Irving Fisher Committee on Central Bank Statistics

BIS

Data science in central banking: unlocking the potential of data

Douglas Araujo, Adam Cap, Ilaria Mattei, Rafael Schmidt, Olivier Sirello, Bruno Tissot[1]

Executive summary

Since 2021, the Irving Fisher Committee on Central Bank Statistics (IFC) has established a series of workshops on "Data science in central banking", gathering statisticians and economists from central banks, international organisations and national statistical offices (NSOs) in addition to other stakeholders. The aim has been to review the adoption of data science in the central banking community and beyond, identify the associated opportunities and challenges as well as exchange best practices. The third edition focused on **how data science can effectively unlock the potential and value of data** by fostering their use, reuse, access and sharing.

Against this setting, **four main takeaways** stand out.

First, innovative techniques can play a central role in **leveraging traditional as well as emerging data sources and types**, with benefits for both producers and users of economic and financial information in central banks. For producers, they offer various tools to streamline statistical processes and improve data quality, especially in terms of accuracy, frequency, timeliness and granularity. For users, they enable the analysis of large, complex and/or high-dimensional data sets to extract meaningful insights or detect new patterns.

Second, data science can support **adequate and secure sharing of data sets** – particularly granular ones – without disclosing sensitive information, for instance by using tools such as privacy-enhancing technologies.

Third, **challenges persist when tapping into the increasing and various amounts of information available** today. A prominent one is the need for robust yet costly information technology (IT) resources, such as those required to handle and share large data sets. Moreover, organisational barriers can prevent the successful integration of new data techniques into existing processes, for example, because of "silos" or a limited coordination between subject matter experts, IT and data scientists. Efficient access to and use of data may also be constrained by still limited standardisation. This is particularly the case for alternative or secondary data sources,

[1] Respectively, Adviser on data science in economics (douglas.araujo@bis.org), Senior Macroeconomic Analyst (adam.cap@bis.org), Senior Financial Market Analyst (ilaria.mattei@bis.org), Head of IT, Monetary and Economic Department (rafael.schmidt@bis.org), Statistical Analyst (olivier.sirello@bis.org), Head of BIS Statistics and Head of the IFC Secretariat (bruno.tissot@bis.org).

The views expressed here are those of the authors and do not necessarily reflect those of the BIS, the IFC or any of those institutions represented at the conference.

We thank Bilyana Bogdanova, Magdalena Erdem, Mário Lourenço and Luís Teles Dias for helpful comments and suggestions.

which often lack common definitions, methodological consistency and alignment with well established standards, in turn limiting their use for compiling official statistics.

The fourth takeaway is that, fortunately, **central banks** as both users and producers of statistics **are well equipped to address these challenges**, notably thanks to their long-standing experience in big data analytics, information standards and, more broadly, data management and governance.

Looking ahead, facilitating effective data access and sharing to make the most of the information available to central banks may call for further progress in the following key areas:

- **Data management and governance** to ensure the quality of statistical information, including its availability, transparency and usability.
- **Interoperability capabilities and statistical standards** to enable effective and accurate use of data, for instance by fostering their integration, findability, access and sharing for reuse, especially by data scientists.
- **Modern, metadata-driven and easily accessible (big) data platforms** within organisations complemented by single access points for the public and data spaces more generally.
- **Strengthened collaboration** among peers and counterparts to advance data science projects, by developing IT solutions for secure and adequate data access and sharing as well as by promoting a structured exchange of experiences across interested stakeholders.

1. Introduction

The ongoing data revolution offers central banks many opportunities to inform their policy decisions with sound economic and financial analyses. As producers of official statistics, they can leverage the increasing and diverse supply of information sources to support their statistical function and fill gaps (De Beer and Tissot (2021); IFC (2023a)). Special attention has been placed on secondary data – that is, information which is not primarily collected for statistical purposes, such as geospatial data, private records or administrative registers (MacFeely (2020)). As data users, central banks can benefit from more real-time, detailed and multidimensional insights to support their mandates (IFC (2024)). To this end, they have been working on better integrating micro data into macroeconomic aggregates which they have traditionally focused on (IFC (2021a)).

Yet **harnessing this wealth of information effectively can be challenging**. Large and/or complex data sets tend to require additional processing before being able to be used for statistical purposes, reflecting their inherently limited quality and poor standardisation, as noted by the Bank of Italy's Deputy Governor Alessandra Perrazzelli. In addition, making insights available to a broader range of users, including researchers, may encounter important operational, technical, legal and ethical barriers – such as confidentiality restrictions, commercial limitations, resource shortages and IT requirements (IFC (2023b)).

Fortunately, **innovation can help overcome many of these constraints**. In particular, data science – which combines statistical, mathematical, IT and subject matter expertise – provides a wide range of tools such as big data analytics, machine learning (ML) and artificial intelligence (AI). These can be instrumental in tapping into large or complex data – such as unstructured information containing text, images, audio and video (IFC (2023c)). Innovative techniques can for instance improve data quality, extract useful insights and securely share them. Recent central bank experience also shows that these tools **can help access and share effectively the information available** to support their statistical function as well as their policy analyses.

However, while data science can be a powerful enabler, **maximising data use calls for strengthening cooperation**, knowledge-sharing and the exchange of best practices, as emphasised by the Head of BIS Statistics Bruno Tissot. Such efforts can take place at various levels, including among central banks, within national statistical systems and internationally. From an operational perspective, they may also require further improvements to the existing global statistical infrastructure, especially by developing common standards, unique identifiers and proper data management frameworks.

The overview of this *IFC Bulletin* draws on the various contributions presented in the third edition of the "Data science in central banking" IFC-Bank of Italy workshop and focuses on **how data science approaches can effectively unlock the potential of data by fostering their use, reuse, access and sharing**. Section 2 discusses the related benefits for central banks in their dual role as data producers and users. Section 3 turns to the key obstacles faced in this endeavour, especially regarding emerging data types and secondary sources. Section 4 concludes with a number of key priorities moving forward, highlighting the importance of international collaboration.

2. Accessing and making data available for further analysis

Data science can enhance the use of existing as well as novel data sources, with benefits for both producers and users of economic and financial information in central banks. For producers, it offers various tools to improve overall data quality, especially in terms of accuracy, frequency, timeliness and granularity, while also streamlining statistical processes. For users, it enables the analysis of large, complex and/or high-dimensional data sets, for instance to extract meaningful insights or detect new patterns. Finally, it supports both producers and users in adequately and safely sharing information, such as through privacy-enhancing technologies.

