Irving Fisher Committee on
Central Bank Statistics

30 July 2025

Guidance note 6 – Data integration: opportunities and challenges for central banks[1]

This guidance note discusses how the growing availability of new information sources can be integrated to maximise their use. Based on central banks' experience, it sheds light on:

- *the scope and relevance of data integration in the context of the ongoing proliferation of information sources and innovative technologies;*
- *the associated benefits for both producers and users of statistics, especially to address the need for timely and multidimensional information while minimising reporting burden;*
- *the key challenges faced in this context, notably in terms of IT infrastructure, statistical methodology and data governance; and*
- *the possible ways forward to address these challenges, highlighting the need for a holistic approach to data in central banks.*

1. Introduction

In recent years, the digitalisation of economies has driven a staggering proliferation of data sources. For central banks, the vast amount of available information has opened new opportunities to support their statistical function, enabling the production of more granular, timely and diverse statistics in a more cost-efficient way. The expansion of sources combined with innovative data techniques has also benefited their internal data users, for instance by enabling them to tap more easily into unstructured data (eg text, images, geospatial) or to better assess the impact of unpredictable events, such as pandemics (de Beer and Tissot (2020); Scotti et al (2024)).

Yet this surge in data supply poses a number of issues. The first risk is **quality**: new sources often lack the consistency, accuracy and independence of official statistics, possibly generating questionable information. In particular, they may feature various statistical biases, including scarce representativeness and incompleteness (Brault et al (2024)). Second, having abundant data at one's disposal does not always help to fill **information gaps**. These often require targeted data collections such as through surveys, which have long been considered preferred primary sources. But much of the newly available information stems from secondary data generated as by-products of other processes, such as big data and administrative records (MacFeely (2020)), which may not always be "fit for purpose" (Schubert (2025)). Third, there is a risk of **bad data crowding out good data**. Reliable statistical sources are typically costly and slow to set up, making them potentially less attractive compared to alternative, "organic" data. Furthermore, a

[1] This note was prepared by Olivier Sirello, Senior Statistical Analyst, Monetary and Economic Department, Bank for International Settlements (BIS) (olivier.sirello@bis.org) and Bruno Tissot, Head of BIS Statistics and Head of the Irving Fisher Committee on Central Bank Statistics (IFC) Secretariat (bruno.tissot@bis.org).

The authors thank Douglas Araujo, Mauro Bucci, InKyung Choi, Archana Dilip, Zlatina Hofmeister, Edward Lambe, François Robin, Luís Teles Dias, Kevin Tracol, Gabor Vincze and colleagues at Banco de Portugal and the European Central Bank for their valuable comments and suggestions.

growing number of users, including policymakers, can be tempted to rely on secondary or unofficial sources for their timeliness at the expense of higher accuracy, risking misuse due to poor awareness or inadequate handling of their caveats.[2]

The above opportunities and challenges have raised increasing interest in the concept of data integration. This process, also known as linking or fusion, basically refers to combining multiple data sources. The main goal is to fill information gaps effectively and efficiently, by leveraging the various existing data while minimising reporting burden. Another key objective is to ensure the production and enhancement of high-quality, reliable and trustworthy statistics. This calls for having adequate methodological frameworks, information standards and infrastructure, as well as, perhaps more fundamentally, a systematic and holistic view of the available data being managed and used.

Central banks' experience shows that data integration has become strategically and operationally important to perform their functions effectively. As key producers in the official data ecosystem, they need to make use of novel information types, such as micro or financial big data, in a cost-effective way (Israël and Tissot (2021)). As data users, they depend on timely and multidimensional indicators to inform their policies in various areas.

This guidance note discusses data integration in the context of central bank statistics, starting by analysing the approaches for combining various data sources (Section 2) and the benefits associated (Section 3). Section 4 sheds light on the various practical, methodological and organisational challenges encountered in this endeavour. Finally, Section 5 outlines some possible ways to make the most of the available information, underscoring the need for an efficient use of the data sources at hand.

2. The multiple facets of data integration

Data integration refers to the process of combining data and spans three key dimensions (Graph 1). First, it encompasses merging diverse sources, typically distinguished as "primary" (ie data collected for statistical purposes in line with international statistical standards, such as the UN NQAF) and "secondary" (ie data that have not been *primarily* collected for statistical purposes and can be a by-product of another activity, such as administrative, regulatory and commercial data; UNECE (2024a)). Second, it may involve linking data sourced from multiple areas, such as social, economic and environmental statistics.[3] A prominent case is the production of national accounts, which relies on a mix of various statistical domains. Third, data integration may combine different information types, such as "traditional" statistical time series with novel data types (IFC (2024a)). As an example, the integration of geospatial information with textual and sensor data can help produce experimental statistics on the environment (OECD (2018)).[4] Another typical case is the reconciliation of micro-level data with aggregate statistics, such as security-by-security databases in the balance of payments (BoP) (Dilip and Tissot (2024)).

In practice, **combining sources requires both data interoperability and a comprehensive information technology (IT) infrastructure.** Both aspects call for using well defined standards as the backbone of the data management architecture needed to support integration.

2 Cf the typical trade-off between the various dimensions of "quality", such as timeliness or relevance vs accuracy and consistency. For example, accurate yet untimely data may be of limited use for decision-making, further highlighting that "good data are (also) used data"; see Everaers (2024).

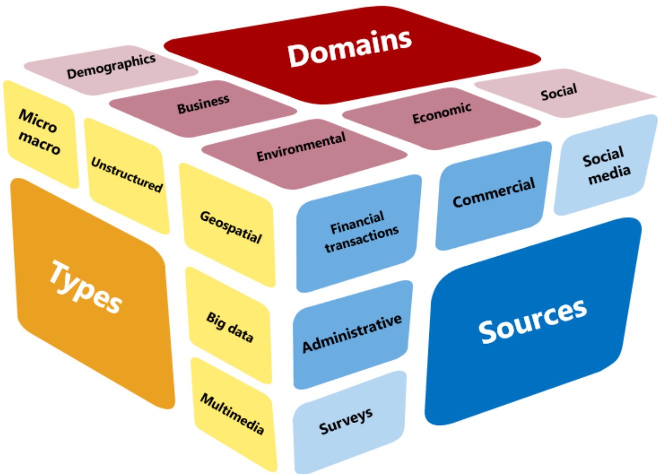
3 Cf the UN Classification of Statistical Activities (2.0), which groups statistics into seven main domains: (i) demographic and social; (ii) economic; (iii) environment; (iv) governance; (v) cross-cutting; (vi) statistical infrastructure and methodology; and (vii) strategic and managerial activities; see unstats.un.org.

4 Information is defined as processed data, for example through statistical techniques. Data correspond to unprocessed representations of facts (ie typically numerical figures) without contextual meaning per se; see MacFeely (2020).

First, **standardisation is a key component of interoperable data**.[5] Specifically, standards can support **semantic alignment** to ensure that the data to be integrated have a clear and unambiguous interpretation (Mitchell (2012)). For example, the Statistical Data and Metadata eXchange (SDMX) standard provides consistent concepts across the various statistical domains (see IFC (2025a), UNECE (2024c) and Box A).[6] Moreover, data integration requires **adequate orchestration of the various data management processes** involved. This is important, for example, to produce harmonised statistical indicators efficiently and reduce reporting burden (Mayerlen (2024)), as shown by central banks' experience in combining statistical and supervisory reporting exercises (Turner (2021)).[7] This can also be facilitated by implementing **standardised processes**, for instance the Generic Statistical Business Process Model (GSBPM),[8] which describes the main data management processes involved in the production of official statistics.

The three key dimensions of data: domains, types and sources

Graph 1



Source: authors' elaboration.

Second, having a **comprehensive IT infrastructure is a key requirement for effectively combining large varieties of data** formats and types. This calls in particular for integrating novel technologies into existing ones,[9] for instance data lakes to handle disaggregated and aggregated data in parallel (IFC (2023a)). Another important goal is to be able to support multi-mode collections, with the use of various data collection methods (eg surveys, observations etc), techniques and tools such as Application Program-

[5] Interoperability is the capacity to accurately make use of information by different parties or systems. It can be semantic (ie accurate interpretation of the data exchanged), structural (ie well defined structure and hierarchy of the information), syntactic (ie common data formats) and system (ie solutions such as protocols for communication). Standards can greatly support these different aspects of interoperability, for example by ensuring harmonised concepts, formats and applications; see UNECE (2024b).

[6] Concepts in SDMX follow a reference governance framework to ensure harmonisation (SDMX Statistical Working Group (2023)).

[7] For instance, the Banks' Integrated Reporting Dictionary (BIRD) seeks to align concepts and establish common transformation rules to comply with reporting requirements in the European Union (EU); see bird.ecb.europa.eu.

[8] See UNECE (2025a).

[9] Case in point, the UN Data Modernization Project aims to link SDMX-formatted resources with other open data formats, such as DDI-Cross Domain Integration (DDI-CDI), schema.org and Simple Knowledge Organisation System (SKOS); see ECOSOC (2025).

