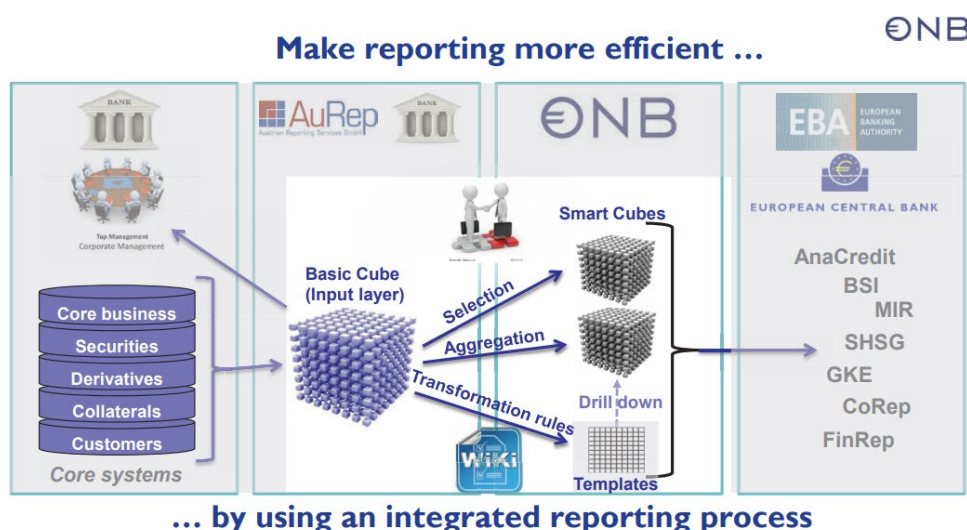
Such an approach can help to improve data quality, achieve synergies and lower reporting costs, and can **represent a "game-changer" with an overall impact on the information organisation framework.** In addition, it facilitates the use of



advanced analytical techniques, as data can be accessed and retrieved more easily. Lastly, one key lesson is that the more granular the reporting requirements, the more effective the approach can be, since the underlying data are not transformed when stored and can thus be re-used for different exercises – minimising the potential impact of additional enquiries by the authorities. Further progress could be obtained if regulators set out their reporting requirements in standardised languages, possibly machine-readable ones. This would allow retrieval (or “pull”) of the data from the reporters’ own internal cubes, instead of asking them to send (or “push”) the data themselves – in turn reducing reporting costs and further increasing data quality and timeliness.

Integrating the data reporting process: the experience of the Central Bank of the Republic of Austria

Chart 3

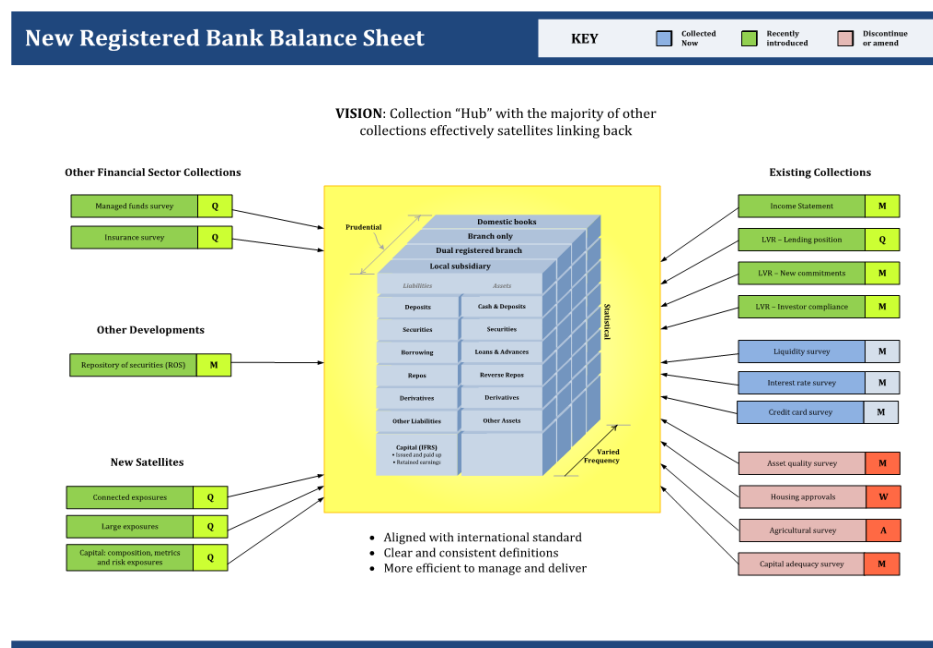


Source: J Turner, “OeNB’s reporting data model as RegTech/SupTech solution”, *IFC Bulletin*, no 53, April 2021.

**The focus on reducing reporting burden is also high on the agenda of the Reserve Bank of New Zealand**, which has developed a “collect once, use multiple times” approach similar to the one adopted in Austria (Chart 4). A main goal has been to enhance the accuracy of the data collected from the financial industry, especially banks, and ensure that all regular data collections are relevant, fit for purpose and cost-effective. In particular, and to avoid misunderstandings in implementing the related instructions, the Reserve Bank ensures that reporting exercises are designed in strong connection with the data providers. Building relationships, providing clarity around data requests and spending time discussing with data providers is not costless but is proving extremely valuable in the long run. Moreover, any new specific data request – covering all prudential, macroprudential and statistical purposes – can be linked back to a single high-level balance sheet or central “hub”. The experience is that this “collect once but use it multiple times” approach can be instrumental in reducing errors in the data sent, addressing new reporting needs with more agility, and facilitating data analytics work.

Single collection for multiple use (prudential, macroprudential and statistical purposes): the experience of the Reserve Bank of New Zealand<sup>1</sup>

Chart 4



<sup>1</sup> As at 2017.

Source: S Schuster and N Humphries, "Collect once, use multiple times – the Reserve Bank of New Zealand's data collection approach", *IFC Bulletin*, no 53, April 2021.