Improving data quality, especially in terms of accuracy, timeliness and processing

A prominent area where data science and ML can be instrumental relates to **data quality**[2] and accuracy specifically, with common applications dealing with **anomaly and error detection**. For example, the Bank of Italy employs Siamese neural networks to identify manufacturing defects and automate quality controls for the production of banknotes. Another example is from the Magyar Nemzeti Bank and Clarity Consulting, which follow a gradient-boosting ML approach to optimise error identification modelling that can be applied to both aggregate and granular data sets. In a similar vein, other work at the Bank of Italy shows a further promising use case of gradient boosting to support anti-money laundering and prudential supervision, by uncovering patterns in company financial statements that can suggest association with criminal activities.

ML can also enable the production of **more granular and timely statistics while also maintaining high accuracy**, for instance by leveraging sources such as big financial data sets or web-based indicators collected in real time. The Central Bank of Chile uses ML to analyse daily transactions from millions of electronic invoices based on their text descriptions and derive more timely consumption and investment metrics. More broadly, ML algorithms can significantly improve the speed and accuracy for processing large amounts of granular information (Liu et al (2008)), in turn facilitating the expansion of central banks' statistical offerings (Brault et al (2024)).

Furthermore, the advanced capabilities provided by **innovative tools can significantly foster the automation of statistical processes**. One prominent area is data **collection**, particularly to extract and structure information from various sources and formats, such as webpages, geodata or textual documents. For instance, the Deutsche Bundesbank uses large language models (LLMs) to extract citations and assess the associated impact of academic papers. Another project showcases how ML can facilitate the identification of financial securities eligible for specific policy purposes, by classifying the textual information extracted from their prospectus. More generally, innovative techniques can streamline processes across the entire data life cycle. For instance, the Sveriges Riksbank is developing a set of automated tools to test, integrate, deploy and orchestrate software to support the compilation of BIS international banking statistics. This has in particular helped automate data **aggregation** and **evaluation**, with consequent benefits in terms of operational continuity and efficiency. Similarly, the UK Financial Conduct Authority has been leveraging innovative supervisory technologies (SupTech) to support data **integration**, for instance by linking multiple sources to detect market abuse.

[2] "Quality" is a multifaceted concept spanning the various characteristics sought for in official statistics and captured by the Fundamental Principles of Official Statistics (UN (2014)). These include accuracy, integrity, security, transparency and trustworthiness. The concept also covers the user side, for instance to ensure that the data are fit for purpose, easily accessible, findable, traceable and reusable, and that they can eventually be deleted adequately. Lastly, it includes ethical aspects, such as assuring impartiality, objectivity, professional independence and social acceptability (ie information collected is not misused) to secure trust in official statistics (IFC (2021b)).

Dealing with large, multidimensional and novel data for analysis

Data science techniques can facilitate the analysis of large and multidimensional data sets. Clustering, neural networks or network analysis can identify uncovered patterns among many variables – a task that traditional statistical methods may struggle with (Mittal et al (2019)). A practical application is demonstrated by the Bank of Italy, which builds network graphs based on a centrality measure to show interconnections between participants in the derivatives market using granular data collected under the European Market Infrastructure Regulation (EMIR).[3] Bank Indonesia presents a similar use case deploying big data analytics and network analysis – particularly the core periphery model – to visualise interbank relationships using granular payments.

Data science can also help **tap into unstructured data such as text through natural language processing** (NLP) and LLMs. Thanks to their ability to capture relationships between words, these tools can assist users in analysing financial reports, news articles and social media (Araujo et al (2024a)).[4] A widely used application in central banking is **sentiment analysis**, a set of methods to quantify qualitative components of textual data to gain better, real-time insights into market mood, economic expectations or public perceptions of policy actions. For example, the Bangko Sentral ng Pilipinas uses FinBERT (Araci (2009)) – an encoder-based model for sentence identification finetuned specifically for financial texts – to construct sentiment indices, effectively complementing and predicting forward-looking indicators such as purchasing managers' indices. Similarly, the Board of Governors of the Federal Reserve System combines tree-structured topic modelling with BERT-based sentiment analysis to analyse developments in equity and credit default swap markets depending on the specific content released during banks' earnings calls. This research underscores the importance of sentiment analysis in understanding market reactions and the broader financial implications of communication by financial institutions.

Making use of textual data is particularly relevant for central banks for at least two reasons. First, they are themselves a key source of rich qualitative statements, for instance through various monetary policy decisions, reports, speeches and press releases. One important example is the BIS database comprising most of the speeches delivered by Governors and other central bank officials worldwide.[5] Second, textual information can be useful for assessing the impact of central bank communication on public expectations and, in turn, for forecasting economic indicators. For instance, the Deutsche Bundesbank has measured the influence of the central bank's narrative on underlying economic conditions and associated risks. Such novel techniques can also improve the forecast accuracy of a number of macroeconomic indicators, at least to some degree. Recent research suggests for

[3] New data techniques may facilitate the use of derivatives data sets, which can be challenging to analyse due to their high granularity and large volume (IFC (2018a)). For additional details on EMIR, see ec.europa.eu.

[4] For additional information on how LLMs can support the compilation of official statistics, see IFC (2025a) and UNECE (2023a).

[5] The database can be accessed at bis.org and through the open source python library *gingado* (Araujo (2023)).

instance that the analysis of the impact of European Central Bank's (ECB) monetary policy statements can improve euro area forecasts (Araujo et al (2024b)). Lastly, **nowcasting** applications often rely on the integration of data from various sources and types to "forecast" the present state of the economy (IFC (2021c)). One example is shown by the Board of Governors of the Federal Reserve System that combines the use of "standard" macroeconomic indicators with more qualitative information – especially the regular commentary in the Beige Book from US Federal Reserve banks, which provide anecdotal information on economic conditions.

Supporting adequate data access and sharing

Another area where data science can yield significant benefits relates to access to and sharing of data. This is relevant because **various obstacles still persist in making statistical information available to a broader audience**. Reasons include operational and resource constraints for producers, on one hand, and difficulties for users to effectively find the information they are looking for, on the other hand. Innovative techniques can be instrumental in both cases by allowing automatic curation and dissemination of the indicators available and, conversely, by facilitating data discovery through state-of-the-art search engines and user-friendly open data portals.[6]

The added value of data science is particularly significant for micro data, which are increasingly used by central banks.[7] Access to and sharing of disaggregated data often involve many challenges, due to their complexity and highly confidential nature.[8] Hence, there is typically a need to balance protection of sensitive information with its secure use by larger audiences, including for non-statistical purposes such as economic policy and research.[9]

Fortunately, novel data techniques, such as **privacy-enhancing technologies (PETs)**, **can mitigate these constraints** by sharing information that would otherwise be unavailable due to data protection protocols (OECD (2023)). First, **data obfuscation** can alter the data at their source by removing individual identifiers through various techniques. Second, **encrypted data processing** allows calculations to be performed while the data remain encrypted, by using techniques such as

[6] For example, the BIS Data Portal leverages the Statistical Data and Metadata eXchange (SDMX) standard to improve dissemination efficiency for producers and data discoverability for users; see Lambe and Park (2024).

