

Doll, Hendrik Christian; Alves Werb, Gabriela

Article — Published Version

Innovation for improving climate-related data—Lessons learned from setting up a data hub

AStA Wirtschafts- und Sozialstatistisches Archiv

Provided in Cooperation with:

Springer Nature

Suggested Citation: Doll, Hendrik Christian; Alves Werb, Gabriela (2023) : Innovation for improving climate-related data—Lessons learned from setting up a data hub, AStA Wirtschafts- und Sozialstatistisches Archiv, ISSN 1863-8163, Springer, Berlin, Heidelberg, Vol. 17, Iss. 3, pp. 355-380, <https://doi.org/10.1007/s11943-023-00326-w>

This Version is available at:

<https://hdl.handle.net/10419/308872>

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.



<https://creativecommons.org/licenses/by/4.0/>



Innovation for improving climate-related data—Lessons learned from setting up a data hub

Hendrik Christian Doll  · Gabriela Alves Werb

Received: 31 March 2023 / Accepted: 16 August 2023 / Published online: 26 October 2023
© The Author(s) 2023

Abstract In this article, we present a framework to assess the challenges in the climate-related data landscape. From our perspective, we describe challenges and opportunities for innovation to close existing gaps in selected data quality dimensions for central banks and statistical offices. We further examine the role of networks in promoting innovation and international collaboration, highlighting practical experiences with a case study of the Sustainable Finance Data Hub at the Deutsche Bundesbank. Finally, we discuss how these lessons can provide avenues to enhance data quality in central banks and official statistics and outline directions for future research.

Keywords Climate risks · Climate-related data · Data gaps · Hub and spoke · Sustainable finance · Data integration

JEL-Classification C81, E59, Q38

1 Introduction

As climate change grows into one of the most urgent and complex challenges of our time, central banks increasingly rely on climate-related data to support decision-

All views expressed in this work are personal views of the authors and do not necessarily reflect the views of the Deutsche Bundesbank or the Eurosystem.

✉ Hendrik Christian Doll
Sustainable Finance Data Hub, Deutsche Bundesbank, Frankfurt, Germany
E-Mail: hendrik.doll@bundesbank.de

Gabriela Alves Werb
Data Service Centre, Deutsche Bundesbank, Frankfurt, Germany
Frankfurt University of Applied Sciences, Frankfurt, Germany

making across their core functions (Campiglio et al. 2018). Central banks use data on these risks to support climate stress tests and disclosures, analyses for banking supervision, financial stability, monetary policy, and statistical purposes (Deutsche Bundesbank 2022b; ECB 2022a, b, 2023; FED 2021).

Similarly, statistical offices face the growing need to provide climate-related data and indicators. More than 60 heads of statistical offices recently defined a set of climate-related indicators to be established at the national level (UNECE 2021a). The United Nations has also promoted similar initiatives to support national statistical offices in this undertaking (United Nations Economic and Social Council 2022). These data are instrumental in producing a comprehensive and internationally comparable set of climate-related indicators and statistics to support policy decisions, foster public discourse, and monitor progress toward fulfilling the Sustainable Development Goals (SDG) or the goals of the Paris Agreement (Destatis 2022; Eurostat 2021).

Consequently, central banks and statistical offices typically require granular data on physical and transition risks to support climate-related analyses and to develop indicators. Transition risks relate to the uncertainty in the shift to a less carbon-intensive economy, which requires societies to reshape their production methods and reduce their reliance on carbon-based assets. Physical risks arise from adverse climate events and extreme weather, such as floods, storms, wildfires, or droughts, as well as disruptions in ecosystems related to climate change.

Despite the undisputed importance of data for assessing climate risks, there are still substantial gaps in their availability, reliability, and comparability (NGFS 2022b). In the absence of unified sources for high-quality data, the current climate-related data landscape is characterized by several fragmented sources, including administrative, commercial, and publicly available data in different frequencies, formats, and units of observation. Each source alone provides a specific subset of measurements and observations with somewhat limited coverage. Therefore, harmonizing these data to leverage them jointly is one of the crucial challenges ahead.