Box A

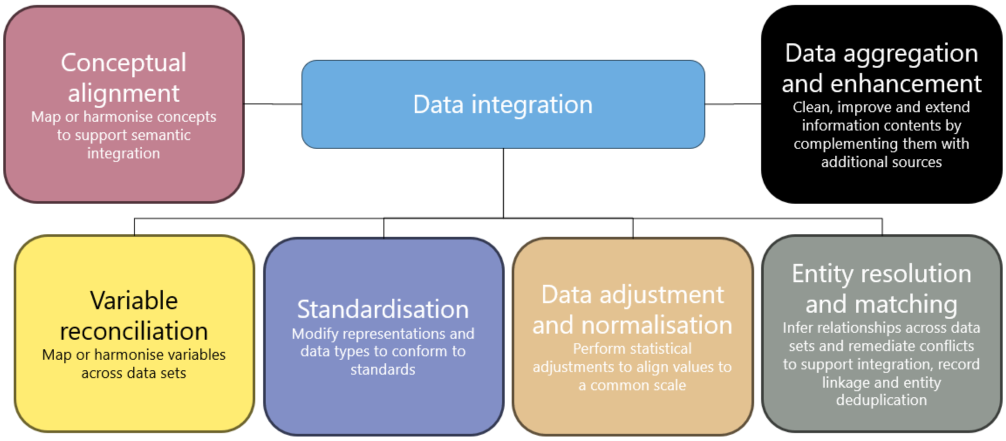
Data integration components: an architectural perspective

As a core capability of a statistical exercise, data integration relies on six main steps (Graph A.1):①

1. **Conceptual alignment** includes the harmonisation of statistical concepts;
2. **Variable reconciliation** maps or harmonises variables across the various data sets to be integrated. An important aspect is to make sure that the data sets to be linked measure identical phenomena;
3. **Standardisation** of formats and types ensures, for example, that data stored in different formats can be integrated. A related challenge is structuring unstructured data, for which AI can be useful;②
4. **Data adjustment and normalisation** involve applying statistical techniques, for instance, to align the data values to a consistent scale;
5. **Entity resolution and matching** enable mapping entries across multiple data sets (to univocally identify economic agents), using methods such as data matching or record linkage;
6. **Data aggregation and enhancement** eventually clean and improve the newly integrated information, for instance, by complementing it with additional sources.

Main operational steps supporting data integration

Graph A.1



Source: UNECE Common Statistical Data Architecture (CSDA) (version 2.0).

Turning to the actual approaches for data integration**,** there are **various IT solutions that orchestrate the integration process end-to-end**. These often involve **physical replication** to move data from various sources into a single storage location.③ For instance, the so-called "extract, transform and load" (**ETL**) pattern transforms data before loading them into storage facilities, such as warehouses. On the other hand, the "Extract, load and transform" (**ELT**) loads the raw data into a location for structuring them later in the process.④ In both cases, structuring the data often involves the use of data standards, such as SDMX. Nonetheless, physical replication can cause delays and inefficiencies for real-time, large-scale data integration processes.

To address this challenge, **virtual integration** patterns combine data at query time without physical replication, reducing costs and latency. A key approach is **data virtualisation**, which uses an abstraction layer to integrate data in their original location, enabling real-time transformations.⑤ It supports architectures like data fabric ⑥ as well as data mesh, which can arguably achieve better integration by decentralising the management of data products.⑦

① According to the Common Statistical Data Architecture (CSDA), data integration relies on four capabilities: (i) data design and description; (ii) information logistics; (iii) information-sharing; and (iv) data transformation; see unece.org. ② Berman (2018). ③ Doan et al (2012). ④ Haryono et al (2020). ⑤ Bogdanov et al (2020). ⑥ Data fabric provides a unified architecture with a single, integrated point of access to facilitate real-time access, sharing and process for users; see Östberg et al (2022). ⑦ Data mesh is a decentralised approach that distributes data access and responsibility across domain-specific teams; see Dehghani (2022).

-ming Interfaces (APIs).[10] Further, centralised registries, for example common metadata repositories, can also play a decisive role in supporting integration processes. Finally, the integration of new data tools into IT systems, including artificial intelligence (AI), has also recently become an area of key focus, particularly for supporting tasks such as data standardisation (IFC (2025b), UNECE (2024d)).

The above considerations suggest that **data integration ultimately requires a multidisciplinary and holistic approach to data**, including a full understanding of the available sources and their adequate sharing across the various stakeholders involved. This calls for developing organisational strategies both to manage the data – for instance using common metadata repositories and data catalogues –[11] but also to secure adequate resources and skills, such as professionals with multidisciplinary expertise (Damouras et al (2021)).[12]

3. The benefits of data integration

Additional analytical insights and statistical agility

Addressing the wide range of user demands for multidimensional information can be facilitated by combining various sources. Climate change statistics are a prime example of the benefits offered by data integration, both in terms of aggregation levels – ie micro and macro – and the types – ie textual documents (eg sustainability reports, press releases), images (eg satellite imagery, street view) and sensor data.[13] Data integration can also help meet policy demands in the areas of financial stability and supervision. For example, detailed assessment of banks' credit exposures often requires linking data on loan activity with business registers to identify counterparties (Wuermeling (2023)).[14] Furthermore, combining various data sources can also be useful for central banks' economic monitoring and forecasting tasks, such as measuring policy uncertainty through text analysis (Ghirelli et al (2021)).

Linking sources also enables additional and more diverse analytical perspectives. In external statistics, combining business registers and BoP data can help map global financial risks (Diz Dias et al (2025)) as well as complement traditional residency-based statistics with those that are nationality-based (McGuire et al (2024)). Linking BoP statistics with mirror data has also proven useful to better track international finance (Pradhan and Silva (2019)). Beyond external statistics, integrating payments data can unlock a wide array of new types of analysis, such as estimating consumption through credit card transactions and tourism statistics, monitoring the financial cycle, estimating loan-to-value ratios and, more broadly, tracking economic conditions through mobility indices (OMFIF and State Street (2024)).

Lastly, integrating micro and macro sources can enhance statistical agility, with benefits for both data users and compilers (Rosolia et al (2021)). For users, it allows answering questions with greater agility, for instance by "zooming in" on specific events without losing sight of macro or system-wide perspectives

[10] A key consideration is the transfer mechanism to access the various data to be integrated. In the traditional "push" model, the source sends data to the target system, while in the "pull" model the target retrieves data from the source. In practice, the IT infrastructure will have to support these two models, although the "pull" approach appears well suited to the development of APIs and present important benefits in terms of timeliness and reporting burden.

[11] This can be facilitated by the creation of dedicated units responsible for organisation-wide data integration and management (cf Banco de Portugal's example in Gonçalves et al (2023) and Moreno (2021)).

[12] This can be in contrast with the task-specific expertise typically associated with other processes like data acquisition and compilation.

[13] For additional references related to data integration and climate change statistics, see Alonso-Robisco et al (2024), Doll et al (2024), Moreno and Caminero (2022) and UNECE (2025b).

[14] For instance, Perrella and Catz (2020) also show the benefit of full integration of credit, security and business registers.

(Israël and Tissot (2021)). For producers, combining micro and macro data can allow the timely compilation of new aggregates, by swiftly rearranging granular inputs (eg “Lego bricks”) instead of setting up new data collections that are typically slower to deploy.[15]

Better use of existing data

Effective data integration allows for maximising the use of existing data sources with three main benefits, namely (i) limiting reporting burden, (ii) filling information gaps – especially in times of statistical disruptions – and (iii) improving timeliness.

As regards reporting burden, data integration can help **identify and potentially streamline redundant data collections.** This is because similar information may be collected multiple times in practice due to the absence of data inventories, inadequate sharing or governance limitations (OMFIF (2023)). A typical case relates to statistical and supervisory data exercises, which often collect identical information from the same reporting agents.[16] To address this issue, the ECB’s Integrated Reporting Framework (IReF) project aims to integrate the Eurosystem’s statistical requirements for banks and, ultimately, statistical and prudential reporting.

Another advantage of data integration is to **close information gaps in a cost-effective way**. For users, their data requests can be more easily met without setting up new data collections. Turning to producers, **integrating** various **data from administrative, credit and business registers can enhance statistical coverage**, potentially moving away from probability samples as the main basis for official statistics (Fosen et al (2025)). Case in point, the Bank of Korea used administrative data to complement the 2020 economic census to enhance its coverage of e-commerce and freelancers (Ha (2024)). Relatedly, Eurostat has developed a micro-data-linking project to leverage data from business registers – notably the EuroGroups Register – to derive additional breakdowns of structural business statistics (Ancona et al (2024)). Further, in the area of external statistics, linking BoP aggregates with security-by-security holdings data can provide additional information on the hidden wealth of households and firms (Bui Quang and Gervais (2019), Diz Dias et al (2024)).[17]

The benefits of data integration to fill information gaps can be particularly critical to cope with sudden stops in data collection and overcome periods of “statistical darkness”. The Covid-19 pandemic was a prime example, also serving as a wake-up call for official statistics to complement traditional sources with alternative data as a way to maintain statistical production during periods of distress (de Beer and Tissot (2020)). This episode underscored in particular the importance of building “rainy day data funds” or “statistical buffers” in case of shortages in the information supply chain, while also allowing for the flexible measurement of phenomena that traditional methods cannot adequately capture in times of crisis (Veronese et al (2020)).