[7] Micro data are "data below the level of aggregation and with a higher likelihood of identifying individual reporting units than in the aggregated data" (Israël and Tissot (2021)). Central banks have been increasingly expanding their use, including for greater precision, flexibility and timeliness compared with traditional macro indicators (Carstens (2016); IFC (2024)).

[8] For instance, the sixth principle of the Fundamental Principles of Official Statistics reads "individual data collected by statistical agencies for statistical compilation, whether they refer to natural or legal persons, are to be strictly confidential and used exclusively for statistical purposes" (UN (2014)).

[9] The Principles governing international statistical activities recommend developing "a framework describing methods and procedures to provide sets of anonymous micro data for further analysis by bona fide researchers, maintaining the requirements of confidentiality"; see unstats.un.org.

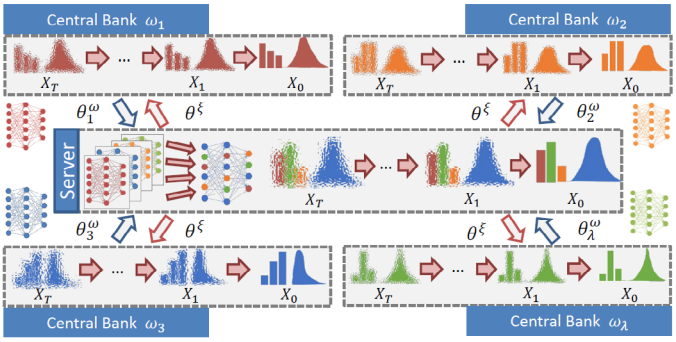
homomorphic encryption and multi-party secure private computing.[10] A third approach features **coordinated and distributed analytics**, such as federated learning that allows for the training of ML models without directly accessing the data (see Box A). Lastly, **data accountability** tools, such as personal data stores and threshold secret sharing, can help information owners and holders to retain greater control over the shared data.

In practice, **central banks have been actively adopting PETs**. A typical **data obfuscation** application is the generation of **synthetic data** to produce "fake" data that mimic the statistical properties of the original sample (Jordon et al (2022)). This technique is indeed increasingly used in official statistics and research as a way to reuse or share anonymised data (UNECE (2023b); Ravn (2025)). As analysed in a Bangko Sentral ng Pilipinas study, various tools can be used for this purpose (eg Gaussian Mixture Model, Tabular Variational Autoencoder, Generative Adversarial Networks). The best choice typically depends on various criteria, such as the quality of the synthetic data generated, the level of privacy protection and the ability to handle large amounts of information.

Central banks have also been applying other types of ML-based PETs. One example is Project Aurora by the BIS Innovation Hub Nordic Centre which relies on **network analysis** of payment data to improve the detection of money laundering activities.[11] In particular, the NSOs of Italy (Istat), Canada (Statistics Canada) and the Netherlands have explored **homomorphic encryption** techniques for analysing sensitive information collected from mobile devices. Finally, the Deutsche Bundesbank and the University of St Gallen illustrate how **federated learning** can be used to train local ML models without sharing the full data set. This approach safeguards privacy and lowers the typically high costs associated with data transfers (Graph 1).

FedTabDiff: training ML models with federated learning — Graph 1

1. Each central bank trains a **local generative model θ_i^{ω}.**
2. The Server aggregates all models into a **global generative model θ^{ξ}** accumulating knowledge from local datasets D_i^{ω}.
3. The global model θ^{ξ} is used to **generate high quality tabular data** without sharing the actual data.



Source: T Sattarov and M Schreyer, "Overcoming data-sharing challenges in central banking: federated learning of diffusion models for synthetic data generation", *IFC Bulletin*, no 64, May 2025.

[10] See Ricciato (2024) for a recent application of multi-party secure private computing in official statistics.

[11] See also Auer et al (2025) for a review of PETs in the context of digital payments data.

Box A

Sharing data with and without accessing them: exploring data science solutions

Data sharing refers to the process of making information available to others, such as individuals, entities or systems, typically through dissemination or transfer platforms. In contrast, **access** is a sharing modality and refers to the process by which authorised users obtain and use data for specific purposes, often governed by legal or commercial agreements.

The varying degrees of combining these two concepts give rise to distinct scenarios. The first occurs when **sharing data aligns with accessing them**. In practice, this option often entails secure transfers, encrypted techniques and fine-grained control of access rights. Yet, this solution is not always feasible due to scalability limitations and security concerns. A second scenario involves **sharing the insights or outputs** derived from data analysis instead of the raw data themselves. One traditional example is aggregation methods where information is shared in a summarised form, such as through averages or totals. However, this often requires centralising the data with a single holder, potentially creating concentration risks. A third scenario enables the **sharing of computations** among data holders, allowing them to retain full control over their data without needing access to others' data. The last scenario may entail **sharing information on the data themselves** (or metadata). A concrete case is making available data catalogues without granting access to the assets. Another example is exchanging information on corporate groups, for instance integrating accounting, fiscal and statistical data to streamline collections and enhance the quality of statistical information (IFC (2023b)).

Data science offers new methods to facilitate sharing without granting access to the underlying data. Federated learning, for instance, allows multiple parties to collaboratively train a machine learning model without exchanging raw data. Instead, only model updates or aggregated parameters are shared, preserving privacy and confidentiality. Another promising technology, actively explored by statistical organisations, is multi-party secure private computing, where parties can jointly compute a result without revealing their individual data, hence ensuring privacy and security throughout the process.

The above developments illustrate how **innovative techniques can support central banks in promoting efficient and secure data sharing**, with two key implications. First, they underscore the increasing focus on exchanging insights or computations without the need to share raw data, allowing for more effective use of sensitive information with significantly reduced privacy risks. Second, they point to the importance of adopting advanced encryption methods, robust cyber security measures and privacy-preserving technologies. At the same time, the emergence of quantum computing may challenge existing cryptographic practices, making early investments in **quantum-resistant solutions** essential.

3. Key obstacles in accessing, sharing and using data

While innovative data techniques can enable better access to and sharing of vast and diverse amounts of information, various challenges remain, including the need for substantial investments in IT resources, organisational barriers and the still limited information standardisation typically associated with novel or secondary sources.