The path forward requires improvements in data collection efforts, which initiatives and legislation have supported in several jurisdictions. In Europe, the Corporate Sustainability Reporting Directive (CSRD) aims to expand the universe of reporting firms and increase the comparability of climate-related information disclosed at the firm level (CSRD 2021). However, the CSRD should only take full effect in 2027, when small and medium enterprises must also report relative to the previous financial year. Furthermore, the Regulation (EU) 2019/2088, also known as the Sustainable Finance Disclosure Regulation (SFDR), and the revision of Pillar 3 disclosures to include climate-related risks (EBA 2022) are expected to improve the quality and structure of disclosed climate-related information in the financial services sector.

There is also an increasing number of joint initiatives from several international institutions. One example is the effort towards a new Data Gaps Initiative (DGI) developed by the Inter-Agency Group on Economic and Financial Statistics (IAG) in joint work with the Financial Stability Board (FSB). In this high-level work plan, climate-related data are one of the four priority areas (IMF 2022). Nevertheless, the recommendations are expected to be implemented within five years of the work plan's launch, so observing their first effects will take a few years. In the meantime,

given the urgency with which climate risks may materialize, there is a growing need for alternative pathways to address these gaps (IPCC, 2022).

In this paper, we outline a framework to assess the current challenges to incorporating climate-related data into central banks' and statistical offices' decision-making and discuss how these create opportunities for innovation. We describe how to improve the quality of these data by leveraging novel data sources and methods to enhance their coverage, timeliness, accessibility, reliability, and integration. Then, we examine the role of collaboration networks in advancing the climate-related data landscape, considering different cooperation models and highlighting the experiences of the Sustainable Finance Data Hub at the Deutsche Bundesbank. Finally, we discuss how these lessons generalize, show avenues to enhance data quality in official statistics, and outline directions for future research.

2 Current challenges in the climate-related data landscape

The current gaps in the climate-related data landscape present several challenges for central banks and statistical offices to base their analyses, disclosures, and decisions on reproducible and reliable information (NGFS 2022b). Following a user-centric approach (Wang and Strong 1996), we present a framework to classify and assess these challenges across several dimensions. By doing so, we aim to capture the main aspects of data quality from the perspective of consumers of climate-related data in central banks and statistical offices. We link current challenges to the FAIR data principles, i.e., findability, accessibility, inter-operability, and re-usability (Wilkinson et al. 2016).

We start by discussing the typical shortcomings of relying solely on administrative and commercial data. While leveraging alternative sources for structured and unstructured data can contribute to addressing them, doing so brings another set of challenges, which we examine subsequently. Figure 1 provides an integrated overview of the current challenges associated with the climate-related data landscape.

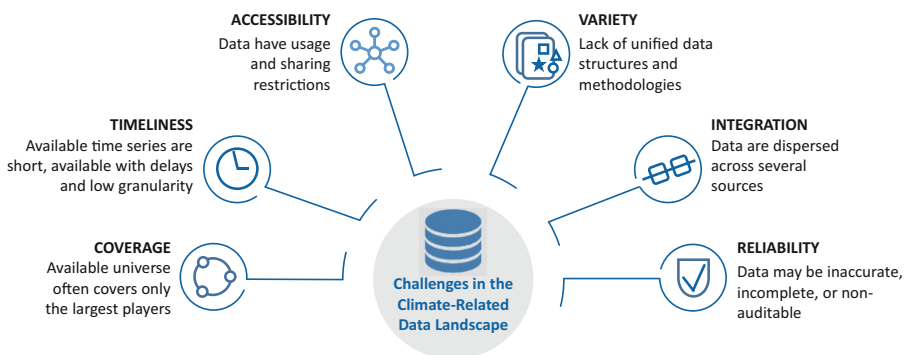


Fig. 1 Proposed Framework to Classify and Assess Challenges in the Current Climate-Related Data Landscape

2.1 Data coverage and timeliness of information remain a challenge

Data coverage limitations are one of the most salient barriers to comprehensive climate-related analyses. Administrative data only cover limited sets of indicators and firms, as policymakers must often make compromises when attempting to reduce the reporting burden for the affected stakeholders. Similarly, other policy-related data sources, such as the European Union Emissions Trading System (EU ETS), only cover firms in specific sectors, such as aviation and selected energy-intensive manufacturing sectors. Novel reporting frameworks such as the CSRD in Europe should substantially expand the number of reporting firms in selected jurisdictions. However, the resulting universe will still mainly comprise the largest firms, as measured by balance sheet totals, net sales, and the number of employees.