Third, **combining sources can also improve timeliness, in particular to address urgent information needs** (Jahangir-Abdoelrahman and Tissot (2023)). For instance, leveraging online indicators offers numerous opportunities thanks to their immediate accessibility and availability (Cavallo (2015)). A well known example relates to the measurement of consumer price indices that can benefit from web-scraped data to nowcast inflation, produce higher frequency indicators and/or complement official

[15] For example, the OECD combined online job advertisements with other official data to offer timely estimates of the impact of AI on the labour market in the United Kingdom (Schmidt et al (2024)).

[16] For instance, there are approximately 200 collections for regulatory data in the United Kingdom, very often for ad hoc exercises which overlap with regular statistical templates; see Benford (2024a).

[17] Another example is the use of mobile phone data to improve surveys on travel and tourism (Carboni, Di Bella and Ciccioli (2024)). Relatedly, the integration of BoP data with granular data on trade and business surveys can help capture cross-border exchanges of creative products (Carboni, Doria and Valletta (2024)).

statistics in those jurisdictions characterised by data collection difficulties.[18] Examples of such advancements also include the compilation of consumer expenditure indicators leveraging scanner data[19] or the analysis of market sentiment by integrating textual data with more traditional indicators (Araujo et al (2025)).

Enhanced data accuracy

Combining information from various sources can be instrumental in supporting quality checks for data collected by a statistical compiler like a central bank.[20] In particular, it can assist with consistency assessments, for example by benchmarking the data against other, well identified sources to ensure that they provide a coherent description of the economic indicator being measured.[21] Relatedly, plausibility checks can also benefit from having multiple sources, especially for detecting outliers or incorrect patterns.[22] Finally, combining information can also facilitate mirror analysis by allowing the comparison of the same phenomena with independently collected observations.[23]

Additionally, **leveraging secondary data as auxiliary sources has also the potential to support overall statistical quality** through adequate assurance frameworks and processes (Brault et al (2024)). They can assist in a broad range of editing tasks, such as locating missing data or filling gaps due to delayed reporting. This can be especially useful to cope with growing non-response rates for surveys (Scotti et al (2024)).[24] Further, they can support post-survey adjustments, such as imputation techniques to estimate missing values (Sakshaug and Steorts (2023)). Finally, mixing data from different sources may play a decisive role in protecting sensitive information, for instance by enhancing statistical disclosure techniques (Woo et al (2009), UNECE (2023), Araujo et al (2025)).

Furthermore, the concept of data integration is not solely an in-house issue. It can indeed be extended to the broader perimeter of the data ecosystem, allowing for instance **central banks to benchmark their data against those produced by national statistical offices (NSOs) and international organisations**. An example is the joint work by the Bank of France and the French National Institute of Statistics and Economic Studies to ensure that their coverage of economic activities is consistent and, in turn, more accurate (Barut-Etherington and Golfier-Chataignault (2024)). Similarly, the Bank of Spain and the Spanish National Statistics Institute collaborate closely to check and improve the quality of data on foreign-controlled corporates by combining multiple sources, such as business registers, the central bank balance sheet office, foreign affiliated trade statistics, national accounts and BoP (García del Riego and Paúl (2024)).

18 The Central Bank of Armenia, for instance, has developed a methodology to produce flash estimates of the consumer price index by integrating online data (Baghdasaryan and Lazyan (2023)).

19 See initiatives by the Dutch and Mexican statistical offices (Metzemakers et al (2024), Pineda (2024)).

20 The leading role devoted to primary sources is emphasised by the UN NQAF that indicates that "statistical outputs are compared with results of other statistical or administrative sources that provide the same or similar information on the same subject matter" (UNSD (2019)). More broadly, integrating different sources to enhance the reliability of the statistical outputs has been a long standing issue in official statistics (Tillé et al (2022)).

21 Cf Statistics Canada's definition of coherence as "the degree to which it can be successfully brought together with other statistical information within a broad analytic framework and over time"; see statcan.gc.ca.

22 Bank of Italy also relies on various sources, such as business registers, to identify erroneous identifiers reported by insurance corporations (La Serra and Svezia (2023)).

23 In external statistics, for example, a foreign direct investment outflow from country B into country A, reported by B, should in principle be recorded symmetrically as an inflow by the recipient country (Jellema et al (2020)).

24 Specifically, Scotti et al (2024) show that the non-response rate for the Bank of Italy's survey of industrial and service firms has tripled since its inception in 1972.

4. Challenges and limitations

Combining multiple data sources and types does not come without challenges and costs. These can be grouped under three broad categories: (i) fragmented information standards and identifiers; (ii) IT infrastructure challenges; and (iii) statistical quality, including ethical, legal and organisational aspects.

Fragmented information standards and identifiers

One key issue limiting effective data integration is the absence of consistent information standards covering the various types of data and statistical phases. For example, SDMX is often used for macroeconomic statistics (IFC (2025a)), while other standards are considered for regulatory data (ie eXtensible Business Reporting Language, or XBRL),[25] cross-border payments (ISO 20022),[26] metadata and surveys (Data Documentation Initiative, or DDI) and geospatial information (ISO 19111 and 19115).[27] While each of them is tailored to specific needs, the presence of several standards may also increase fragmentation, costs and maintenance difficulties for data compilers.[28] Ultimately, this complexity can hinder efficient data use (Gregory et al (2024)).

Another challenge is the limited coverage of common identifiers. These are critical to join different sets of information, for instance through record linkage (see Box B). Over the past few decades, there have been a number of international initiatives to make progress in this regard, such as the Global Initiative on Unique Identifiers for Businesses (GIUIB),[29] the Legal Entity Identifier (LEI)[30] or the International Securities Identification Number (ISIN) for traded securities.[31] Yet, in practice, global identifiers still have many limitations. First, they typically rely on voluntary adoption. For example, the LEI is not mandatory across all jurisdictions and includes only firms accessing financial markets, so that it may poorly represent countries with limited financial services (Pilgrim and Ang (2024)). Second, global identifiers often compete with existing regional or national ones[32] which underpin most business registers and administrative data sets but may not be consistent. Finally, most private data brokers, such as Bloomberg ticker and Refinitiv Instrument Code (RIC), have their own identifiers making data linking more complicated.

25 For instance, see the collaboration between the European Banking Authority (EBA), the European Insurance and Occupational Pensions Authority (EIOPA) and the ECB (EBA et al (2024)).

26 See CPMI (2025).

27 Cf the survey conducted by the UNECE INGEST Task Force on Standards Issues of statistical offices on the use of standards for the integration of geospatial data with statistical information across the GSBPM; see Appendix 2 of the report at unece.org.

28 Certainly, there are a number of converters for supporting interoperability among standards, such as the one between XBRL and SDMX, but they may bring additional complexity.

29 The GIUIB aims to promote synergies between business and statistical registers; see UNSC (2023) and unstats.un.org.

30 The LEI is a 20-character, unique and global code to identify legal entities involved in financial transactions that was endorsed by key international regulators after the Great Financial Crisis of 2007-09 (FSB (2024)); see gleif.com.

31 The ISIN is an ISO standard which uniquely identifies traded securities with a 12-digit code globally.

32 Examples for securities are the Stock Exchange Daily Official List (SEDOL) – mostly used in the United Kingdom – and the Committee on Uniform Security Identification Procedures (CUSIP) – popular in Canada and the United States.

Box B

Data matching: a key method for integrating sources

Linking sources requires shared attributes such as identifiers for joining data sets. Data matching, or record linkage, is a common method for doing so.[1] Depending on whether common identifiers are available, either deterministic or probabilistic record linkage is used.

Deterministic record linkage uses common identifiers to connect sources. For example, the Eurosystem has developed a unique identifier for financial and non-financial entities through the Register of Institutions and Affiliates Database (RIAD)[2] to link micro data sets.[3] RIAD also ensures integration with other unique identifiers, such as tax codes and the Legal Entity Identifier (LEI), allowing for further integration with external sources. This type of record linkage can also be used for other applications, such as data quality management, as shown by the Bank of Italy to spot outliers in insurance data.[4]

Probabilistic record linkage integrates sources lacking identifiers. This method relies on soft or qualitative variables to build sets of potential pairs and can help fill gaps in the compilation of micro data sets.[5] For example, in the ECB's Centralised Securities Database (CSDB), securities issuers lacking a unique identifier are linked through matching algorithms that connect records by identifying similar strings in entity names.[6] Another example involves linking record-level data on financial entities, such as postal addresses, with geospatial data to produce experimental statistics on climate.[7] Matching can also help cluster time series based on their metadata and spot anomalies.[8] Finally, text mining probabilistic techniques, such as fuzzy matching, can facilitate tapping into the data contained in corporate reports.[9]

However, data matching presents some important challenges, for instance as regards the accuracy of the probabilistic techniques deployed. Moreover, confidentiality is a prominent issue, as this technique often requires the identification of individual statistical units. Fortunately, mitigation solutions include privacy-enhancing technologies, such as the generation of synthetic data.[10] Additionally, innovative data techniques may provide new opportunities, for example deep-learning-based record linkage or multi-party secure private computing, which enables computing statistical outputs without disclosing the input data.[11]

[1] D'Orazio et al (2006). [2] See ECB guideline 2018/876. [3] Perrella and Catz (2020). [4] La Serra and Svezia (2023). [5] Leulescu and Agafitei (2013). [6] Pérez and Huerga (2016). [7] Aurouet et al (2023). [8] Bogdanova et al (2023). [9] For instance as regards the ability to identify and measure the quality of climate disclosures by banks, cf Moreno and Caminero (2022). [10] Araujo et al (2025). [11] Ranbaduge et al (2024) and Ricciato (2024).