The need for substantial IT resources

Obviously, **effective use of data science techniques requires state-of-the-art IT resources**. This is particularly the case when handling large and complex data sets

and facilitating their access by many stakeholders. Fortunately, central banks appear to have significantly upgraded their IT and data infrastructures over the past few decades, notably in response to the data revolution as well as to the Great Financial Crisis of 2007–09 (IFC (2020)). In particular, they have been deploying modular, scalable and metadata-driven solutions, such as data lakes, hubs and big data platforms, to efficiently store, manage and share novel or complex – including unstructured – data (Azzabi et al (2024)).

Upgrading IT infrastructures has often entailed significant transitory investments, not least because of the need to both upgrade existing technological stacks and maintain legacy solutions. For instance, a joint Bank of England and IFC survey showed that moving away from "traditional" proprietary systems such as FAME[12] has led to significant short- and medium-run costs (IFC (2023d)). One prominent reason for this is that transitioning from integrated tools can be difficult without disrupting existing setups that already work well. Another is that migrating to new solutions may require, at least initially, **additional maintenance and technical expertise** compared with more traditional applications implemented by central banks. Yet, migrations also bring important benefits, including solutions better suited for data science applications, especially those based on open source software (OSS) such as R and Python.[13] Such OSS solutions can indeed yield greater customisation, less need for (costly) user licensing requirements and reduced vendor lock-in (Araujo et al (2023)).

More fundamentally, new data science techniques require a thorough reassessment of IT resources to support central banks' activities in the longer run. One notable reason is the growing demand for **high-performance computing and efficient platforms for managing structured and unstructured data** (BIS (2024)). In particular, the use of resource-intensive applications, such as ML and big data analytics, often requires high-performance hardware (IFC (2020)).[14] One possible strategy – increasingly explored by central banks – is to rely on **cloud services**, at least in a hybrid approach combined with on-premises solutions tailored to business and security requirements, as shown by Bank Indonesia. The actual design of the IT infrastructure will in practice depend on a number of factors. For instance, cloud services offer flexible, modular and scalable solutions, especially for using customised hardware on demand (UNECE (2024a)). They can also provide **centralised storage solutions**, which may facilitate data sharing within and across organisations. Yet, the cloud may also present various limitations, starting with increased dependency on external providers. Other challenges include arguably weak data protection and sovereignty, especially when hosting facilities are located in foreign jurisdictions, raising potential legal and geopolitical risks (IFC (2025b)).

[12] FAME (Forecasting Analysis and Modelling Environment) is a proprietary statistical software and database for time series.

[13] As opposed to closed source or proprietary software, the source code of OSS can be freely edited, reproduced and redistributed; see opensource.org/osd.

[14] Such hardware includes notably graphics, tensor or neural processing units (GPUs, TPUs and NPUs, respectively) as well as clusters of computers to efficiently distribute and parallelise computations.

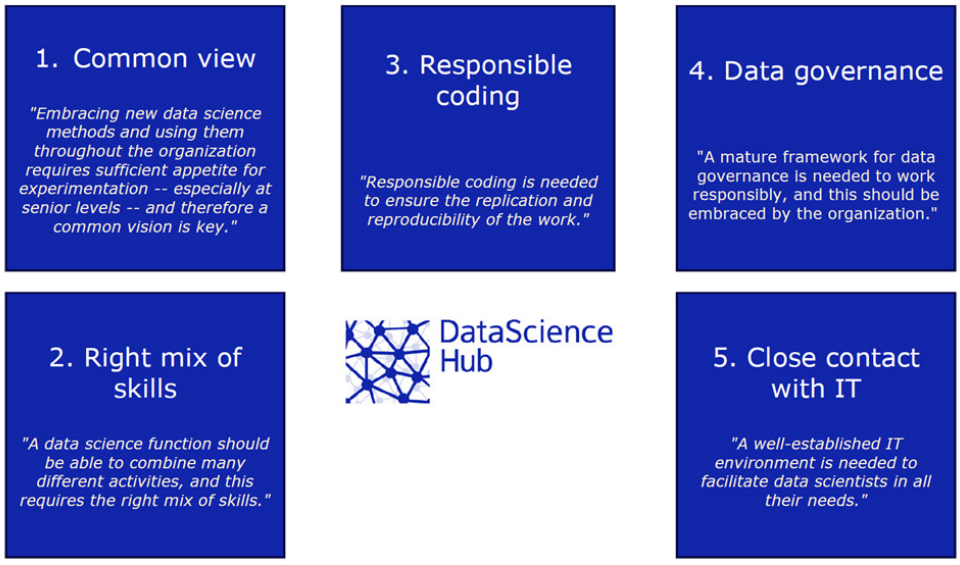
Organisational barriers

Despite the opportunities provided by innovative techniques, **the use, access to and sharing of data may face various organisational barriers** across different areas of central banking. A first challenge comes from inherent "**data silos**" (OECD (2019)). These arise because data collections and ownership are often structured organically and closely aligned with separate functional domains, such as statistics, payments, market surveillance and supervision. As a result, data tend often to be managed independently by individual business units in isolated infrastructures (IFC (2023b)). This, in turn, can limit internal user access to data and reduce awareness of available information sources (UNECE (2021)).

Limited internal coordination may be another issue. Data science fundamentally requires multidisciplinary teams (IFC (2023c)). Yet, in practice, many organisations face difficulties in promoting collaboration on projects involving subject matter experts, IT and data science teams (UNECE (2022, 2024b)). Hence, effectively meeting end users' information requirements calls for fostering synergies between business areas, which may be challenging given the diverse objectives pursued by central banks. The experience of De Nederlandsche Bank's Data Science Hub suggests that success will often depend on the following key aspects: (i) a common strategy to embrace data science; (ii) the right mix of skills; (iii) adequate IT practices, including "ethical/responsible coding" by developing software that, for instance, ensures replicability, privacy protection or accessibility; (iv) a sound data governance framework; and (v) close collaboration between stakeholders, especially data scientists, IT developers and business users (Graph 2).

Essential elements for supporting data science integration in the organisation

Graph 2



Source: P Duijm and I van Lelyveld, "Experiences, essentials and perspectives for data science in the heart of central banks and supervisors", *IFC Bulletin*, no 64, May 2025.