Against this scenario, it is not uncommon for central banks to source proprietary data from commercial providers. For example, the European System of Central Banks (ESCB) recently announced the joint procurement of climate-related data from multiple providers, equipping member central banks with broader indicators for assessing physical and transition risks (Deutsche Bundesbank 2022a). In official statistics, a similar need is increasingly recognized for incorporating proprietary data (e.g., Eurostat 2022; UN 2017). Nevertheless, proprietary data from commercial providers also have substantial coverage drawbacks, as they typically only focus on large corporations in Europe and North America. Smaller firms that are regionally strong or even internationally highly successful and might therefore be systemically relevant for climate-related analyses are typically not represented in the universe provided by these data.

Furthermore, additional issues characterize the market for proprietary data, such as long-run dependencies, the potential of increasing prices due to monopolies, and a lack of published methodological and quality frameworks (Eurostat 2022). When relying on climate-related data from proprietary sources, these issues are to an extent mitigated by competition among data providers with similar offers regarding transition and physical risk data. Also the multi-source approach is an attempt to reduce over-reliance on any one source (Deutsche Bundesbank 2022a).

As the reporting frameworks for climate-related data are relatively new, only a few firms disclosed a narrow set of indicators on an infrequent basis in the past decades before being legally required to do so. The essential balance between the data needs and the reporting burden for affected firms also culminated in reduced data granularity and reporting frequency in administrative data. Consequently, the available time series are still relatively sparse and short for most firms, sectors, and indicators. When aiming for FAIR data, these issues represent a lack of (re-)usability, especially taken together with changing disclosure formats.

Furthermore, firms typically only report climate-related data yearly for the previous financial year, in a process that may take several months after the respective financial year-end. Consequently, data timeliness is also a significant concern, as there may be a substantial lag between when firms or policymakers take a given action and the time in which the most current information about the respective indicators is available. In practice, this temporal mismatch prevents the timely incorporation of recent developments into decision-making. As the effects of interest

might also have complex dynamics, aggregating the data at the year level potentially further diminishes their analytical potential.

Because commercial data providers also rely to some extent on voluntary disclosures from firms, they also suffer from this shortcoming. As these providers also require time to process the information contained in the disclosures before incorporating it into their proprietary data products, in their case, timeliness is typically an even more prominent issue. It is, therefore, not uncommon for the most current proprietary data to have a gap of two years or more.

2.2 Climate data come in a variety of formats and underlie restricted access

Another barrier central banks and statistical offices face when handling administrative data is that these often underlie a strict access control regime within the institution. Depending on the legal basis for obtaining the data from the reporting parties, data access might be restricted to the department responsible for the assessment or supervision, which precludes, for example, data reuse for statistical purposes. Furthermore, indicators of interest or their proxies may be dispersed across multiple reporting frameworks and institutions. In this case, institutions face even more significant legal, organizational, and technical obstacles to accessing and sharing these data.

While commercial data typically do not underlie these same constraints, the licensing agreement terms with the commercial data provider set the boundaries for using and sharing them. The license may refer to a user, department, or institution and typically does not allow sharing of the data partly or in its entirety with third parties, even if in a summarized or aggregated manner. These access restrictions exist despite the often publicly available nature of climate-related data at their original source and can potentially hinder data accessibility, which is also a relevant aspect in the FAIR data framework.

For example, the Research Data and Service Centre of Deutsche Bundesbank provides access to Bundesbank microdata for internal and external researchers. Currently, administrative microdata is available for banks, securities, firms, and households. Linking these data with the commercial climate-related data that is internally available would provide a great analytical potential. However, this possibility cannot currently be extended to external researchers due to the aforementioned licensing restrictions.

Furthermore, the long-standing lack of unified standards aligned with the methodological and conceptual differences in recent reporting frameworks leads to relatively low comparability of climate-related data across different sources. One notable example is the case of Scope 3 greenhouse gas emissions, which denote the emissions measured throughout the firm's value chain except those originating from the firm's assets or energy consumption. They include upstream and downstream activities, such as those associated with the product's supply chain, usage, and end-of-life treatment (Greenhouse Gas Protocol 2011).