IT, security and resources issues

Linking data can raise various IT challenges. Key ones include investing in adequate solutions to handle various and complex data formats as well as to manage, store and process large data sets (Doerr et al (2021)). This raises the issue of how to migrate from legacy systems – typically designed to handle macro and structured data – to more agile, scalable and flexible solutions that can offer higher performance computing capabilities (IFC (2020)). Yet, migrations may also lead to a heterogeneous IT environment, with the (costly) need to maintain dual-system operations at least on a temporary basis and increased security risks (IFC (2023a)).

One associated challenge relates to the option of adopting cloud services, but this poses difficult trade-offs. On the one hand, the increased interest for the cloud by central banks reflects the need to find cost-effective solutions for storing and handling diverse sources like web-scraped text, scanner data and mobile phone records, which often require high scalability, computational power and, increasingly, sophisticated tools that are not available on premises (IFC (2025b)). On the other hand, cloud adoption also raises important risks, particularly in terms of potential data leaks and violations of sovereignty as well as concerns about excessive (and often quite expensive) third-party dependency (OMFIF (2023), UNECE (2024e)).

A second major issue is ensuring adequate IT security, particularly to protect restricted and/or micro level information. This typically involves developing and implementing solutions to prevent data breaches by hardening cyber security measures, which can, however, be costly to maintain. The fast-paced development of **quantum computing** may also threaten existing cryptographic techniques and, more broadly, current practices to safeguard data confidentiality and authentication (Auer et al (2025)).[33]

A third challenge is that the implementation of IT solutions may face a **shortage of in-house skills**, calling for recruiting or training qualified staff to keep up with the pace of technological developments. It also demands developing new, "hybrid" professional roles that have been traditionally rare in central banks and statistical offices, such as data engineers and data scientists (Antonucci et al (2023)). In low- and middle-income countries, the lack of staff resources, combined with intense global competition for IT skills, may in particular critically hinder the modernisation of their IT infrastructure (Hammer et al (2017)).

Fit for purpose and consistent integrated data

One key methodological concern with secondary sources is whether they are fit for the intended statistical purpose, which is a golden rule when producing high-quality statistics.[34] Since alternative sources are typically generated as a by-product of other processes, there is no guarantee that their collection is aligned with existing methodological frameworks and best practices (UNECE (2021a)). Moreover, a significant part of these indicators are made available by private providers with limited transparency and metadata information. The absence of clear documentation can create ambiguity regarding how to integrate the data, potentially leading to inconsistent outcomes, even when identical analyses are performed on the same original data sources.

The lack of robust methodology underpinning most alternative sources might exacerbate a number of statistical biases. A prominent one is related to big data samples, whose representativeness can be limited due to the biased coverage of most non-traditional sources (Tissot (2019), Cipollone (2022)).[35] Another one is hallucination, with the possibility of compounding errors by integrating inaccurate data sources. This clearly raises the risk of polluting good data with bad ones by incorporating inaccurate information into well established statistical processes (the "garbage in – garbage out" or "data contamination" syndrome).

One particular issue also relates to the **incorporation of micro-level information into existing statistical frameworks** that often rely on macroeconomic data. This integration can result in major discrepancies, not least because of inconsistencies in terms of sector classification. Further, micro data are generally organised around legal units, such as consolidated global groups spanning across several jurisdictions, rather than institutional units identified in specific sectors of the domestic economy (Tissot (2016)).

Dealing with these caveats may call for **developing and implementing adequate statistical methods and practices**. These include acknowledging data limitations in a transparent way (Gelman and Henning (2017)), providing adequate methodological guidance[36] and communicating clearly how

33 One latent danger to data confidentiality and integrity potentially posed by quantum computing is the so-called "harvest now, decrypt later" scenario, where malicious actors would harvest encrypted data today to decrypt them as soon as quantum computers become sufficiently mature to do so.

34 The *Fundamental Principles of Official Statistics* recall that the "fit for purpose" paradigm is critical to understand and meet users' needs; see UN (2014) and unstats.un.org.

35 Telling examples are the use of credit card transactions and social media texts for estimating consumption behaviour and sentiment indices, respectively. Certain strata of the population, such as younger people, are likely to be overrepresented (Einav and Levin (2013)).

36 For instance, the International Monetary Fund has developed a methodological and operational toolkit for compilers to facilitate the use of Google data to derive high-frequency statistics (Austin et al (2021)).

experimental indicators are manufactured, especially when these do not fit established statistical standards to (partially) address users' immediate needs.[37]

Statistical continuity

Combining multiple data sources could increase the likelihood of discontinuities, in turn undermining the overall quality of the statistics produced and generating breaks in final outputs, which can be hard to explain. One example of such limitations is that big data sets are often highly volatile and can create inexplicable jumps in time series (Braaksma and Zeelenberg (2015)). Another is the typically limited historical depth of a number of secondary sources, which might prevent cross-time consistency and, in turn, the compilation of long series (Woloszko (2024)). A final issue relates to the risks posed by the discontinuation of the business processes generating alternative indicators, for instance when a social media or mobile phone company stops its activities.

One possible response to this challenge is to **continuously evaluate the stability of the sources being integrated**, by engaging closely with all the stakeholders involved. Having formal data-sharing agreements with explicit clauses on data provision continuity could be another mitigation factor. In any case, it may be essential to work on contingency plans, limit overreliance on certain providers and promote the diversification of sources (Ascari et al (2020)).

Siloed structures

Compartmentalised organisational structures may hinder effective linking of data, as the information tends to be guarded by its owners in "data silos". While this can reflect important considerations (eg confidentiality protection or institutional arrangements), it also prevents any holistic approach to the information available within the institution.

In practice, **the presence of silos can generate a number of challenges for data integration** projects as well as overall data reusability. First, fragmented data collections may lead to inconsistent formats and sources across reported data, complicating the integration process, especially when there is scarce semantic and system interoperability. Second, silos can result in inefficiencies, such as with redundant data stored in multiple locations, leading to poor scalability and high storage costs. And third, silos may undermine collaboration between data users and producers, causing potentially costly errors, flawed analytical outcomes and users' redundant data requests.

Ethical and legal issues: balancing utility with social acceptability and independence

Linking sources can raise privacy, confidentiality and ethical concerns. For one, integration could lead to centralising multiple information sources,[38] resulting in negative public perceptions should individual data not be adequately safeguarded, inappropriately used and/or handled without consent (Sexton et al (2018)). A potential way to solve this issue is to remove the identifiability of individual records, but this might also significantly decrease the usefulness of the data, in turn undermining further integration – for instance through record linkage methods (see Box B above).

[37] Given the increasing role of experimental indicators, the EU has introduced the concept of "statistics under development"; see article 17g at eur-lex.europa.eu.

[38] Typical cases include using medical security records to produce better household surveys (IMF et al (2024)), mobile phone data for tracking cross-border displacements (Aydoğdu et al (2025)) or granular financial data and tax records for better estimating consumption (Rivas and Crowley (2018)).

These considerations call for establishing **robust and clear rules underpinning data integration exercises, not least to protect sensitive information**. At the international level, the *Principles governing international statistical activities* by the Committee for the Coordination of Statistical Activities (CCSA) indicate that "individual data […] are to be kept strictly confidential and used exclusively for statistical purposes or for purposes mandated by legislation".[39] Another option is developing legal arrangements to facilitate access to administrative data while ensuring the protection of privacy rights.

In addition, it is important to take adequate **operational measures to protect data confidentiality at the operational level**. A number of central banks have put in place information security management guidelines and notices, for example to prevent breaches through stringent protocols for access and control (IFC (2023c)). Further, they have created data centres or labs, particularly for researchers, to enable the integration of various granular data sets in a secure and protected environment (Brault et al (2024)). Privacy-enhancing technologies such as aggregation, synthetic data generation or anonymisation have also been developed to enhance data protection while also allowing access to a broader range of users (UNECE (2021b), Araujo et al (2025)).

Finally, the use of multiple sources – notably administrative and private ones – for statistical purposes has to be carefully weighed against the **risk of excessive reliance on external providers, which could impact – directly or indirectly – the professional independence** of compilers (Ljones (2011)). Strengthening the adherence to the Fundamental Principles of Official Statistics and safeguarding the independent role of central banks' entities responsible for producing official statistics is crucial in this regard.