## 2. Accessing and using micro data sets

**Once the (micro) data sets are collected, the next steps are to enable easy and effective access of this information (overcoming the challenges related to size and complexity) and to make sense of it.** In other words, it is not sufficient to "collect" the dots, what is essential is to "connect" them (Caruana (2017)) – basically meaning central banks should perform relevant analyses that can support both the preparation and monitoring of their policies. These include monetary policy, typically at the heart of central banks' missions, but also increasingly macro- (since 2011 for the ECB) as well as microprudential policies (since 2014 for the ECB in the area of banking supervision). The need for adequate information supporting these three different functions has clearly reinforced the interest in collecting granular data. As stated by the former President of the ECB, well established macro statistics will continue to provide the "big picture" but "we should also offer a magnifying glass" (Draghi (2016)).

**However, the information provided by very rich granular data sets can be overwhelming. Making it usable requires the support of ad hoc techniques** so as to transform simple "data points" into "knowledge" (Drozdova (2017)). A key success factor is therefore to have adequate staff skills (eg data scientists) as well as appropriate IT tools (eg big data analytics; Wibisono et al (2019)). Another important

lesson from central banks' work with micro data is the importance of combining different approaches, requiring the setup of multidisciplinary teams involving IT staff, subject matter statisticians and economists. One reason is the large number of different types of micro data sets that require specific competencies to be properly accessed and analysed; for instance, commercial data, administrative data and social media data sets. In addition, the interest of central banks' users has also expanded, so that projects dealing with granular data have become more frequent and serve increasingly multiple purposes – from the traditional research area to using micro data as an auxiliary input (or even sometimes as the core input) for the fine-tuning of macro policies and micro surveillance processes. Furthermore, statistical dissemination has become more complex, as an impressive and increasing wealth of information is made available to internal users and the research community. This has important consequences for the organisation of central banks' data work. One such consequence is that information architectures need to allow for combining multiple data sets, sharing data throughout the institution, and supporting various tasks such as regular statistical compilation, ad hoc data exploration and experimental activities. Another is that close cooperation is required among the various stakeholders involved, so that different sources of information can be used in a complementary way and/or for different purposes.

**These issues can pose important challenges in practice, not least because of the resource constraints faced by public institutions.** Hence, while the use of aggregated data sets has become routine for those authorities in charge of monitoring macroeconomic developments and implementing related policies, progress in extracting insightful developments from micro data sets has been much more limited. In particular, while one can relatively easily conduct standard macro aggregations out of individual data appoints, the identification of patterns of interest out of heterogeneous developments combining large granular data sets can be a much more complex task.

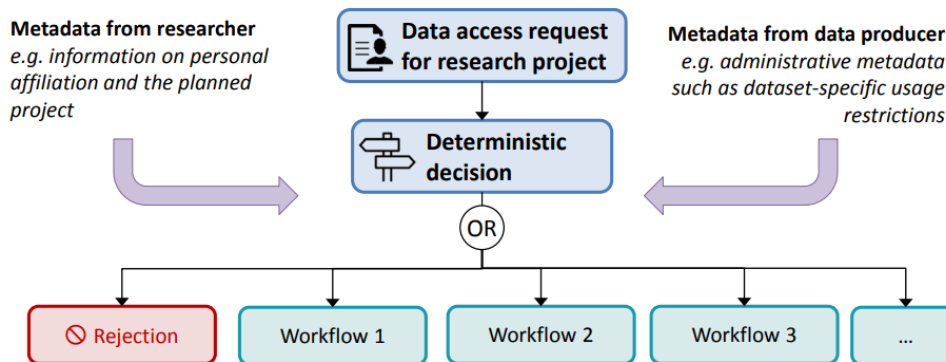
**One solution to these challenges is knowledge-sharing.** To facilitate that, several central banks, together with a number of international organisations and statistical offices, have formed the International Network for Exchanging Experience on Statistical Handling of Granular Data (INEXDA), supported by the BIS (Bender et al (2018)). A main objective of this initiative was to identify best practices in facilitating access to, and sharing of, micro data sources, for instance to develop data access procedures for researchers, as well as to explore new analytical possibilities, eg to analyse the effect of policies. This review has highlighted the benefits of automating decisions in workflows, by: (i) relying on effective “metadata schema” that describe how administrative data sets can be managed and used from a legal, organisational or technical point of view; and (ii) having workflows that follow such metadata information deterministically (Chart 5). Lastly, the international statistical community, including central banks, is also actively engaged in public sharing of IT tools in the context of SDMX (IFC (2016a)) as well as software codes through secure online software repositories (eg based on the free, open source system Git). This can greatly support the re-use of codes in open source languages such as Python and R for accessing and exploring micro data sets without requiring the actual exchange of the underlying data.

Yet a key lesson is that the devil is often in the details: the various pilots implemented by central banks have faced **concrete implementation challenges** – including the lack of standardisation of the information contained in the data sets and of the related workflows (“metadata reference”), the difficulties in setting up

adequate facilities allowing external users to access this information, and the management of confidentiality risks when results drawn from the work on micro data have to be disseminated.