As Ducoulombier (2021) notes, existing reporting practices do not support cross-firm comparisons. First, each firm may rely on a different mixture of primary and secondary data to estimate the emissions in its value chain. Second, the allocation

of the emissions across the different suppliers, inputs, and outputs will invariably follow different assumptions and methodologies, rendering the resulting estimates hardly comparable. Third, firms may change their data sources, assumptions, or methodologies across reporting periods, impairing comparisons for the same firm over time. Taken together, these shortcomings represent the often prevalent lack of inter-operability in the climate-related data domain, an aspect the FAIR framework strives for.

This issue becomes more pronounced in the case of commercial data because commercial data providers often rely on proprietary algorithms to estimate missing data points and calculate aggregated scores. However, they typically do not disclose the exact variables and weights driving these calculations. Empirical data exhibit substantial variation in the climate-related indicators from different providers for the same firm-year pairs, suggesting that each data provider relies on a different, undisclosed set of assumptions and methodologies. These characteristics obstruct the usage of these data for reproducible analyses and research. Finally, the variety of formats and sources and the prevalence of restrict access regimes often render the FAIR dimension of findability difficult to apply for climate-related data.

2.3 In the absence of unified data, integration and reliability are essential

There is a demand for promoting mutual understanding and transparency in the assumptions and methodological decisions in the current variety of reported and estimated data. Despite the international efforts in different jurisdictions, such as the European Taxonomy Regulation (Regulation (EU) 2020/852) or the steps from the International Sustainability Standards Board towards establishing global standards (IFRS 2022), more dialogue is required to break silos across countries and disciplines.

However, it is noteworthy that such processes are typically lengthy. One example is the international discussion about general data quality standards, which started decades ago with several supporting countries. Despite all efforts, the ISO 8000 standard was only published more than 20 years later and remains a work in progress with many remaining disputes (Cai and Zhu 2015).

Luckily, we may tackle some of the challenges currently associated with administrative and commercial climate-related data by relying on the vast amount of information available from alternative data sources. These may include documents and reports from firms, press releases, spreadsheets with crowdsourced data, user search behavior, user-generated content, or satellite images, among many others. Nevertheless, due to the wide variety of these data, which often include several data types and complex data structures, leveraging them brings additional obstacles.

For example, unstructured data require advanced data extraction, processing, and modeling methods. Furthermore, handling the complexity of such varied data involves orchestrating a portfolio of structured, semi-structured, and unstructured data products from different sources. This paradigm requires institutions to move away from data silos and establish appropriate data lakes, supporting a richer constellation of data types and structures. However, central banks and statistical offices must use a proper data governance framework to accompany this process. Otherwise, data

lakes can quickly become “data swamps” (Hai et al. 2016), and data will not be usable.

Furthermore, tackling the difficulties in integrating data from a potentially large number of sources requires a suitable metadata management concept, which provides the necessary context and semantic information that allows users to establish connections between the different data sources, types, and units of observation. These aspects are especially crucial in this context due to the lack of unique identifiers to integrate the data across all sources.

Finally, while leveraging several data sources should deliver complementary information and foster a more comprehensive perspective of climate-related risks, doing so also presents the risk of facing multiple instances of inaccurate or incomplete data. Central banks and statistical offices must therefore develop methodologies to handle situations of conflicting information, which may rely on algorithms that assess the reliability of individual sources and establish a consensus mechanism. Such a consensus-driven approach is especially relevant in this context, given that the data, in many cases, might be non-verifiable or non-auditable by third parties. Furthermore, this approach fosters the FAIR dimensions of data inter-operability and re-usability.

These numerous obstacles in extracting, integrating, and sharing climate-related data represent an unparalleled chance to foster innovative approaches. In the next section, we provide an overview of how these challenges present opportunities for interdisciplinary innovation. Furthermore, we discuss the potential of alternative data sources and novel methods to close gaps in selected dimensions of climate-related data quality.

3 A roadmap for innovation in climate-related data

The outlined challenges for climate-related data map out areas with crucial need for innovation. In this section, we describe promising avenues to increase data coverage and improve data access and integration by relying on unstructured data, alternative data sources, and machine learning methods. Doing so supports FAIR climate-related data, particularly in the dimensions accessibility, inter-operability and (re-)usability.