5. Maximising data use through multisource statistics

Maximising data use and value through integration while mitigating the associated challenges calls for making further progress in four areas: (i) data governance and management; (ii) adequate data access and sharing; (iii) data quality; and (iv) international cooperation.

Data governance

Sound **data governance can play an important role in enabling multisource integration**, for at least five reasons.[40]

First, it helps define comprehensive **assessments**, identifying the various roles, responsibilities and accountabilities involved when combining data sources both at the organisational and broader ecosystem levels (MacFeely et al (2025a)). These include clear rules on who controls the assets, who can access them and how they can use and reuse them.

Second, it also underpins the setup of adequate **structures** to facilitate coordination of data activities and eventually break down organisational silos. One example is the growing relevance of data stewards in addition to the role of data owners to facilitate the use and reuse of data across the organisation (UN

[39] Aligned with the *Fundamental Principles of Official Statistics*, these principles apply to a large number of statistical and international organisations worldwide; see unstats.un.org. Relatedly, the Conference of European Statisticians has endorsed principles and guidelines on the confidentiality aspects of data integration (UNECE (2009)).

[40] Data governance is a multifaceted concept, which can be referred to as "a system of decision rights and accountabilities for the management of the availability, usability, integrity and security of the data and information to enable coherent implementation and coordination of data stewardship activities as well as increase the capacity to better control the data value chain"; see UNSD (2025). More generally, governance aims to secure the overall quality of statistical information.

(2020)). In addition, several central banks have taken actions to foster a common bank-wide coordination function for consolidating and streamlining data activities across the various stakeholders involved.[41]

Third, a data governance framework can help **coordinate data management processes**, in particular to prevent redundancies in collections, licences and associated costs, as well as inadequate storage solutions and limited access.

Fourth, its implementation requires developing **clear strategies for data (re)use**. This calls for recognising the importance of **data as a strategic asset** to generate high-quality information. More broadly, it also underscores the fully fledged role of data resources in modern societies as a public good, not least because of their fundamental implications in terms of representation, equity and trust in public policies. As such, (high quality) data should foster transparency in decision-making and promote inclusive access to public information (MacFeely et al (2025b)).

Fifth, a data governance framework helps to properly and actively manage data throughout their entire lifecycle (the so-called "data curation" process), ensuring their preservation, quality, reliability, integrity and access (Križman and Tissot (2022)). This should ultimately strengthen the role of **reliable statistics as a key public good infrastructure** in an increasingly global competitive information environment (UN-CEB (2023)).[42]

Data management

Integrating various sources to compile official statistics also calls for sound data management to execute and supervise the organisation, maintenance and improvement of the necessary IT infrastructure. One first action point is to develop and document **data catalogues** as unique inventory of the information assets available in the organisation. Such catalogues are key instruments helping users identify, search and make use of existing data sets.[43] Another focus point is to implement well recognised **standards** to ensure that data can be properly found, exchanged, accurately interpreted and reused, for instance by fostering semantic **interoperability**. Case in point, SDMX has been widely adopted by statistical organisations, including central banks, to support the various stages of the data life cycle, including harmonisation of metadata through common registries (IFC (2025a)).

Additionally, developments in AI are also opening new avenues for data management, especially for **data discovery applications**. Search engines are being improved to capture users' inputs through semantics rather than predefined keywords. New AI-based tools can also enhance the overall quality and accuracy of classifications, which is essential for ensuring the reliability of data catalogues. Taken together, these developments can foster better data integration processes.

[41] One solution is to establish a chief data officer function, frequently assigned to the head of statistics (see the examples of the Bank of England (Benford (2024b), Deutsche Bundesbank (Deutsche Bundesbank (2024), ECB (Stapel-Weber and Fischer (2023) and IMF (IMF (2018)) or, in contrast, clearly disentangled (cf the cases of the Board of Governors of the Federal Reserve System (FRB OIG (2023)) and Bank of France (Bank of France (2020)). Another solution is to develop oversight boards (instead of a dedicated functional role), such as the UK Transforming Data Collection Advisory Board comprising representatives from the Bank of England, the Financial Conduct Authority and the banking industry (Benford (2024a)).

[42] Cf the data strategies being developed in many jurisdictions, such as the US Federal Data Strategy, which promotes a 10-year vision for using data to serve the public interest while also safeguarding data sovereignty, privacy and confidentiality. Similarly, the European data strategy – and in particular the EU regulation 2023/2854 ("European Data Act") - aims to make data more widely used across the EU; see eur-lex.europa.eu.

[43] Cf the ECB consolidated catalogue of micro data sets within the Eurosystem, available since 2024; see ecb.europa.eu.

Data access and sharing

Fostering appropriate **access to and sharing of information is essential to facilitate data reuse** and make the most of the benefits of data integration.

As regards **data access**, central banks and other statistical organisations have already developed various solutions (de Carvalho et al (2024)). These notably include easily accessible **data portals** to foster the findability and discoverability of the disseminated statistics (IFC (2024c)); **common data platforms** that pool data together across multiple institutions;[44] and, finally, **single access data points** to foster the reuse of information by making it available to the public through centralised platforms (Araujo et al (2025)).

Central banks have been actively promoting organisation-wide **data sharing** (eg between statistical and supervisory functions), across the national statistical system (eg between the central bank and the NSO) as well as internationally (cf the GSIBSs data collection undertaken under the BIS International Data Hub; Bese Goksu and Tissot (2018)). Specific projects include the development of **open data interfaces**,[45] **data libraries**[46] and **data spaces** which support data sharing in secure and privacy-preserving environments,[47] as well as the creation of **data centres** serving as entry points for all data assets in the organisation, including micro data (Bender et al (2019)).[48] Statistical offices and central banks are also leveraging innovative data techniques, such as privacy-enhancing technologies to promote better access to and sharing of information while also safeguarding confidentiality (Kim et al (2025)).

Moreover, central banks have been engaging with other statistical organisations to promote the exchange of best practices and recommendations on data access and sharing at the international level. An example is **recommendations 13 and 14 of the third phase of the G20 Data Gaps Initiative**, which aim to establish internationally agreed principles and taxonomies on data sharing and access, particularly to guide the collaboration of statistical organisations with private companies.[49] Nevertheless, there are still important and understandable operational, legal and ethical factors constraining the scope of data sharing.

Quality assurance and curation

To be successful, data integration exercises need to incorporate the various data sources (especially secondary ones) into well established **statistical and data quality frameworks** through rigorous assurance and curation processes. One key issue is to address the arguably lower quality of emerging alternative data to be used for statistical purposes(Viggo Sæbø and Hoel (2023)). Additionally, the ongoing adoption of AI is underscoring the importance of a wide range of data quality aspects to be fostered, including availability, openness and machine readability, particularly of metadata (MacFeely (2025)).

Addressing these quality issues calls for focussing efforts in three areas. The first one is **documenting how non-official sources can be used for statistical purposes**, for instance, by making

[44] One example is the Data Commons, an open source platform by Google which aggregates public data sets, notably from the UN, the World Bank and US federal agencies.

[45] Cf Base dos Dados in Brazil; see basedosdados.org.

[46] For example, see Tobin (2024) for the National Data Library in the United Kingdom.

[47] A recent initiative in this regard is the Common European Data Spaces, launched by the European Commission as part of the European strategy for data (European Commission (2024)).

[48] For example, the IMF Big Data Center was established in 2024 to take stock of existing data assets and fully leverage them with advanced data science techniques (Maduako (2024)).

[49] The level of access to private and administrative data by central banks and statistical offices taking part in Recommendation 13 of the third phase of the G20 DGI is regularly tracked in annual progress reports, which can be accessed at imf.org.

recommendations[50] or publishing dedicated handbooks based on current experiences.[51] Central banks can play an important role in this endeavour, not least because their various functions naturally call for integrating data across multiple domains. A second area is **advancing metadata quality frameworks**, particularly to better respond to the emerging needs for AI-ready data (eg the "FAIR-R" principles; cf Verhulst et al (2025)).[52] A third objective is to further promote the development and use of **data maturity frameworks** that can help better identify the opportunities and challenges posed by data integration. Such frameworks provide organisations with a high-level benchmark to review their various data processes against different levels of sophistication (ie the maturity degrees). One recent example is the maturity framework developed by the DGI to assess institution-wide capabilities in accessing and sharing data in official statistics.[53, 54]

Collaboration and cooperation

Integrating the large variety of available data sources calls for **strengthened collaboration with all the stakeholders involved**.

Fortunately, central banks can leverage their well established **partnerships within the statistical system**, notably with NSOs, other government bodies and international organisations. Such collaborations can **broaden the palette of available sources**, particularly administrative data,[55] in turn helping to **measure complex phenomena**. The example of large cases units (LCUs) – ie specialised units collecting data on multinational enterprises – showcases how cooperation between NSOs, central banks and other government institutions can lead to a better "profiling" of international firms and understanding of the global value chains. Moreover, collaboration between central banks and other statistical actors can be instrumental in **optimising data collection**. An example is the partnership between the NSO, the Bank of England and the Financial Conduct Authority in the United Kingdom, which has helped to maximise the use of existing data reported from financial institutions (Benford (2024a)).