Finally, **reluctance to adopt data science tools in various corners of the organisation** may also raise issues. This resistance can be attributed to the diversity of profiles, with some arguably having less exposure to and/or interest in innovation. Risk aversion may also play a role, as experimenting with new data techniques can raise various challenges without yielding the expected benefits. Moreover, low levels of buy-in can result from the perceived poor transparency associated with ML techniques, often referred to as a "black box". As analysed by the International Monetary Fund (IMF), these challenges emphasise the importance of **continuous monitoring**, **upskilling** and **training** of staff. They also call for addressing complexity through measures such as **human review of AI/ML outputs** ("human-in-the-loop"). These considerations imply that adoption of data science in the organisation takes time and persistence, putting a premium on a clearly communicated, strategic and long-term approach (Duijm and van Lelyveld (2025)).

Limited standardisation hindering user access to data

Despite their benefits, the **deployment of new data science techniques is often hindered by limitations in the data themselves**. In fact, central bank users may typically need to access large granular data sets "as is", which often present complex issues (eg high multidimensionality, uncoded variables) and may **lack formal structures** compared with aggregated statistics (IFC (2024)). Moreover, many big data sets are characterised by variables formatted as free text. This can make their integration with other information sources difficult, since matching free-text variables is harder than using coded identifiers.

These difficulties are reinforced by the **limited standardisation and volatility of novel data**, at least in comparison with more traditional macroeconomic aggregates which leverage well established standards, both in terms of modelling – such as the SDMX (IFC (2025c))[15] – and methodology – such as the System of National Accounts (SNA). As noted by Banco de Portugal, micro data typically require greater modelling flexibility and customisation to handle uncoded patterns, manage complex entity relationships and navigate hierarchies to drill down from aggregates ("zoom in") or the reverse ("zoom out"). Another example relates to geospatial information, which is increasingly used by NSOs despite the lack of a fully harmonised approach across the entire data life cycle (UNECE (2024c)). Turning to methodological aspects, novel data might face sudden changes in their compilation and sources, raising inconsistencies and causing potential disruptions in the statistical chain (Benedetto et al (2025)).

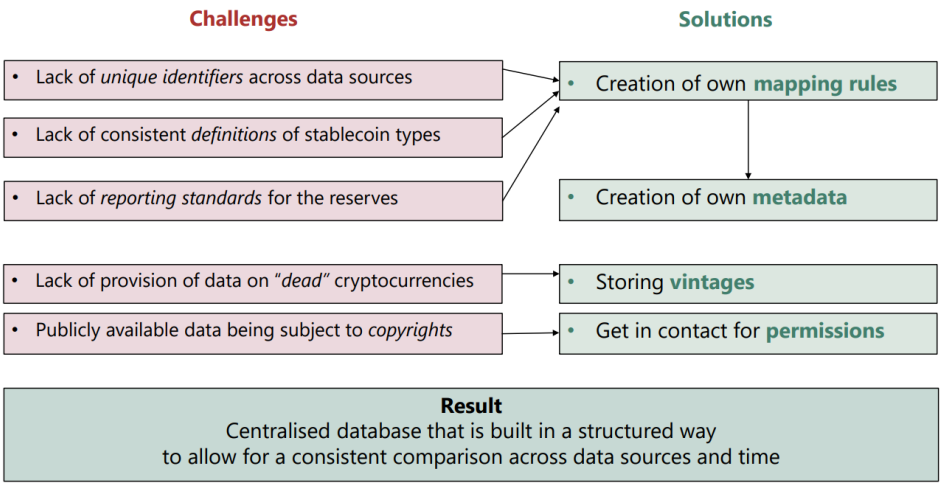
An additional challenge arises when tapping into secondary data sources to complement reference data sets, something that has become increasingly important for official statisticians. These sources often lack **unique identifiers** that are consistent across providers and domains, even though these identifiers are crucial for

[15] SDMX is an ISO-standard that facilitates the exchange, production and dissemination of statistical data and metadata between organisations. It is sponsored by eight international organisations (BIS, European Central Bank, Eurostat, International Labour Organisation, IMF, Organisation for Economic Co-operation and Development, United Nations and World Bank) and widely adopted by central banks, statistical offices and international organisations worldwide; see IFC (2025c), ECOSOC (2025) and sdmx.org.

supporting many tasks, including quality assurance and data integration. One example documented by the BIS is the absence of **common definitions and identifiers** for cryptocurrencies, which compromises consistency and comparability across multiple data sources (Graph 3). Fortunately, a number of efforts are under way to address these challenges, for instance by developing adequate metadata as well as promoting the adoption of up-to-date and easily accessible global unique identifiers, such as the Legal Entity Identifier (LEI; FSB (2024)).[16]

Linking data sets: the experience of building a database on cryptocurrencies

Graph 3



Source: A Illes and I Mattei, "Building a database on cryptocurrencies", *IFC Bulletin*, no 64, May 2025.

4. Making data access and sharing more effective

While data science can surely help, achieving effective and adequate data access and sharing may require further progress in four key areas: (i) data management and governance; (ii) interoperable information standards; (iii) common data platforms; and (iv) increased collaboration among the various stakeholders involved.

[16] The LEI is an ISO international standard to uniquely identify institutions involved in financial activities; see leiroc.org.

Advance data governance and management frameworks

Robust data governance is an essential condition to access the sheer volume of complex information available today and share insights appropriately.[17] It not only helps to ensure data quality, availability, usability, integrity and security by laying out clear sets of principles, policies and organisational frameworks. It also contributes to the adequate management of resources, particularly through the appropriate use of data science techniques. Three important benefits are worth highlighting from this perspective: access to high-quality information, its usability and appropriate sharing.

First, ensuring **access to high-quality information** is crucial, especially in terms of accuracy, completeness, lineage, traceability and transparency (IFC (2021b)). In practice, this may involve a number of solutions where data science can help. These include the development of trustworthy machine-actionable documentation of data and metadata to inform users about the sources, transformation processes, tools and methods employed. Another approach could be developing a citation standard for data, similar to current practices for software or academic research (Schwalbach and Mauer (2025)).

The second aspect is **data usability**. An organisation-wide or holistic approach to data can comprise solutions such as data inventories or catalogues to support users, including data scientists, to efficiently explore and reuse available information. For example, Banco de Portugal has adopted a master data management approach to harmonise statistical concepts across data sets in a common data dictionary, such as lists of unique codes and concepts. This strategy can also help with setting up a centralised data warehouse and a unified statistical production platform. Another important focus is to align data dissemination practices with well established guidance, such as the FAIR principles.[18]

A third consideration is to enable proper **data sharing** between interested stakeholders **while safeguarding data security and privacy**. Protecting sensitive information is very often a regulatory requirement but it is also key to securing trust among data reporters, compilers and users. Data science techniques can be useful in this regard, as illustrated by Bank Indonesia's project to protect individual data in the face of digitalisation and cyber vulnerabilities. This initiative highlighted the merits of developing a comprehensive data access framework, with adequate controls and a detailed set of information protection criteria.