3.1 Increasing data coverage and timeliness by leveraging unstructured data

One promising avenue to improve the coverage of climate-related data is to draw on publicly available data, which may come from several sources, such as firms, (inter)governmental agencies, civil society organizations, or individual citizens. These data have the advantage of being easily accessible and not being subject to usage restrictions, which enables interested stakeholders to acquire, merge, and share them.

For example, many firms disclose information about the exposure of their assets to climate risks, the environmental impact of their activities, and their efforts to reduce their environmental footprint. Nevertheless, this information is typically unstructured and dispersed across multiple sources, such as annual, integrated, or sustainability

reports, investor presentations, dedicated websites, press releases, or social media (Alves Werb and Doll 2023).

Often, these sources contain data in multiple modalities. For example, sustainability reports typically combine the information in text, tables, and images containing charts in a single document. Because they potentially carry additional or complementary information, it would be beneficial to acquire all of them. However, automatically identifying and extracting these different modalities in the reports is challenging because they are typically not disclosed in a machine-readable format. Instead, most firm disclosures are in PDF format, following different layouts (e.g., single, double, or multi-column).

Furthermore, contrary to other types of annual firm disclosures, sustainability reports have no standard document sections. Consequently, there is substantial variation in the information presented and its sequence, both within sustainability reports of a single firm and across reports of different firms. These aspects represent obstacles to adopting a standard approach to extract each data format from an extensive collection of sustainability reports. When handling similar use cases, we have observed that open-source or proprietary solutions that deliver good results for standardized documents typically do not perform well or do so only for a limited subset of reports.

Possible solutions include training a new supervised model to identify and extract text, tables, and charts from sustainability reports. However, this solution requires large volumes of annotated data, which is, in practice, a resource-intensive undertaking. Another alternative is to leverage the power of transfer learning by relying on pre-trained models. In the domain of deep learning, transformer-based architectures offer state-of-the-art models to extract one or more modalities and fine-tune them to the particular use case of sustainability reports. One example is to fine-tune pre-trained transformer models to identify topics in the text contained in sustainability reports.

However, instead of separately addressing each modality, we foresee a promising avenue in jointly extracting and leveraging the multiple data modalities embedded in sustainability reports, in a process known as multimodal deep learning. Such models can harness the potential of several modalities and are at the core of the current developments in deep learning. Consequently, we argue that there is a high, yet unexplored potential, to apply them in the context of climate-related data, particularly with sustainability reports.

3.2 Improving data accessibility with alternative data

In recent years, an increasing body of literature in economics has drawn on alternative data to improve inflation, economic activity, labor market outcomes, or financial risk models (Beręsewicz 2017; Froidevaux et al. 2022; Kapetanios and Papailias 2018). In this context, alternative data refers to data originally collected by institutions or individuals for other purposes than that of the analysis, such as credit card spending, search engine queries, or webscraped price data from online shops. Several studies rely on combining traditional statistics with alternative data to fill the gaps in series with a lower frequency, such as yearly or quarterly.

Central banks and statistical offices also followed this movement and have progressively incorporated alternative data in their research and analyses (Chen et al. 2022). As doing so results in a challenge for traditional statistical methods, these applications typically rely on non-parametric machine learning methods as a robust alternative to jointly leveraging these data.

Nevertheless, central banks often rely on aggregated risk indicators from commercial providers in their analyses involving physical and transition risks. One possible reason is that such indicators consolidate several dimensions and abstract from a few methodological challenges while typically providing identifiers that facilitate the integration with administrative data. However, these data often preclude a comprehensive analysis of the different climate-related aspects and their possible interplay with the economic variables or units of observation of interest. Furthermore, due to their proprietary nature, they are typically subject to strict usage and sharing restrictions.

In the context of climate-related data, alternative data for central banks and statistical offices may include structured ocean, weather, and atmospheric data that are frequently publicly available from environmental, space, and other local and inter-governmental agencies at a jurisdictional or geographic level. Meteorologists and climate scientists use these data extensively for weather forecasting and complex climate modeling. Given their substantial analytical value, we expect that leveraging them will enable more robust models and provide novel insights. Potential use cases include estimates of asset or property insurance coverage and their potential gaps, estimates of the impact of extreme weather on labor supply and productivity, and forecasts of government spending for disaster response, among several others.