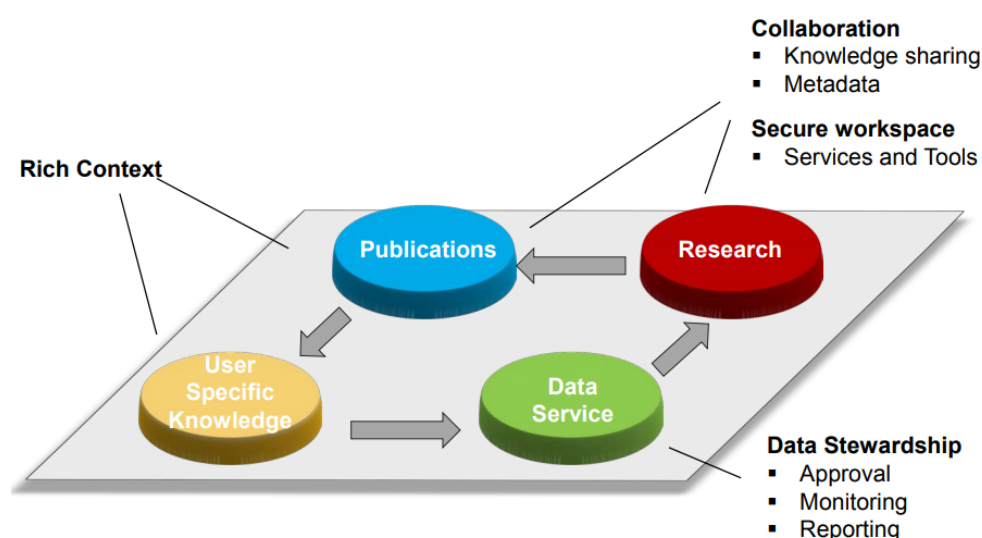
## Towards automation of decisions in workflows: INEXDA best practice recommendation

Chart 5



Source: S Bender, "Best practice recommendations on facilitating access to microdata: outcomes from the INEXDA working group on data access", *IFC Bulletin*, no 53, April 2021.

In fact, a number of successful national initiatives are worth highlighting. One relates to **dealing with external research projects using micro data, an area in which the Deutsche Bundesbank has developed significant expertise in recent years**. A key lesson is the importance of enhancing user knowledge of the data sets in question by providing rich contextual information (or "metadata"). This can be achieved through multiple steps forming a feedback loop, replacing previous one-directional information flows (Chart 6). First, by collecting information on the usage of the respective data sets; for instance, to know who has been using a specific type of data in the past and for what purpose. This can be achieved by, first, moving from "classical" to "data-centric metadata" that includes information on the use of the data and on the related administrative workflows. Second, by learning from this experience in an automated way, eg by relying on ad hoc ML techniques; for instance, to inform a user working on a specific topic what data sets are also of interest for other researchers conducting similar work – providing insights such as "people working in your area are most frequently interested in these fields". And third, by actively engaging with data users to incorporate their specific knowledge and findings and, in turn, improve the data sets' metadata information, so as to facilitate future work – eg information related to the data viewed by the user, similar data of interest, what others have done with similar data ("recipes") and what alternative recipes may be.



Source: C Hirsch, "Enhancing user knowledge by providing (richer) context to microdata", *IFC Bulletin*, no 53, April 2021.

**Another interesting initiative has been to facilitate the working of teams with different competencies to make the most of the new data landscape.** The objective is to have specialised staff, represented by different "colours", joining forces and leading to the assembling of mixed business and technology skills in composite groups of "purple people", ie those specialised staff whose skills overlap. To ensure such an adequate blending of staff skills, the Banco de Portugal has developed a tailored internal training programme in partnership with a university, noting that attracting external specialised skills like data scientists had proved to be an unsuccessful task (Chart 7). Similarly, the Netherlands Bank has developed a "Data and Technology" traineeship with the goal of developing "datapreneurs".

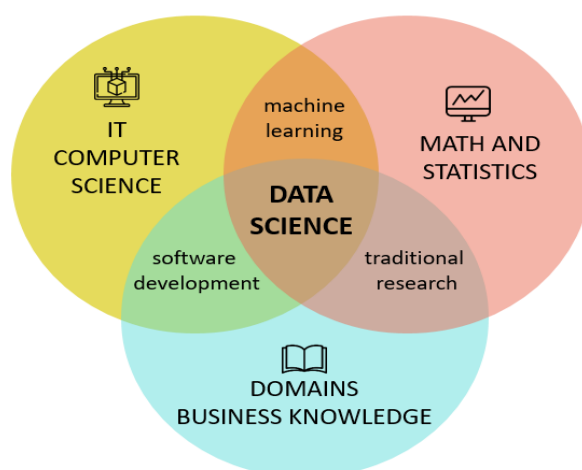
Another important point is **collaboration**. The new Integrated Data Management initiative – undertaken by Banco de Portugal as part of its strategic planning aims – promotes the creation of added value by making users, statisticians, data scientists, IT staff, supervisors etc work together. This collective approach is reported to have enhanced the use of micro data in the institution, removed cultural/organisational obstacles, and rationalised the various processes involved in terms of data collection/processing/sharing. It was also a key element supporting the Bank's internal governance model, with data stewards allocated to all business areas and tasked with overseeing the quality and fitness for purpose of the data assets. Turning to the Netherlands Bank, the focus has been on promoting discussions between data scientists and supervisors and policy officers, with the organisation of joint projects in a "hub and spoke" model.

---

The Banco de Portugal School of Data Science: the internal training programme to provide business areas with a blend of staff skills

---

Chart 7



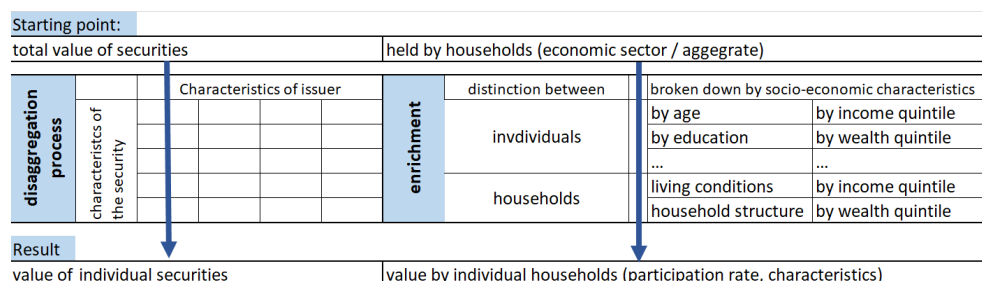
Source: L Teles Dias, "Post-crisis skills landscape: The emergence of 'Purple People'", *IFC Bulletin*, no 53, April 2021.