[17] Data governance can be defined 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 co-ordination of data stewardship activities as well as increase the capacity (technical or otherwise) to better control the data value chain, and the resulting regulations, policies and frameworks that provide enforcement"; see UNECE (2024d).

[18] These principles promote information discovery and reuse by requiring data to be findable (detailed and searchable metadata), accessible (open and standardised), interoperable (common vocabularies) and reusable (clear licences, provenance and attributes); see Wilkinson et al (2016).

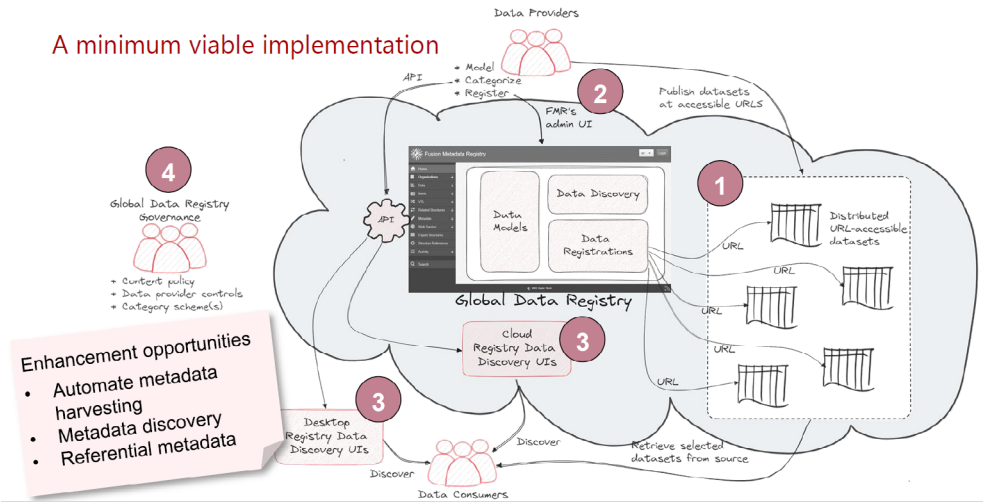
Promote interoperable standards

The development of interoperable information standards is another key factor in enabling effective and accurate use and reuse of data, especially by data scientists.

A first step is to **set up adequate data and metadata standards** to ensure accurate interpretation and seamless information exchange (UNECE (2024e)). Indeed, initiatives such as the SDMX standard, actively promoted by the central bank community, are instrumental in providing a common governance framework for data exchange, management, dissemination and discovery. In particular, and as shown by the BIS, a key strength of SDMX is the availability of a unified global metadata registry – a repository that serves as a unique reference point and source of cross-domain concepts at the international level (Graph 4; Tice et al (2025)).[19]

The SDMX global data registry to support cross-domain statistical work

Graph 4



Source: G Tice and M Nelson, "Data sharing using a global data registry, on a place to discover global structured time series, macro and micro data, *IFC Bulletin*, no 64, May 2025.

A second related consideration is to **promote and ensure interoperability among the various data standards and systems**. Key benefits include smoother data integration and reuse, reduced reporting burden and overall greater process efficiency. Central banks along with statistical offices and international institutions have been actively developing solutions, such as converters between SDMX – largely used for modelling and exchanging time series – and other standards such as XBRL – often employed for financial and micro supervisory reporting. Another progress relates to improved support for increasingly used data types, such as micro, geospatial and unstructured information.[20] Lastly, ongoing efforts are exploring how common implementation standards, such as SDMX, the Data Documentation

[19] For example, to train ML models, as done through the *gingado* library (Araujo (2023)).

[20] For example, SDMX has included support for micro and geospatial data with its version 3.0 (Nikoloutsos and Sirello (2023)).

Initiative (DDI) and Validation and Transformation Language (VTL), can help statistical organisations design metadata-driven production systems by using reference information standards, such as the Generic Statistical Business Process Model (UNECE (2024f)).

Relatedly, **interoperability within and across standards can also facilitate the implementation of decentralised data architectures** to better address business requirements ("**data mesh**"). This paradigm of data management enables domain-oriented decentralisation of analytical data operations, with distributed access, applications and responsibilities across teams (Dehghani (2022)). In addition, business areas can better tailor data processes and products to their needs while leveraging a common data modelling language. For this to be effective, standards must not only be compatible with each other but also inherently designed for interoperability, for example by ensuring that different versions of a standard can be employed to work seamlessly across teams and systems.

Finally, **standards can act as AI enablers and facilitate data findability**, as argued by Eric Anvar (OECD) in his keynote speech. For instance, the interactions between SDMX and LLMs are especially useful to enhance data discovery via natural language search while ensuring a consistent understanding across domains and business units and over the entire data life cycle.[21] Perhaps more significantly, standards can support AI systems in achieving greater accuracy through the provision of metadata or documentation about the data, hence mitigating the so-called "garbage in, garbage out" issue often faced by AI/ML solutions (IFC (2025c)).[22] Finally, standardised data could also facilitate machines in accessing information, for instance to feed AI models and digital twins (IFC (2025a); Hassani and MacFeely (2025)).

Develop common data platforms and access points

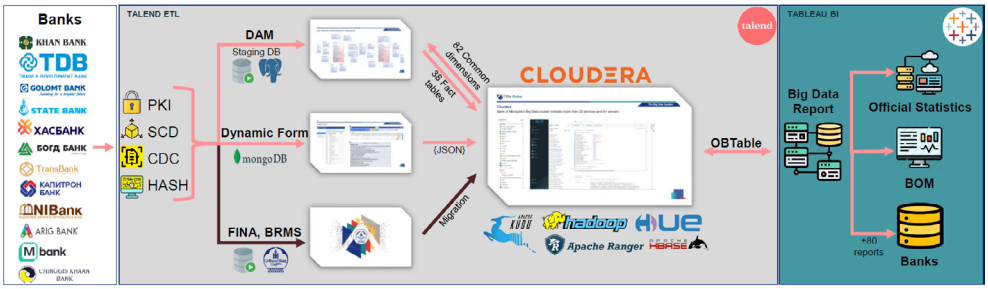
The above considerations suggest that making the most of large, diverse and complex data sets calls for setting up **modern, metadata-driven and easily accessible data platforms** within the organisation. For example, the Bank of Mongolia has developed such a platform (FinPulse) to monitor and analyse the country's financial system for regulatory purposes (Graph 5). By creating a standardised, unified regulatory database, this initiative helped **reduce reporting lags and improve timeliness** and **efficiency**. Another advantage was the strengthening of users' data access and analytical capabilities by allowing a real-time view of business intelligence reports and key indicators. Yet, building and maintaining such platforms often require massive investments in both technology and expertise. This highlights that a **strategic, institution-wide approach is essential** for successfully adopting, implementing and managing big data infrastructure, not least to address business requirements and resource constraints effectively (IFC (2020)).