Furthermore, numerous international efforts, such as those spearheaded by the European Space Agency (ESA) or the National Aeronautics and Space Administration (NASA), produce an increasing volume of geospatial data, such as satellite imagery and remote sensing data from specialized instruments. These data are available in an open-source format and represent a substantial breakthrough in computing indicators for changes in vegetation, water bodies, erosion, or desertification, as well as in detecting expansion and potential vulnerabilities in agricultural or urban areas. In addition, they enable assessments of the potential or actual impact of natural hazards, such as wildfires, floods, or storms.

The solutions from international collaborations keep growing with new data products, such as the Multi-Mission Algorithm and Analysis Platform (MAAP) for biomass, which integrates data from ESA and NASA missions (Albinet et al. 2019). In contrast to administrative or commercial data, these platforms offer invaluable resources that central banks and statistical offices can leverage, integrate, and share with little to no confidentiality restrictions.

Furthermore, anonymized location-based data, such as mobile phone data, can provide real-time and more accurate information about mobility choices in each administrative region and provide a valuable resource for studying commuting effects (Daas et al. 2011). It is also possible to use it to enrich the existing data to estimate the transition risk due to urban car dependency. These data have a much higher frequency than vehicle registrations and are less costly to obtain than data about individual mobility preferences, which typically come from surveys.

Similarly, we identify a unique window of opportunity to reuse administrative, proprietary, and crowdsourced data in novel ways to increase our understanding of climate-related risks. There are many use cases for improving their usability to enrich existing data sources with additional information on the exposure of assets, firms, or administrative regions to transition and physical risks.

However, we must address the methodological and integration challenges to explore this avenue. Given the multiple data sources, remote sensing parameters, instruments, and resolutions in which the data are available, it is essential to understand how these aspects and their interplay impact the subsequent inferences (e.g., Rashkovetsky et al. 2021; Tavakkoli Piralilou et al. 2022). There has also been an active discussion in the economics community on how to deal with methodological challenges. One example is the advance in accounting for the role of extreme events when incorporating granular ocean, weather, and atmospheric data into economic models (Auffhammer 2018; BIS 2021; Pindyck 2017).

Alternative data also require a solid understanding of the phenomena in their generating process. While efficient metadata supports users in understanding a few data properties, these schemas cannot represent many essential considerations and specific domain knowledge. For example, the data may follow non-traditional sampling or be influenced by other non-trivial factors. Addressing these challenges is crucial to improve alternative data's usability and unlock their promising analytical potential.

To do so, we require close collaboration with the disciplines that study these phenomena to understand the data's representativeness and what inferences they allow. In addition, there are increasing efforts to leverage machine-learning methods for data integration and knowledge representation with encouraging results. We discuss these aspects in detail in the next section.

3.3 Enhancing data integration and reliability with machine learning

In the previous sections, we built upon the vast potential of unstructured data and alternative data sources to discuss avenues to improve selected quality dimensions in climate-related data. However, as the number of explored data sources, formats, units of observation, and frequencies grows, it becomes increasingly difficult to integrate and jointly leverage these data.

One of the first steps is to conceptualize and implement a software solution to orchestrate the data, document its lineage, and make it available for subsequent analyses. To do so, we also require a methodology to integrate the data models, which may include establishing relationships among the multiple measurement units, frequencies, and units of observation present in all data sources. For example, data may refer to a project, an industrial facility, a firm's activities in a particular geographic area, or a geographic coordinate. Consequently, if we are to use these different aspects to measure a firm's exposure to climate-related risks, it is crucial to define how to represent them jointly.

The next challenge is record linkage, that is, to identify the same entity or a group of related entities (e.g., ownership structures, branches, or subsidiaries) across multiple data sources. In the field of information retrieval, researchers also refer to this

problem as entity resolution. The absence of unique common identifiers to do so is a frequent problem faced by central banks and statistical offices when handling multiple sources of survey and administrative data. Historically, they have tackled this problem with a mixture of deterministic and probabilistic approaches (Bakker et al. 2014; Gessendorfer et al. 2018; Schnell 2010). However, there has been increased interest in machine learning methods to address data quality issues in official statistics (UNECE 2021b).