---

**The good news is that these efforts can rapidly pay off, with central banks increasingly able to develop informative use cases based on the micro data sets stored in their vaults** (La Cava (2015)). Moreover, experience shows that this information can improve understanding of the economy's functioning – eg to facilitate macroeconomic forecasting exercises that are based on the modelling of economic agents' behaviour (Holmquist and Hume McIntosh (2020) for a US illustration). In Austria, for instance, efforts to "bridge" micro results with developments in macro aggregates have supported the analysis of the financial behaviour of households. While macro statistics based on financial accounts had been able to document significant trends in recent years (the "what"), such as observed shifts in deposits' maturity and financial instruments, more granularity was required to understand the factors at play behind these phenomena (the "why"). In contrast, the micro level information drawn from the Consumption and Financing Survey can help identify the types of households behind observed aggregated portfolio changes (Chart 8). In particular, this approach highlighted that the shift from Austrian deposits with longer maturity to overnight deposits could be observed for nearly all types of households; but less than 10% of them (those with higher net wealth) were responsible for the aggregated shift from debt securities to investment fund shares. The work to enrich aggregated information for socioeconomic characteristics has also emphasised the need to ensure methodological and scope consistency between the different types of data.

## Enriching macro aggregates by socioeconomic characteristics using distributional information from micro-level household survey data

Chart 8



Source: J Turner, "Financial behaviour of households under the lenses of distributional aspects: the OeNB's experiences", *IFC Bulletin*, no 53, April 2021.

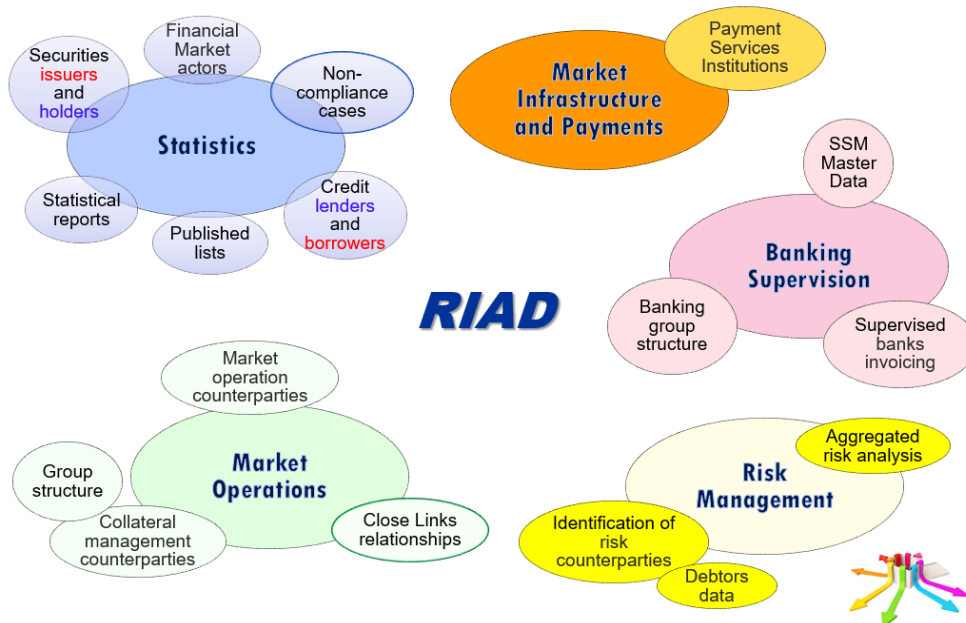
**Another successful example relates to ECB work on granular data on credit and credit risk.** Instead of developing new ad hoc surveys that may lead to inconsistent pictures, the idea has been to pull together all the granular information of potential interest in one data set (collected from AnaCredit) that can support different analyses and policy uses (Israël et al (2017)). For instance, such a rich data set can enable analysis of supply and demand credit factors depending on market conditions, small and medium-sized enterprises' specificities, the transmission channel of monetary policy, sectoral risks in the central bank's asset management operations etc. This wide range of analyses can thus assist various functions, eg monetary and macro- and microprudential policies as well as emerging policy needs related to environmental, social and governance factors. Moreover, the exercise provides value by allowing different data sets to be connected, with a holistic approach favouring consistency and interoperability across data domains. Furthermore, it was an opportunity to develop new / enhance existing statistics; for instance, data on credit proved to be more accurate when linked to a reliable business register.

To this end, the **Register of Institutions and Affiliates Database (RIAD)** was set up as the backbone of all granular statistics for the European System of Central Banks, with a unique master data set, a shared platform, high-quality checks and rules, and a comprehensive data model including reference data on individual units and relationships among them (Chart 9). However, successfully combining different data sets requires clear data definitions and methodologically sound reference information (metadata), in turn highlighting the importance of establishing a good dialogue with the financial industry (see Section 1).



Standardising and integrating existing frameworks across domains: the ECB RIAD register to integrate micro and granular data

Chart 9



Source: J-M Israël, "Linking micro datasets to better service policy-making and analyses", *IFC Bulletin*, no 53, April 2021.

### 3. New analytics for working with micro data

Due to their inherent complexity and size, granular data sets are difficult to work with, as they can require substantial time and IT capacity. This is indeed why the main applications on micro data pursued in the past have basically related to the extraction of aggregated indicators, typically using simple Structured Query Language (SQL)-based queries designed for "traditional" relational systems. **Yet increased computing power capacity and new analytical tools such as ML techniques have clearly changed the picture**, as it has become easier to work directly on the underlying data points to perform various purposes, from pure statistical production tasks to more in-depth and/or real-time complex analyses. In turn, the insights gained in understanding, for instance, the functioning of complex financial systems, assessing potential vulnerabilities and analysing contagion effects can help to effectively prepare and monitor policy actions.