[21] As a recent example, the IMF has developed an AI chatbot "StatGPT", enabling users to search for statistical data using natural language; see Kroese (2024).

[22] One related question is how to make data FAIR-R (FAIR^2) or findable, accessible, interoperable, reusable and "ready for AI"; that is, how to ensure their responsible use in AI applications for public good. See Verhulst et al (2025).

Data platform architecture: the example of the FinPulse project at Bank of Mongolia

Graph 5



The Big Data platform uses staging databases (i.e., parallel warehouse) that are kept on the bank's premises to hold raw, granular-level data using the BoM's Data Acquisition Model (DAM) format which is encrypted and transmitted to the BoM through a closed-circuit connection, and offers a single, unified source of information for both supervisors and banks to access in near real-time. Besides DAM, the platform also collects data from commercial banks and licensed institutions through manually filled Dynamic form.

1. **Data Acquisition Model (DAM):** The DAM generates an automated report daily with access in near real-time, eliminates data redundancy and inconsistency both interdepartmentally and system-wide, and can acquire and assimilate more forms of data.
2. **Dynamic Form:** Data that is not readily available through DAM will be collected using a web-based dynamic questionnaire management solution connected to the big data system. Dynamic form uses customizable data templates at a fixed frequency and is easy to make additional changes and updates to the reports.

Source: T Dorjpurev, "Big data platform (FinPulse) initiative", *IFC Bulletin*, no 64, May 2025.

Offering single access points can be another option to facilitate access, especially when data ownership is split across various organisations and domains. The European Single Access Point (ESAP) is a concrete example that connects national data platforms via a common data registry.[23] As the analysis by the Business Reporting – Advisory Group shows, the ESAP – whose initial rollout is currently planned for 2027 – offers the advantage of providing data in a standardised and machine-readable format. Relatedly, the development of "**data spaces**" is another solution for enhancing data access and reuse through common data infrastructures and governance frameworks with a clear structure for participants to access, share and collaborate on their data assets. One key example is the Common European Data Spaces (European Commission (2024)).

Lastly, central banks in particular have developed specific data access solutions to improve academic researchers' use of their micro data (IFC (2024)). Examples of such **research data centres** include on-site infrastructures such as the Deutsche Bundesbank's Research Data and Service Centre or the Bank of Italy's LabBI, which provide secure environments for accessing confidential information.[24] Other possibilities offer **remote execution systems**, such as the Bank of Italy's REX, allowing external participants to perform statistical operations on granular records without having direct access to or visualisation of the underlying data.

Foster collaboration through the sharing of tools and experience

Central banks have been leading the way to promote collaboration among their peers and counterparts to advance their data science projects as well as develop solutions

23 See the ESAP regulation at consilium.europa.eu.

24 The ECB maintains a list of the research data centres and a catalogue of available micro data sets within the European System of Central Banks; see ecb.europa.eu.

for better data access and sharing. Two main approaches have typically been followed, namely a technical one – by developing commonly available IT solutions – and an organisational one – by promoting a structured exchange of experience across interested stakeholders.

First, at the technical level, **making software available as an open source** can play a key role in fostering collaboration among developers, data scientists and, broadly, users. Tangible benefits include economies of scale, more creativity, innovation and transparency (Araujo et al (2023)). It can also help reduce the risk of bias or errors in the methods and algorithms used, although it might increase complexity.

In practice, **central banks have been increasingly producing OSS**, including for economic modelling and official statistics. As documented by the BIS, they are already sharing software codes by **open sourcing macroeconomic models**. Their experience provides a useful benchmark for new projects, based on criteria such as accessibility, documentation and replicability (Graph 6). This work also offers practical advice for enhancing code availability. In parallel, a number of **open source tools for official statistics** have been developed through the BIS OpenTech and sdmx.io initiatives.[25] Key ones are the Fusion Metadata Registry, Fusion Workbench and SDMX Dashboard Generator, which support SDMX processes across the data life cycle. More generally, leveraging OSS and common data standards can be key for central banks as producers of official statistics to develop a comprehensive open data strategy, as analysed by Regnology.

Benchmarking macroeconomic model open sourcing — Graph 6

Core criteria

- Open access
  - Easy to find
  - Free access
- Documentation
  - Academic description
  - API documentation
  - Input data documentation
  - Output documentation
- Replication
  - End-to-end execution
  - Instructions/tutorial/vignette
  - Minimum requirements

Additional criteria

- Software
  - Software open availability
  - Testing
  - Technology neutrality
  - Availability of past versions
  - Explicit code versioning
- Contribution
  - Contact with software maintainers
  - 3rd party contributions possible
  - Contribution guidelines
- License
  - Explicit open source license

Source: D Araujo, "Open sourced central bank macroeconomic models", *IFC Bulletin*, no 64, May 2025.

Second, from an organisational perspective, collaboration may require a coordinated involvement of the stakeholders willing to share data, knowledge and best practices. To this end, central banks have been actively developing a **community of practice** to exchange ideas continuously, not least in the context of the IFC. A number of initiatives have also aimed to extend such knowledge-sharing beyond the

[25] See bis.org and sdmx.io, respectively.

central banking community, including to academia and international organisations. An example is the **International Network for Exchanging Experience on Statistical Handling of Granular Data** (INEXDA), which facilitates the exchange of information on micro data issues (IFC (2018b)).

Collaboration can be strengthened by the ongoing development of **common international data frameworks**. Important and promising developments in this regard are the revisions of core statistical methodologies (such as the newly revised SNA 2025 and Balance of Payments Manual - BPM7), the data collections initiated in the context of the **Data Gaps Initiative** (DGI)[26] and the various actions undertaken to strengthen the global statistical infrastructure – including in terms of registers, identifiers and data standards. Taken together, these initiatives can be instrumental in facilitating better access to alternative data for official statistics and fostering adequate sharing of economic and financial information for the public good.

[26] The DGI is an international initiative endorsed by the G20 and supported by several countries. Launched in the aftermath of the Great Financial Crisis, it entered its third phase in 2022 to specifically address four statistical areas: (i) climate change; (ii) household distributional information; (iii) fintech and financial inclusion; and (iv) access to private sources of data and administrative data, and data sharing. See IMF et al (2023) and imf.org.