Existing literature suggests that supervised learning methods applied to record linkage provide superior results than those from unsupervised methods, such as K-means (e.g., Christen 2012). Recent advances include combining graph convolutional networks and siamese networks to leverage the relationships and contextual information in knowledge graphs (Krivosheev et al. 2021). Nevertheless, supervised approaches require a substantial volume of high-quality training data, which might be difficult and too expensive to obtain in practice, as the data distribution is highly unbalanced towards negative (non-matches) pairs.

In the past years, there have been many attempts to circumvent this issue, most of which relying on neural networks (Barlaug and Gulla 2021). For example, Ebraheem et al. (2018) use pre-trained word embeddings to train recurrent neural networks that require less training data, whereas Kasai et al. (2019) propose a mixture of transfer learning and active learning. In their approach, the neural network learns from similar domains with large volumes of labeled data. Then, it relies on user feedback for a limited number of informative samples to fine-tune the transferred model to the field of interest.

On the other hand, Wu et al. (2020) propose an unsupervised method based on Gaussian Mixture Models that reaches performance levels comparable to some supervised approaches. We sustain that applications for climate-related data should build on these recent developments, as labeled training data is typically unavailable for applications in this domain. Another advantage is that their implementations are typically easily accessible with publicly available source codes and documentation in dedicated libraries (e.g., Hosseini et al. 2020).

Furthermore, we argue that novel machine learning applications in this context should increasingly focus on the fusion of several data sources and modalities (e.g., sensor, satellite images, textual, or tabular) to deliver more robust inferences and increase data reliability. This paradigm shift should enable us to take full advantage of the extensive landscape of climate-related data available. One example would be to jointly leverage historical tabular data on local flood incidence and severity, current weather data, textual descriptions of a firm's reforestation efforts, and satellite images to monitor the risk of floods in a firm's primary production location.

In the context of multimodal learning, neural networks became popular for their performance and suitability for multiple modeling purposes (Baltrušaitis et al. 2018). For example, it is possible to train the network to perform a specific task, such as predicting the risk of floods. Alternatively, we may rely on an unsupervised approach, such as with an autoencoder architecture, to provide an abstract data representation. The resulting model can later be used to find similar observations using a siamese architecture or as a first step to fine-tune the result for a specific prediction task.

More recent research has also shed light on the effect of linkage errors on the results of subsequent statistical analyses based on the matched data (Chambers and Diniz da Silva 2020; Han and Lahiri 2019). Particularly relevant are scenarios in which the probability of obtaining a match correlates with an unobserved variable that is potentially associated with the variables of interest.

Given the typically unknown sampling and complexity of many alternative data sources (Beręsewicz 2017) and potential biases in the representativeness of the administrative and commercial data (Meyer and Mittag 2021), the interplay between machine learning and data quality remains an active field for further research in several disciplines (Ilyas and Rekatsinas 2022). Consequently, we argue that major developments in this research frontier will require intensive collaboration across interdisciplinary research communities and institutions. In the next section, we present a framework for developing a collaborative climate-related data network.

4 Collaborating within and across organizations

A large body of literature investigates innovation in organizations, which has remained a topic of high interest to management and strategy researchers over the decades. Previous research indicates that corporate culture, technology, and the external market environment are important determinants of organizations' innovation paths and success (e.g., Kitchell 1995; Tellis et al. 2009). However, Damanpour (1996) finds in a meta-analysis with more than 20 studies that their effect depends significantly on how we define and measure these constructs.

One topic of particular interest is how collaboration networks can foster innovation within and across organizations. While the importance of networks is indisputable, there is much debate about their optimal configuration. Baum et al. (2010) find that heterogeneity and complementarity in the knowledge background of the network members promote a more productive network. Other studies suggest that networks should avoid geographic concentration in a single country and foster "structural holes" that avoid static ego networks and allow for exchange with novel partners (e.g., Kumar and Zaheer 2018).

In our domain of interest, central banks have a track record of decades of international collaboration with well-established global networks. Furthermore, given their dual role in generating and consuming climate-related data, they have a great incentive to drive innovation to improve its quality and availability (Artman et al. 2023; Rosolia et al. 2021). However, because of this endeavor's novel challenges, it is necessary to intensify global collaboration among central banks and extend it to other stakeholders involved in climate-related data, such as academic institutions, non-governmental organizations, and commercial data providers (IFC 2021; NGFS 2022b).