**However, several challenges have to be addressed.** First is data access: one has to identify adequate sources, set up appropriate sharing arrangements, and implement data governance procedures, not least as regards the ownership of, and access to, sensitive information. A second issue is whether to opt for on-premises or cloud-based solutions. The latter are increasingly popular, not least because of their scalability and flexibility. In addition, while they can entail security risks, most notably leakages of sensitive data that may have high reputational costs for central banks (IFC (2020)), the recent experience with the Data Science Hub set up by the

Netherlands Bank shows that ways can be found to save confidential data in the cloud. A third issue is the “black box” syndrome: one has to have a good understanding of the techniques involved in order to ensure that analyses are not only accurate but also “interpretable” – so that specific explanatory causes or factors can be identified and communicated as policy-actionable “causal stories”. This again emphasises the importance of the human factor.

Despite these challenges, **one important avenue explored by central banks in making using of new analytical tools relates to their role as producer of official statistics. In particular, this is used to enhance data quality management (DQM) processes** – performed in general through both automated checks (to verify pre-determined relationships) and plausibility checks (by relying on “acceptance regions”) – eg by setting up imputation techniques, detecting outliers and duplicates. Yet, and in contrast to traditional macro statistics, the checking of large granular data sets often characterised by heterogeneous reporting patterns can be cumbersome, complex and almost impossible manually.

**Automated techniques can greatly facilitate these tasks**, as shown by the experience of the Bank of Italy in conducting ML-based quality checks on granular banking data (Chart 10). The approach, based on a supervised learning algorithm (Quantile Regression Forests), was to detect abnormal patterns in payment services data reported by banks. The results, cross-checked with reporting agents, show that this can help detect “new outliers” (ie those which had not been initially identified by existing data quality management processes) more precisely and with reasonably high accuracy. Moreover, the setup of dynamic acceptance thresholds and periodical training of the algorithms allow for an automatic update as new data come in, reducing analysts’ involvement.

A supervised learning algorithm to detect outliers in payment services data reported by banks: the potential for more precise quality checks

Chart 10

#### Cross check of outliers with banks

	<i>PI</i> <sub>1</sub>	<i>PI</i> <sub>2</sub>	<i>PI</i> <sub>3</sub>
Prediction intervals:	[ <i>q</i> <sub>0.01</sub> , <i>q</i> <sub>0.99</sub> ]	[ <i>q</i> <sub>0.025</sub> , <i>q</i> <sub>0.975</sub> ]	Inter-quartile range
<b>a-Total number of potential outliers</b>	<b>373</b>	<b>489</b>	<b>457</b>
b-Anomalies detected and revised (“true positives”)	289	312	292
c-Confirmed observations (“false positives”)	84	177	165
<b>d-Precision b/a (%)</b>	<b>77.5%</b>	<b>63.8%</b>	<b>63.9%</b>

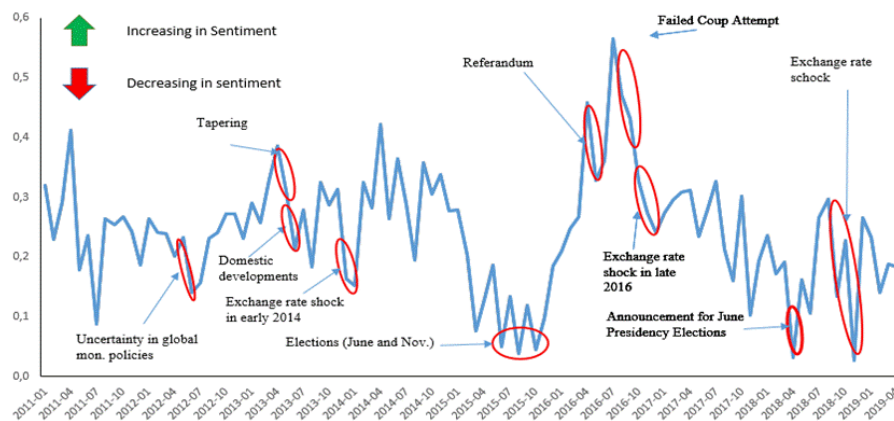
Source: F Zambuto, “Quality checks on granular banking data: an experimental approach based on machine learning”, *IFC Bulletin*, no 53, April 2021.

**Another important application is to make sense of the vast amount of textual information to extract summary indicators**, for instance economic

sentiment. This kind of approach has been implemented at the Central Bank of the Republic of Turkey to predict economic growth based on subjective information such as emotion, opinion and attitude (Chart 11). The input information was extracted from published statements based on a lexical approach and the use of ML techniques like natural language processing (NLP) for the automated processing and analysis of large amounts of text – with the setup of a term matrix to generate a vectoral representation of the document of interest, and the subsequent extraction of a summary sentiment indicator.

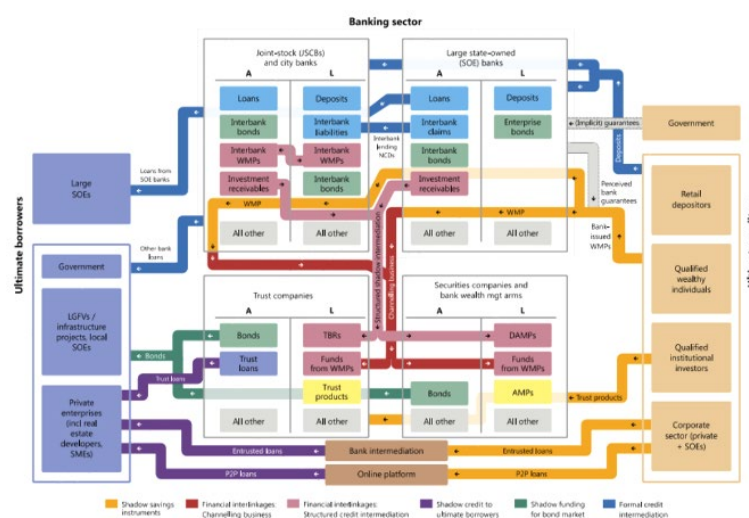
Evolution of the sentiment index for the Turkish economy