Collaboration may also expand to other important stakeholders in the data ecosystem, especially the **private sector, academia and citizens**. The starting point is to recognise that companies are generating immense amounts of data, which can be used for statistical purposes (UNECE (2025c)). For example, information from mobile network firms can be leveraged to produce external statistics (IFC (2024b)). Regarding academia, collaborative alliances with statistical authorities can offer various advantages to promote a better use of novel data and leverage experimental techniques, while also fostering statistical literacy (Hansen et al (2024)). Turning to the general public, its involvement ("**citizen science**") can contribute to statistical work, for instance by using social media information in the sustainable development area (Fritz et al (2019)). Similarly, citizens' participation in public data collections through **crowdsourcing** may enhance official statistics, for instance on payments (by leveraging digital transactions) and migration (ie geo records from smartphones; UNECE (2025d)).

50 Cf "Recommended practices on the use of non-official sources in international statistics" by the CCSA; see unstats.un.org.

51 The "RoadmAp to Multi-SourcE Statistics" (RAMSES) project by the UNECE High-Level Group for the Modernisation of Official Statistics is gathering use cases on multi-source statistics to be released in a handbook in 2026; see unece.org.

52 FAIR-R (FAIR2) or Findable, Accessible, Interoperable, Reusable and "Ready for AI".

53 This tool is aligned with other frameworks, such as the UN NQAF, and has been presented in the 2025 DGI Global Conference; see imf.org.

54 Another example is the UN Big Data Task Team maturity matrix for assessing the quality of big data and organisational readiness for using them (UNGWG (2020)). At the national level, countries such as the United Kingdom have also developed guidance for an organisational data maturity assessment framework; see gov.uk.

55 For instance, EU regulation 2024/3018 introduces the possibility of sharing non-confidential data between the European Statistical System and the European System of Central Banks through a secure data-sharing platform provided by Eurostat; see eur-lex.europa.eu.

While beneficial, collaboration with private stakeholders may raise some challenges. One question is how to tap into corporate information while also guaranteeing an adequate level of **accessibility, equity and inclusiveness**, particularly for countries with limited statistical capacity (Amutorine et al (2024)). Another challenge involves crowdsourcing efforts that require providing incentives to encourage citizen participation in data collections while safeguarding their **privacy**. Finally, statistics generated by the private sector may fall short of the rigorous **quality** standards of official statistics, emphasising the need for users to enhance statistical literacy to better identify limitations.

Interdisciplinary skills and literacy

Maximising the use of various data sources obviously calls for **fostering cross-functional data expertise for both users and producers**, especially to achieve synergies across data domains and, at an operational scale, across units (Nelson et al (2024)). For example, the collection of loan-by-loan information can serve the needs of both monetary policy and macroprudential supervision. But, to be effective, it may also require compilers to combine knowledge of monetary and financial statistics with prudential as well as accounting expertise. Furthermore, users of this information also have to understand the various methodologies involved and develop adequate skills to perform data analysis.

Hence, a key lesson moving forward is that the integration of the constantly evolving **data landscape requires a generalist approach** to effectively handle multiple sources and techniques. The emergence of hybrid profiles, such as data scientists – at the intersection of IT specialists, statisticians and subject matter experts (Araujo et al (2023)) – is perhaps one solution to this issue (Antonucci et al (2023)). More generally, central banks as data-dependent organisations are likely to be increasingly interested in cultivating **flexible and interdisciplinary skills** among staff, for instance by investing in a broad set of learning resources (Damouras et al (2021)).[56]

[56] The use and development of AI applications in central banks also call for advancing data and AI literacy (IFC (2025b)).

6. References

Alonso-Robisco, A, J Bas, J M Carbo, A de Juan and J M Marques (2024): "Where and how machine learning plays a role in climate finance research", *Journal of Sustainable Finance & Investment*, vol 14, no 2, pp 456–97.

Amutorine, M, N Lawrence and J Montgomery (2024): "Increasing data sharing and use for social good: lessons from Africa's data-sharing practices during the Covid-19 response", *Data and Policy*, vol 6, no e53, November.

Ancona, F, F Borrelli and T Mrlianova (2024): "New breakdowns for selected structural business statistics indicators using information from the EuroGroups Register", presentation at the United Nations Economic Commission for Europe Meeting of the Group of Experts on Business Registers, 15–18 October.

Antonucci, L, A Balzanella, E Bruno, C Crocetta, S Di Zio, L Fontanella, M Sanarico, B Scarpa, R Verde and G Vittadini (2023): "Data science skills for the next generation of statisticians", *Statistical Journal of the IAOS*, vol 39, no 4, pp 773–82.

Araujo, D, G Bruno, J Marcucci, R Schmidt and B Tissot (2023): "Data science in central banking: applications and tools", *IFC Bulletin*, no 59, October, pp 1–28.

Araujo, D, A Cap, I Mattei, R Schmidt, O Sirello and B Tissot (2025): "Data science in central banks: unlocking the potential of data", *IFC Bulletin*, no 64, May, pp 1–22.

Ascari, G et al (2020): "Quality of multisource statistics – the KOMUSO project", *The Survey Statistician*, vol 81, pp 36–51.

Auer, R, D Dodson, A Dupont, M Haghighi, N Margaine, D Marsden, S McCarthy and A Valko (2025): "Quantum-readiness for the financial system: a roadmap", *BIS Papers*, no 158, July.

Aurouet, D, A De Sanctis, D Del Giudice, J Franke, J Herzberg, M Osiewicz, R Peronaci and C Willeke (2023): "Indicators of granular exposures to climate-related physical risks for central banks' analytical purposes", in S Arslanalp, K Kostial and G Quirós-Romero (eds), *Data for a greener world: a guide for practitioners and policymakers*, IMF, Chapter 4, April, pp 59–77.

Austin, P, M Marini, A Sanchez, C Simpson-Bell and J Tebrake (2021): "Using the Google Places API and Google Trends data to develop high-frequency indicators of economic activity", *IMF Working Papers*, no 295, December.

Aydoğdu, B, Ö Bilgili, S Güneş and A Ali Salah (2025): "Mobile phone data for anticipating displacements: practices, opportunities and challenges", *Data and Policy*, vol 7, no e5, January.

Baghdasaryan, T and G Lazyan (2023): "Using non-traditional data sources for nowcasting CPI in Central Bank of Armenia", presentation at the United Nations Economic Commission for Europe meeting of the Group of Experts on National Accounts, Geneva, 25–27 April.

Bank of France (2020): *Annual Report 2019*, July.

Barut-Etherington, M-L and C Golfier-Chataignault (2024): "Three good reasons for central banks to (also) produce statistics", *Bank of France Bulletin*, no 250, article 4, March.

Bender, S, B Hausstein and C Hirsch (2019): "An introduction to INEXDA's metadata schema", *IFC Bulletin*, no 49, January.

Benford, J (2024a): "Enhancing our national asset: maximising the value of data collections across the financial sector", speech at the Regtech Summit, London, 3 October.

——— (2024b): "A weathervane for a changing world: refreshing our data and analytics strategy", speech at Big Data and AI World, part of Tech Show London, 7 March.

Berman, J (2018): "Providing structure to unstructured data", in *Principles and practice of big data*, second edition, pp 15–51.

Bese Goksu, E and B Tissot (2018): "Monitoring systemic institutions for the analysis of micro-macro linkages and network effects", *Journal of Mathematical and Statistical Science*, April.

Bogdanov, A, A Degtyarev, N Shchegoleva, V Khvatov and V Korkohv (2020): "Evolving principles of big data virtualization", *Computational Science and its Applications – ICCSA 2020*, pp 67–81.

Bogdanova, B, B Ligani, A Maurin, I Mustafi and O Sirello (2023): "Time series outlier detection using metadata and data machine learning in statistical production", presentation at UNECE Machine Learning for Official Statistics Workshop, Geneva, 5–7 June.

Braaksma, B and K Zeelenberg (2015): "Re-make/re-model: should big data change the modelling paradigm in official statistics?", *Statistical Journal of the IAOS*, vol 31, no 2, pp 193–202.

Brault, J, M Haghighi and B Tissot (2024): "Granular data: new horizons and challenges for central banks", *IFC Bulletin*, no 61, July.

Bui Quang, P and E Gervais (2019): "How to identify 'hidden securities assets' in the balance of payments: methods of Bank of France", *IFC Bulletin*, no 49, January.

Carboni, A, M Di Bella and F Ciccioli (2024): "International trade in creative economy: an analysis with granular data", *IFC Bulletin*, no 61, July.

Carboni, A, C Doria and V Valletta (2024): "Mobile phone data and the 'travel' item in the balance of payments: the Bank of Italy's experience", presentation at the 12th biennial conference of the IFC, Basel, 22–23 August.

Cavallo, A (2015): "Scraped data and sticky prices", *NBER Working Paper Series*, no 21490, August.