References

Araci, D (2009): "FinBERT: Financial sentiment analysis with pre-trained language models", arXiv.

Araujo, D (2023): "gingado: a machine learning library focused on economics and finance", *BIS Working Papers*, no 1122, September.

Araujo, D, S Doerr, L Gambacorta and B Tissot (2024a): "Artificial intelligence in central banking", *BIS Bulletin*, no 84, January.

Araujo, D, N Bokan, F Comazzi and M Lenza (2024b): "Word2Prices: embedding central bank communications for inflation prediction", *CEPR Discussion Paper*, no 19784, December.

Araujo, D, S Nikoloutsos, R Schmidt and O Sirello (2023): "Central banks as users and providers of open-source software", Box 1 in "Data science in central banking: applications and tools", *IFC Bulletin,* no 59, October, pp 8–10.

Auer, R, R Böhme, J Clark and D Demirag (2025): "Privacy-enhancing technologies for digital payments: mapping the landscape", *BIS Working Papers*, no 1242, January.

Azzabi S, Z Alfughi and A Ouda (2024): "Data lakes: a survey of concepts and architectures", *Computers*, vol 13, no 7, 183.

Bank for International Settlements (BIS) (2024): "Artificial intelligence and the economy: implications for central banks", *BIS Annual Economic Report*, chapter III, pp 91–127.

Benedetto, C, S Crestini, A de Gregorio, M de Leonardis, A del Monaco, D Gulino, P Massaro, F Monacelli and L Rubeo (2025): "Applying artificial intelligence to support regulatory reporting management: the experience at Banca d'Italia", *Bank of Italy Occasional Papers*, no 927, April.

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

Carstens, A (2016): "Micro-data as a key input to designing macro-prudential policy: the Mexican experience", *ECB Conference on Statistics*, July.

De Beer, B and B Tissot (2021): "Official statistics in the wake of the Covid-19 pandemic: a central banking perspective", *Theoretical Economics Letters*, vol 11, no 4, August.

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

Duijm, P and I van Lelyveld (2025): "Data science for central banks and supervisors: how to make it work, actually", *Harvard Data Science Review*, vol 7, no 1.

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

Financial Stability Board (FSB) (2024), *Implementation of the Legal Entity Identifier: progress report*, October.

Hassani, H and S MacFeely (2025): "Digital twin and official statistics: a paradigm shift?", *Statistical Journal of the IAOS*, April.

International Monetary Fund (IMF), Inter-Agency Group on Economic and Financial Statistics and Financial Stability Board Secretariat (2023): *People planet economy: delivering insights for action*, G20 Data Gaps Initiative 3 Workplan, March.

Irving Fisher Committee on Central Bank Statistics (IFC) (2018a): "Central banks and trade repositories derivatives data", *IFC Report*, no 7, October.

——— (2018b): "INEXDA – the granular data network", *IFC Working Papers*, no 18, October.

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

——— (2021a): "Micro data for the macro world", *IFC Bulletin*, no 53, April.

——— (2021b): "Issues in data governance", *IFC Bulletin*, no 54, July.

——— (2021c): "Use of big data sources and applications at central banks", *IFC Report*, no 13, February.

——— (2023a): "Post-pandemic landscape for central bank statistics", *IFC Bulletin*, no 58, June.

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

——— (2023c): "Data science in central banking: applications and tools", *IFC Bulletin*, no 59, October.

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

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

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

——— (2025b): "Cloud services and AI in central banks: balancing benefits and risks", *IFC Report*, no 18, Box C, February.

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

Israël, J-M and B Tissot (2021): "Incorporating micro data into macro policy decision-making", *IFC Bulletin*, no 53, April.

Jordon, J, L Szpruch, F Houssiau, M Bottarelli, G Cherubin, C Maple, S Cohen and A Weller (2022): "Synthetic data – what, why and how?", *ArXiv*.

Kroese, B (2024): "Exploring official statistics using generative AI and SDMX", presentation at the 72nd plenary session of the Conference of European Statisticians, Geneva, 20–21 June.

Lambe, E and T Park (2024): "The BIS Data Portal project – delivering the next generation platform for BIS statistics", *IFC Bulletin*, no 60, April.

Liu, F, K Ting and Z-H Zhou (2008): "Isolation forest", *Eighth IEEE International Conference on Data Mining*.

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

Mittal, M, L Goyal, D Hermanth and J Sethi (2019): "Clustering approaches for high-dimensional databases: a review", *WIREs Data Mining and Knowledge Discovery*, vol 9, no 3, e1300, January.

Nikoloutsos, S and O Sirello (2023): "SDMX, an international standard for micro data", presentation at the SDMX Global Conference 2023, November.

Organisation for Economic Co-operation and Development (OECD) (2019): *Enhancing access to and sharing of data: reconciling risks and benefits for data re-use across societies*, OECD Publishing, Paris.

——— (2023): "Emerging privacy-enhancing technologies: current regulatory and policy approaches", *OECD Digital Economy Papers*, no 351, March.

Ravn, L (2025): "The fabrication of synthetic data promises: tracing emerging arenas of expectations and boundary work", *Big Data & Society*, vol 12, no 1, January.

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, March.

Schwalbach, J and R Maurer (2025): "Sharing digital trace data: researchers' challenges and needs", *Big Data & Society*, vol 12, no 1, March.

Tice, G, M Nelson, S Nikoloutsos and X Sosnovsky (2025): "SDMX metadata-driven approach: the BIS experience", Box C, *IFC Report*, no 17, February, p 11.

United Nations (UN) (2014): "Resolution adopted by the General Assembly on 29 January 2014: 68/261 – Fundamental Principles of Official Statistics", January.

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

——— (2022): *Machine learning for official statistics*, March.

——— (2023a): "Large language models for official statistics", *High-Level Group for the Modernisation of Official Statistics White Paper*, December.

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

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

——— (2024b): *Organisational aspects of implementing ML based data editing in statistical production*, High-Level Group for the Modernisation of Official Statistics, September.

——— (2024c): "INGEST Task Force on Standards Issues: a path towards the use of common standards to support the integration of geospatial and statistical information", *Working Paper Series on Statistics*, no 11, August.

——— (2024d): *Data stewardship and the role of national statistical offices in the new data ecosystem*, ECE/CES/STAT/2023/4, April.

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

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

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

Wilkinson, M et al (2016): "The FAIR guiding principles for scientific data management and stewardship", *Scientific Data*, vol 3, 160018, March.