Chart 11



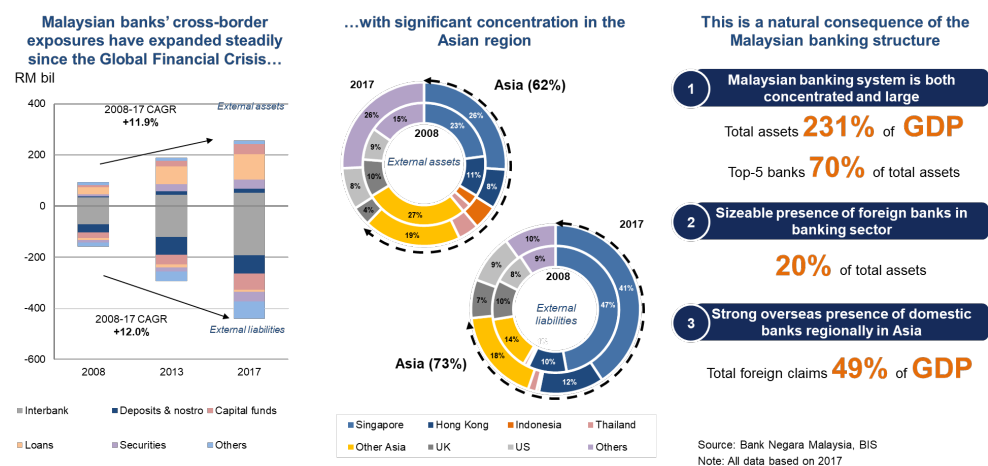
Source: B Akisoglu, "Predicting economic growth using machine learning techniques and sentiment analysis", *IFC Bulletin*, no 53, April 2021.

**A third important area relates to the analysis of financial networks** with the identification of their topological characteristics, such as the likelihood of connection between specific nodes ("connectivity") or neighbours ("clustering"), the distribution of the shortest paths among nodes ("betweenness centrality"), and the relative importance of links and nodes in the system ("average degree"). Micro-level information can be particularly suited for forming a comprehensive map of an ecosystem and its interconnections, for instance to analyse the financial system of China (Chart 12). Moreover, and as experimented by the Data Science Hub set up by the Netherlands Bank, it is an important building block for conducting stress test exercises. These typically involve designing severe circumstances that have yet to materialise, and the flexible combination of granular data from different sources to perform the assessment of such hypothetical scenarios. For instance, individual banks' data on stocks (balance sheets) and flows (transactions) can help identify troubled firms, sectors or markets, allowing the propagation of potential shocks to be traced through the financial sector.



Source: T Ehlers, S Kong and F Zhu, "Mapping shadow banking in China: structure and dynamics", *BIS Working Papers*, no 701, February 2018.

The Central Bank of Malaysia has followed a similar approach, with a **more global perspective so as to analyse international linkages and cross-border contagion risk** given the growing regionalisation of domestic banks and the significant presence of foreign banks in Malaysia (Chart 13). The work was based on granular data provided by banks which were used to develop stress-tests to: (i) assess banks' vulnerability to external shocks at the macro and micro level; (ii) estimate contagion paths across banks (that is, how the shocks propagate themselves); and (iii) identify potential systemic linkages and cross-border domino effects between internationally active banks and Malaysian banks. The results show that the solvency contagion risk of Malaysian banks to external shocks has declined in recent years, but that potential sources of contagion remain. The approach also suggests that a micro-based network analysis approach, combined with counterfactual simulation, could usefully complement macroprudential surveillance tools and support the development of effective arrangements for crisis containment, management and resolution in case unexpected shocks happen. Lastly, the modelling highlights the benefits of cooperation in terms of information-sharing and surveillance to effectively address cross-border contagion risks.



Source: H Sababathy and L Sheng Ling, "Cross-border contagion risk to the Malaysian banking system", *IFC Bulletin*, no 53, April 2021.

#### 4. Bridging the gap between macro- and micro-level statistics

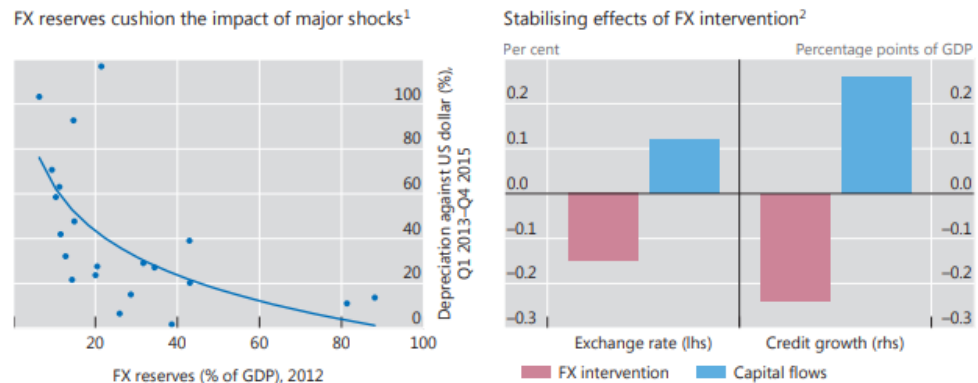
There has been a growing interest among central banks and financial supervisors in directly conducting analyses in a very granular way, especially to better prepare the actions of public authorities and help measure their effectiveness, as documented above. **Yet a key point of interest for policymakers is how to link developments on the individual institution or household level to the more general picture.** Granular data sets can surely not be a substitute for aggregated statistics, but their input can add value for analysing macroeconomic issues, for instance the functioning of the complex financial system, the evolution of inequalities, or sectoral issues.