Cipollone, P (2022): "The experience of 10 years of data in central banking – from gathering real-time data and big data to challenges like storage or skills", speech at the international conference "Future of central banking", Vilnius, 29 September.

Committee on Payments and Market Infrastructures (CPMI) (2025): "BIS CPMI takes further steps to promote ISO 20022 harmonisation for enhanced cross-border payments", press release, 7 January.

D'Orazio, M, M Di Zio and M Scanu (2006): *Statistical matching: theory and practice*", Wiley.

Damouras, S, A Gibbs and S MacFeely (2021): "Training official statisticians for adaptive statistical practice", *Statistical Journal of the IAOS*, vol 37, no 3, pp 887–98.

de Beer, B and B Tissot (2020): "Implications of Covid-19 for official statistics: a central banking perspective", *IFC Working Papers*, no 20, November.

de Carvalho Campos, L, L M Nunes, B Tissot and M S Trincão (2024): "Communication on central bank statistics: unlocking the next level", *IFC Bulletin*, no 60, April.

Dehghani, Z (2022): *Data mesh: delivering data-driven value at scale*, O'Reilly Media.

Deutsche Bundesbank (2024): *Annual Report 2023*, February.

Dilip, A and B Tissot (2024): "Development and maintenance of a security-by-security database", *IFC Guidance Note*, no 5, May.

Diz Dias, J, M C Gómez Ramos, E Martín de Andrés, F Pastoris, M Schmitz, O Sirello and B Tissot (2025), "External statistics in a fragmented and uncertain world: addressing novel analytical challenges", *Statistical Journal of the IAOS*, vol 41, no 2, May.

Diz Dias, J, J F Silva, F Pastoris, M Ryzhenkov and M Schmitz (2024): "Where are the hidden securities in external statistics?", *IFC Bulletin*, no 62, August.

Doan, A, A Halevy and Z Ives (2012): *Principles of data integration*, Elsevier.

Doerr, S, L Gambacorta and J M Serena Garralda (2021): "Big data and machine learning in central banking", *BIS Working Papers*, no 930, March.

Doll H C, E Kormanyos, S Walter and G Alves Werb (2024): "The climate data iceberg – a depth of information to integrate", presentation at the 12th biennial conference of the IFC, Basel, 22–23 August.

Einav, L and J Levin (2013): "The data revolution and economic analysis", *NBER Working Papers*, no 19035, May.

European Banking Authority (EBA), European Central Bank (ECB) and European Insurance and Occupational Pensions Authority (EIOPIA) (2024): *Memorandum of understanding on the establishment of a common data point model governance framework – DMP Alliance*, March.

European Commission (2024): *Commission staff working document on common European data spaces*, SWD (2024) 21, January.

Everaers, P (2024): "Good data are used data: interview with Stephan Schweinfest", *Statistical Journal of the IAOS*, vol 40, no 2, May.

Federal Reserve Board Office of Inspector General (FRB OIG) (2023): "The board can enhance enterprise practices for data management roles and responsibilities", *Evaluation Report*, no 1, January.

Financial Stability Board (FSB) (2024): *Implementation of the legal entity identifier: progress report*, October.

Fritz, S et al (2019): "Citizen sciences and the United Nations Sustainable Development Goals", *Nature Sustainability*, vol 2.

Fosen, J, A Holmberg, I Jansson, D Krapavickaitė and T de Waal (2025): "Preface to the special issue on integrating data from multiple sources for production of official statistics", *Journal of Official Statistics*, vol 41, no 2, May.

García del Riego, M and I Paúl (2024): "Spanish approach on FCC statistics: MNE statistics at sight?", presentation at the OECD Conference on National Economic and Financial Accounts, November.

Gelman, A and C Henning (2017): "Beyond subjective and objective in statistics", *Journal of the Royal Statistical Society Series A: Statistics in Society*, vol 180, no 4, pp 967–1033.

Ghirelli, C, S Hurtado, J Pérez and A Urtasun (2021): "New data sources for central banks", in S Consoli, D Reforgiato Recupero and M Saisana (eds), *Data science for economics and finance*, Springer, pp 169–94.

Gonçalves, A, M Lourenço, D Sousa and T Verheij (2023): "New strategy of data sharing and data access in statistics: the view from Banco de Portugal", presentation at the IFC-Bank of Italy Workshop on "Data science in central banking: enhancing the access to and sharing of data", 17–19 October.

Gregory, A et al (2024): "Cross-domain interoperability framework (CDIF)", *WorldFAIR: Global cooperation on FAIR data policy and practice*, March.

Ha, N (2024): "2020 benchmark revision of the Korean national accounts", presentation at the OECD Conference on National Economic and Financial Accounts, November.

Hammer, C, D Kostroch and G Quiros-Romero (2017): "Big data: potential, challenges and statistical implications", *IMF Staff Discussion Notes*, no 6, September.

Hansen, J C, L M Sanchez Cespedes, L Trujillo Oyola, X K Dimakos, B Walsh, R Souza Bueno, A T Kabo-Bah, O Seidu and V Oestreich Nielsen (2024): "Collaboration between national statistical offices and academia: benefits, conditions, areas of collaboration and practical level experience in countries", *Statistical Journal of the IAOS*, vol 40, no 1, February.

Haryono, E, Fahmi, A Tri W, I Gunawan, A Hidayanto and U Rahardjia (2020): "Comparison of the E-LT vs ETL method in data warehouse implementation: a qualitative study", *2020 International conference on informatics, multimedia, cyber and information system*, pp 115–20.

International Monetary Fund (IMF) (2018): "Overarching strategy on data and statistics at the Fund in the digital age", *IMF Policy Papers*, March.

IMF, Inter-Agency Group on Economic and Financial Statistics and Financial Stability Board Secretariat (2024): *Third phase of the G20 Data Gaps Initiative (DGI-3) progress report*, October.

Irving Fisher Committee on Central Bank Statistics (IFC) (2020): "Computing platforms for big data analytics and artificial intelligence", *IFC Report*, no 11, April.

——— (2023a): "Central banks' use of time series products", *IFC Guidance Note*, no 4, December.

——— (2023c): "Data-sharing practices", *IFC Guidance Note*, no 3, March.

——— (2024a): "Granular data: new horizons and challenges", *IFC Bulletin*, no 61, July.

——— (2024b): "External statistics in a fragmented and uncertain world", *IFC Bulletin*, no 62, August.

——— (2024c): "Communication on central bank statistics", *IFC Bulletin*, no 60, April.

——— (2025a): "SDMX adoption and use of open source tools", *IFC Report*, no 17, February.

——— (2025b): "Governance and implementation of artificial intelligence in central banks", *IFC Report*, no 18, April.

Israël, J-M and B Tissot (2021)*: "Incorporating microdata into macro policy decision-making*", *Journal of Digital Banking*, vol 6, no 3.

Jahangir-Abdoelrahman, S and B Tissot (2023): "The post-pandemic new normal for central bank statistics", *Statistical Journal of the IAOS*, vol 39(3), pp 559–72.

Jellema, T, F Pastoris and C Picón Aguilar (2020): "Using synthetic indicators to assess the quality of macroeconomic statistics via mirror data", *ECB Statistics Paper Series*, no 34, March.

Kim, S, R Jansen, M Jug, D Buckley, R de Fondeville, A C Fadel, A Saxena, J Stahl and W Hsiao (2025): "A Delphi study on the role of privacy enhancing technologies in data sharing ecosystems", *Statistical Journal of the IAOS*, July.

Križman, I and B Tissot (2022): "Data governance frameworks for official statistics and the integration of alternative sources", *Statistical Journal of the IAOS*, vol 38, no 3, pp 947–55.

La Serra, V and E Svezia (2023): "A robust record linkage approach for anomaly detection in granular insurance asset reporting", *Bank of Italy Occasional Papers*, no 821, December.

Leulescu, A and M Agafitei (2013): "Statistical matching: a model based approach for data integration", *Eurostat Working Papers*.

Ljones, O (2011): "Independence and ethical issues for modern use of administrative data in official statistics", *Statistical Journal of the IAOS*, vol 27, no 1-2, February.

MacFeely, S (2020): "In search of the data revolution: has the official statistics paradigm shifted?", *Statistical Journal of the IAOS*, vol 36, no 4, pp 1075–94.

——— (2025): "The role of generative AI in generating official statistics", speech at the UNECE workshop on generative AI and official statistics, Geneva, 12 May.

MacFeely, S, A Me, R Baeven, J Costanzo, D Passarelli, C Rossini, F Schueuer, M Veerappan and S Verhulst (2025a): "Further reflections on the journey towards an international framework for data governance", *Statistical Journal of the IAOS*, vol 41, no 2, pp 337–46.

MacFeely, S, A Me, F Schueuer, J Costanzo, D Passarelli, M Veerappan and S Verhulst (2025b): "Towards a set of universal data principles", *Statistical Journal of the IAOS*, vol 41, no 1, pp 150–55.