**In order to better understand the financial system, bridging the gap between micro- and macro-level statistics can be instrumental for assessing financial stability risks, a key issue for central banks** (Crockett (2000)). The reason is that a financial crisis is by definition a system-wide, macro event that can be caused by very specific, micro factors, for instance at the level of financial institutions that have systemic importance. The challenge is therefore to extract signals that can have more general implications from the wealth of information on idiosyncratic firms. Research conducted at the BIS on predicting and preventing financial crises suggests that a good starting point is to assess the risk of individual bank distress by using bank loan data from credit registries to identify credit demand and supply factors. From there, one can draw useful insights on the general effects of policies, including on their respective contributions (Chart 14). For instance, the use of countercyclical macroprudential tools may address financial imbalances in a targeted way, while monetary policy instruments can affect the economy in a more indistinct way because of their general impact on credit or debt service ratios. Moreover, the granularity of the data can also be helpful in qualifying the results observed at the global level, eg to assess the special challenges faced by emerging market economies. One example relates to the Latin American region, where the effectiveness of financial stability

policy actions is reported to be stronger when monetary and macroprudential policies are synchronised.

## Using micro data to facilitate the understanding of the effects of macroprudential and monetary policies

Chart 14



<sup>1</sup> Based on 21 EMEs. <sup>2</sup> Coefficients from a panel regression analysis for 20 EMEs from 2000 to 2017. The dependent variable is the percentage change in the bilateral exchange rate against the US dollar (increase denotes an appreciation) and the change in the ratio of domestic credit to GDP, respectively. The regressors are the accumulation of FX reserves as well as net capital inflows, respectively, as a ratio to GDP. The control variables comprise the lagged dependent variable, the short-term interest rate spread against the United States, the log change in the VIX, the log change in the CRB commodity price index and country fixed effects.

Source: BIS, "Monetary policy frameworks in EMEs: inflation targeting, the exchange rate and financial stability", *Annual Economic Report*, 2019, Chapter II.

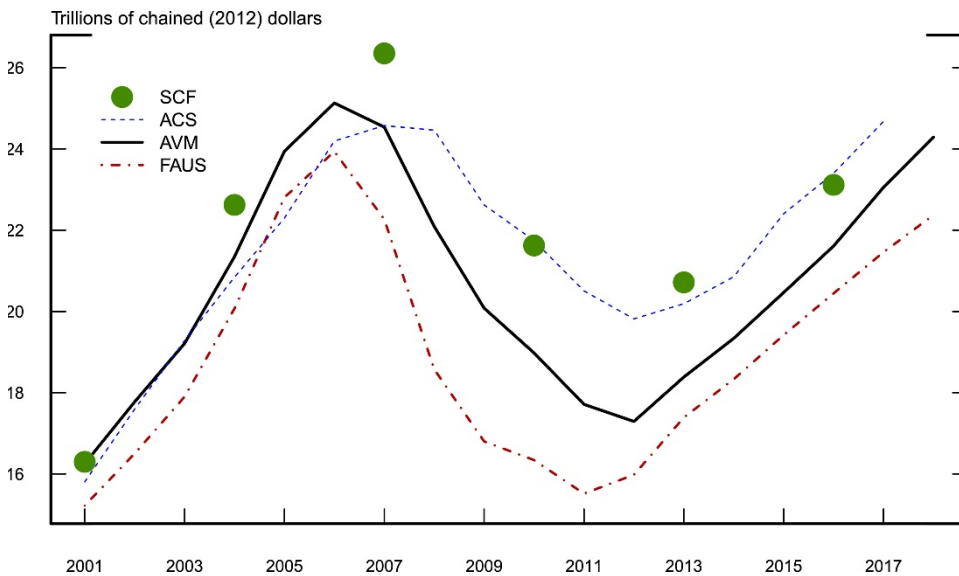
**Turning to distributional issues, the current macro framework presents notable shortcomings.** Certainly, the 2008 SNA clearly states the importance of considering the skewed distribution of income and wealth across households. But it recognises that getting this information is "not straightforward and not a standard part of the SNA" (2008 SNA, no 24.69) and that "there would be considerable analytical advantages in having micro databases that are fully compatible with the corresponding macroeconomic accounts" (2008 SNA, no 1.59). One illustration of this point is the work conducted by the Board of Governors of the US Federal Reserve System to develop Distributional Financial Accounts (DFAs), by using micro data from the Survey of Consumer Finances to distribute the aggregate household balance sheet information available from the Financial Accounts (Batty et al (2019)). Another initiative to complement macro aggregates with micro-level information has been to enhance the measurement and geographical distribution of aggregate US housing wealth, a major component of total household wealth and also an important variable to consider in quantitative macroeconomic models and analyses (Chart 15). Granular "big data" information on housing values generated by ML-based valuation models from Zillow, a private real estate and data analytics firm, was aggregated with additional data from the US Census to estimate and make public long monthly series of aggregate housing wealth (Gallin et al (2021)). Moreover, additional detailed geographical breakdowns were also publicly released. The method has proved to be robust and accurate, providing much insight into the specificities observed at local level and in a timely manner.

---

## Measuring US aggregate housing wealth based on various micro and macro data sources<sup>1</sup>

---

Chart 15



<sup>1</sup> ACS: American Community Survey (US Census Bureau); FAUS: Financial Accounts of the United States; SCF: Survey of Consumer Finance (triennial); AVM: Automated Valuation Model of the Board of Governors of the Federal Reserve System based on big data collected by Zillow.

Source: Gallin et al (2021).

---

**A final issue of interest is the better assessment of developments that are neither at the micro nor the macro level, for instance by economic sector.** In Europe, work has been conducted at Eurostat to improve understanding of the factors driving sectoral productivity developments, based on a similar approach developed by the University of Groningen (Netherlands). Taking below macro-level information (Chart 16) into account has helped produce estimates of labour productivity, supporting a wide range of analyses. A key issue was the adjustment of sectoral measures for purchasing power parities (PPPs), which are indicators of price level differences across regions or countries – knowing that, traditionally, PPPs are not estimated for sectors but from the expenditure side of GDP. The International Comparison Program was set up to calculate such price level indices for a wide range of goods and services, to be used as inputs for various studies such as real GDP comparisons or the measurement of poverty rates.