Maduako, I (2024): "Shedding light on economic activities: GDP nowcasting with satellite data", presentation at the United Nations Network of Economic Statistics "Sprint on Artificial Intelligence and Data Science for Economic Statistics", November.

Mayerlen, F (2024): "The Eurosystem's granular data on securities, loans and entities – how we use such information today and what we may develop differently next time", *IFC Bulletin*, no 61, July.

McGuire, P, G von Peter and S Zhu (2024): "International finance through the lens of BIS statistics: residence vs nationality", *BIS Quarterly Review*, March, pp 73–88.

Metzemakers, P, O Hilverda, T Kelch, R de Jong, M van Veen and L Beele (2024): "Use of scanner data to improve the quality of consumer expenditure", presentation at the OECD Conference on National Economic and Financial Accounts, November.

Mitchell, H (2012): *Data fusion: concept and ideas*, Springer.

Moreno, M (2021): "Data governance: an orchestra of people, processes and technology", *IFC Bulletin*, no 54, July.

Moreno, A-I and T Caminero (2022): "Application of text mining to the analysis of climate-related disclosure", *International Review of Financial Analysis*, vol 83, October.

Nelson, A, P Hogle, S Zanti, S Proescholdbell and J Tenenbaum (2024): "A governance and legal framework for getting to 'yes' with enterprise-level data integration", *Data and Policy*, vol 6, no e31, August.

Official Monetary and Financial Institutions Forum (OMFIF) (2023): *Central banking and cloud services: the new frontier*.

OMFIF and State Street (2024): *Central banks in the digital age: bringing data into focus*.

Organisation for Economic Co-operation and Development (OECD) (2018): "Which strategies for NSOs in the digital era? Towards 'smart data' strategies", Committee on Statistics and Statistical Policy, August.

Östberg, P-O, E Vyhmeister, G Castañé, B Meyers and J Van Noten (2022): "Domains models and data modeling as drivers for data management: the ASSISTANT data fabric approach", *IFAC Papers*, vol 55, no 10, pp 19–24.

Paúl, I (2024): "The use of financial statements on NFC to estimate the valuation of unlisted shares and other equity of resident units in financial accounts", presentation at the OECD Conference on National Economic and Financial Accounts, November.

Pérez, A C and J Huerga (2016): "The Centralised Securities Database (CSDB) – standardised micro data for financial stability purposes", *IFC Bulletin*, no 41, May.

Perrella, A and J Catz (2020): "Integrating microdata for policy needs: the ESCB experience", *ECB Statistics Paper Series*, no 33, February.

Pilgrim, G and S Ang (2024): "The OECD-UNSD multinational enterprise information platform", *OECD Statistics Working Papers*, no 1, February.

Pineda, F (2024): "The development of a timely indicator for private consumption", presentation at the OECD Conference on National Economic and Financial Accounts, November.

Pradhan, S-K and J-F Silva (2019): "Using mirror data to track international banking", *IFC Working Papers*, no 19, December.

Ranbaduge, T, D Vatsalan and M Ding (2024): "Privacy-preserving deep learning based record linkage", *IEEE Transactions on Knowledge and Data Engineering*, vol 36, no 11, November.

Ricciato, F (2024): "Steps toward a shared infrastructure for multi-party secure private computing in official statistics", *Journal of Official Statistics*, vol 40, no 1, pp 3–15.

Rivas, L and J Crowley (2018): "Using administrative data to enhance policymaking in developing countries: tax data and the national accounts", *IMF Working Papers*, no 175, August.

Rosolia, A, S Stapel-Weber and B Tissot (2021): "New developments in central bank statistics around the world", *Statistical Journal of the IAOS*, vol 37, no 4, pp 1055–60.

Sakshaug, J and R Steorts (2023): "Recent advances in data integration", *Journal of Survey Statistics and Methodology*, vol 11, no 3, pp 513–17.

Schmidt, J, G Pilgrim and A Mourougane (2024): "Measuring the demand for AI skills in the United Kingdom", *OECD Artificial Intelligence Papers*, no 25, September.

Schubert, A (2025): "The eternal quest for quality in official statistics" in U Schipper, R Kirchner, and J Walter (eds), *Measuring international economics*, Springer, pp 171–80.

Scotti, C, C Rondinelli, M Bottone, E Mattevi and A Neri (2024): "Central bank business surveys: version 2.0", *VoxEU*, 9 December.

Sexton, A, E Shepherd, O Duke-Williams and A Eveleigh (2018): "The role and nature of consent in government administrative data", *Big Data & Society*, vol 5, no 2, December.

Stapel-Weber, S and B Fischer (2023): "The future speaks data – rethinking data governance and management in central banks: the challenges of creating a data office", presentation at the 64th ISI World Statistics Congress, Ottawa, 17 July.

Statistical Data and Metadata eXchange (SDMX) Statistical Working Group (2023): "A reference framework for SDMX structural metadata governance", *Statistical Guidelines*, March.

Tillé, Y, M Debusschere, H Luomaranta, M Axelson, E Elvers, A Holmberg and R Valliant (2022): "Some thoughts on official statistics and its future", *Journal of Official Statistics*, vol 38, no 2, June.

Tissot, B (2016): "Globalisation and financial stability risks: is the residency-based approach of the national accounts old-fashioned?", *BIS Working Papers*, no 587, October.

——— (2019): "Making the most of big data for financial stability purposes", in M Strydom and S Strydom (eds), *Big data governance and perspectives in knowledge management*, IGI Global, pp 1–24.

Tobin, J (2024): "King's speech 2024: science, technology and innovation", *House of Lords Research Briefing*, July.

Turner, J (2021): "OeNB's reporting data model as regtech/suptech solution", *IFC Bulletin*, no 53, April.

United Nations (UN) (2014): "Fundamental Principles of Official Statistics", resolution adopted by the General Assembly, A/RES/68/261.

——— (2020): *The Secretary General's data strategy: for action by everyone, everywhere with insight, impact and integrity*.

United Nations Economic and Social Council (ECOSOC) (2025): *Report of the Statistical Data and Metadata eXchange (SDMX) sponsors*, 56th session of the United Nations Statistical Commission, New York City, 4–7 March.

United Nations Economic Commission for Europe (UNECE) (2009): *Principles and guidelines on confidentiality aspects of data integration undertaken for statistical or related research purposes*, June.

——— (2021a): *Measuring the value of official statistics: testing and developing a measurement framework*, January.

——— (2021b): *Guide to sharing economic data in official statistics*, February.

——— (2023): *Synthetic data for official statistics – a starter guide*, January.

——— (2024a): *In-depth review of linking data across domains and sources*, Conference of European Statisticians, ECE/CES/2024/5, April.

——— (2024b): *Data governance framework for statistical interoperability (DAFI)*, High-level Group for the Modernisation of Official Statistics, March.

——— (2024c): *Statistical implementation standards in the context of GSBPM*, High-level Group for the Modernisation of Official Statistics, October.

——— (2024d): "Use of generative AI in statistical organizations: CES survey results", August.

——— (2024e): *Cloud for official statistics*, March.

——— (2025a): *Generic Statistical Business Process Model*, May.

——— (2025b): *Data for climate action: how national statistical offices can contribute*, ECE/CES/STAT/2024/4.

——— (2025c): *The future of national statistical offices: a call to action*, May.

——— (2025d): "Measuring migration and human mobility", 73rd Conference of European Statisticians, ECE/CES/2025/25, June.

United Nations Global Working Group (UNGWG) (2020): *Big data maturity matrix*, background paper by the Big Data Task Team on Training, Competencies and Capacity Development, November.

United Nations Statistical Commission (UNSC) (2023): "Report by the Committee of Experts on Business and Trade Statistics", *Items for discussion and decision: business and trade statistics*, fifty-fourth session, E/CN.3/2023/9.

United Nations Statistics Division (UNSD) (2019): *United Nations national quality assurance frameworks manual for official statistics*, ST/ESA/STAT/SER.M/100.

——— (2025): *Handbook on management and organization of national statistical systems*, May.

United Nations System Chief Executives Board for Coordination (UN-CEB) (2023): *International data governance – pathways to progress*, CEB/2023/1/Add.2, May.

Verhulst, S, A Zahuranec and H Chafetz (2025): "Moving toward the FAIR-R principles: advancing AI-ready data", *SSRN*, no 5164337, March.

Veronese, G, A Rosolia, F Venditti and C Biancotti (2020): "Saving economic data from Covid-19", *VoxEU*, 12 April.

Viggo Sæbø, H and M Hoel (2023): "Official statistics: quo vadis?", *Statistical Journal of the IAOS*, vol 39, no 1, pp 39–45.

Woloszko, N (2024): "Nowcasting with panels and alternative data: the OECD weekly tracker", *International Journal of Forecasting*, vol 40, no 4, pp 1302–35.

Woo, M-J, J Reiter, A Oganian and A Karr (2009): "Global measures of data utility for microdata masked for disclosure limitation", *Journal of Privacy and Confidentiality*, vol 1, no 1, pp 111–24.

Wuermeling, J (2023): "Unstructured data hold a wealth of potential for banks", speech at BaFinTech 2023, 19 September.