Prices:

- [Prodcom](#): unit values for over 3000 manufactured goods
- [PPP Programme](#): consumer goods and services, equipment, construction, education, health, collective services
- [Agriculture prices and price indices](#): producer prices for 128 products
- "Proxy PPPs"

Weights:

- [Prodcom](#): production values at product level (below BH level)
- [Structural business statistics](#): GVA for 4-digit NACE branches (above BH level)
- [National accounts](#): GVA for 64 NA branches (further aggregation)

Source: L Olislager, "Key indicators for sectoral productivity analysis", *IFC Bulletin*, no 53, April 2021.

---

## References

- Batty, M, J Briggs, K Pence, P Smith and A Volz (2019): "The distributional financial accounts", *FEDS Notes*, Board of Governors of the Federal Reserve System, 30 August.
- Bender, S, C Hirsch, R Kirchner, O Bover, M Ortega, G D'Alessio, L Teles Dias, P Guimarães, R, Lacroix, M Lyon, E Witt (2019): "INEXDA – the Granular Data Network", *IFC Working Paper*, no 18, October.
- Bese Goksu, E and B Tissot (2018): "Monitoring systemic institutions for the analysis of micro-macro linkages and network effects", *Journal of Mathematics and Statistical Science*, vol 4, April.
- Borio, C (2013): "The Great Financial Crisis: setting priorities for new statistics", *Journal of Banking Regulation*, vol 14, pp 306–17. Also published as *BIS Working Papers*, no 408, April.
- Broeders, D and J Prenio (2018): "Innovative technology in financial supervision (suptech) – the experience of early users", *FSI Insights on policy implementation*, no 9, July.
- Caruana, J (2017): "International financial crises: new understandings, new data", speech at the National Bank of Belgium, Brussels, February.
- Crockett, A (2000): "Marrying the micro- and macroprudential dimensions of financial stability", speech to the 11th International Conference of Banking Supervisors, 21 September, Basel.
- De Beer, B and B Tissot (2020): "Implications of Covid-19 for official statistics: a central banking perspective", *IFC Working Papers*, no 20, November.
- Draghi, M (2016): "Central bank statistics: moving beyond the aggregates", speech at the Eighth Statistics Conference, Frankfurt am Main, 6 July.
- Drozdova, A (2017): "Modern informational technologies for data analysis: from business analytics to data visualization", *IFC Bulletin*, no 43, March.
- Ducharme, L M, J Tebrake and Z Zhan (2020): "Keeping economic data flowing during COVID-19", *IMF Blog*, 26 May.
- European Central Bank (ECB) (2019): *AnaCredit Reporting Manual Part I – General Methodology*, second edition, May.
- European Commission (2020): "A European strategy for data", *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions*, COM (2020) 66 final.
- European Commission, International Monetary Fund, Organisation for Economic Co-operation and Development, United Nations and World Bank (2009): *System of National Accounts 2008*.
- Financial Stability Board (2020): *Holistic review of the March market turmoil and COVID-19 pandemic: financial stability impact and policy responses*, November.
- Financial Stability Board and International Monetary Fund (2009): *The financial crisis and information gaps*.
- (2015): *The financial crisis and information gaps – sixth implementation progress report of the G20 Data Gaps Initiative*.

Gallin, J, R Molloy, E Nielsen, P Smith and K Sommer (2021): "Measuring aggregate housing wealth: new insights from machine learning", *Journal of Housing Economics*, vol 51(C).

Greenwood, R and D Scharfstein (2013): "The growth of finance", *Journal of Economic Perspectives*, vol 27, no 2, spring, pp 3–28.

Holmquist, E and S Hume McIntosh (2020): "Policy uses of financial accounts: the US Fed's perspective", in *IFC Bulletin*, no 51, February.

Hume McIntosh, S (2021): "Measuring aggregate housing wealth: new insights from automated valuation models", *IFC Bulletin*, no 53, April.

IFC (2016a): "Central banks' use of the SDMX standard", *IFC Report*, no 4, March.

——— (2016b): "Combining micro and macro statistical data for financial stability analysis", *IFC Bulletin*, no 41, May.

——— (2017a): "Big data", *IFC Bulletin*, no 44, September.

——— (2017b): "Uses of central balance sheet data offices' information", *IFC Bulletin*, no 45, October.

——— (2017c): "Data needs and statistics compilation for macroprudential analysis", *IFC Bulletin*, no 46, December.

——— (2018): "Central banks and trade repositories derivatives data", *IFC Report*, no 7, October.

——— (2020): "Computing platforms for big data analytics and artificial intelligence", *IFC Report*, no 11, April.

Inter-Agency Group on Economic and Financial Statistics (IAG) (2017): *Update on the Data Gaps Initiative and the outcome of the Workshop on Data Sharing*, March.

Israël, J-M, V Damia, R Bonci and G Watfe (2017): "The analytical credit dataset: a magnifying glass for analysing credit in the euro area", *ECB Occasional Paper Series*, no 187, April.

La Cava, G (2015): "The development of databases linking micro and macro data – an Australian perspective", in *IFC Bulletin*, no 40, November.

Legal Entity Identifier Regulatory Oversight Committee (LEIROC) (2016): *Collecting data on direct and ultimate parents of legal entities in the Global LEI System – Phase 1*, March.

Schubert, A (2016): "AnaCredit: banking with (pretty) big data", *Central Banking Focus Report*.

Tissot (2018): "Providing comparable information to assess global financial stability risks", *Eurostat Statistical Reports*, KS-FT-18-001, January.

United Nations (2013): "Fundamental Principles of Official Statistics", resolution adopted by the Economic and Social Council, E/RES/2013/21, 28 October.

Wibisono, O, H Ari, A Widjanarti, A Zulen and B Tissot (2019): "The use of big data analytics and artificial intelligence in central banking", *The Capco Institute Journal of Financial Transformation: Data Analytics*, 50th edition, November.