Because tackling the challenges with climate-related data requires joining efforts from several disciplines and perspectives, there is a promising potential for co-innovation among central banks, academic, and industry partners (NGFS 2020). This scenario yields a complex landscape where many central banks strive to align with each other and other partners. Nevertheless, similar cases of successful international

collaborations between the industry and academic institutions in other disciplines can provide valuable insights and best practices for central banks' efforts (Owen-Smith et al. 2002; Rybníček and Königsgruber 2019). Many central banks also keep close academic ties for joint research, which they can further extend for other projects.

In a similar context, a recent survey from the World Bank about experiences of international innovation activities in banking supervision examined different organizational setups for combining the expertise of supervisors and data scientists (World Bank 2021). As banking supervision standards are often developed globally—e.g., with the Basel Committee on Banking Supervision and the Financial Stability Board—and national banking supervisors tend to be closely interwoven, particularly within Europe's Single Supervisory Mechanism (SSM), this domain offers incentives to collaborate internationally. The survey identified three main prevalent approaches to organizing such innovation units.

One approach is a centralized data science or innovation unit, where all data scientists work together in the same department to innovate and improve analytics across the organization. This model is also known as a "center of excellence". While this configuration enables dedicated teams to focus on selected innovation initiatives, the independence from business units may limit the contact with users and their demands.

An alternative is decentralizing data science teams and placing them across different business areas. The goal is to promote collaboration between business experts and data scientists for a strong user focus. However, this setup poses challenges for innovation flows, as it may foster silos between departments, reducing the opportunity for knowledge sharing.

Finally, another possibility is the hub and spoke model, which combines elements of centralized and decentralized teams. A centralized business unit acts as a hub coordinating with experts in all business areas in this setup. The goal is to enable information flows across the different teams. The hub and spoke model can mediate complex interactions across stakeholders, increasing value in larger organizations.

The hub and spoke network model first became popular in logistics organizations, such as package delivery and airline transportation, and originally referred to a system in which goods flow between two points via a central node. This setup contrasts with point-to-point networks, where goods are sent directly between two points (Alderighi et al. 2007). Because of the fewer connections, this model reduces complexity while ensuring an efficient network. When n is the number of dots in a network, linking all dots requires $n - 1$ connections. In contrast, a point-to-point network requires $(n - 1) / 2$ connections to achieve the same result.

The hub and spoke model gained traction in recent years to describe the flow of information in networks and orient organizational units' setup as hubs. Longo et al. (2013) describe this model's application for virtual communities, where firm-based hubs connect to spokes in external web-based communities.

Several organizations report applying a hub and spoke setup in the context of banking supervision to facilitate innovation with a mixture of centralized teams and interdisciplinary teams (World Bank 2021). The SSM in Europe also offers examples of best practice for joint innovation with several countries and stakeholders. In

the complex context of the European banking supervision, a hub connects all stakeholders working on innovation projects in the SSM, identifying synergies and needs across them. The hub structures the work in agile teams composed of members of the European Central Bank (ECB) and staff from national central banks (McCaul 2022).

In the climate-related data domain with this organizational pattern the national climate-related data centers are the hubs and the various in-house and external climate-related data stakeholders are the spokes. Furthermore, there is a second layer of hub and spoke collaboration in international cooperation, as outlined in Fig. 2. In this layer, dedicated organizations represent the international hubs and the national central banks are the spokes.

This organizational hub and spoke setup seems to be gaining traction. For example, the ECB founded the Climate Change Center in 2021, combining expertise from diverse business areas in one hub (Lagarde 2021). Similarly, the Banca d'Italia's Climate Change and Sustainability Hub is mandated to enable analyses and facilitate information flows among participants in national and international groups (Banca d'Italia 2022).

In an example of collaboration with academia, the Monetary Authority of Singapore has three hubs that foster collaboration with academic institutions on sustainable finance—one of which aims to acquire and translate granular climate-related data to measure firms' sustainability performance (MAS 2022).

Internationally, central banks launched a series of projects and working groups to jointly explore innovation, such as the Network for Greening the Financial System (NGFS). In the NGFS, 121 central banks and financial supervisors currently collaborate in several working groups to accelerate and spread best practices related to measuring and monitoring climate risks. It has quickly become an important forum

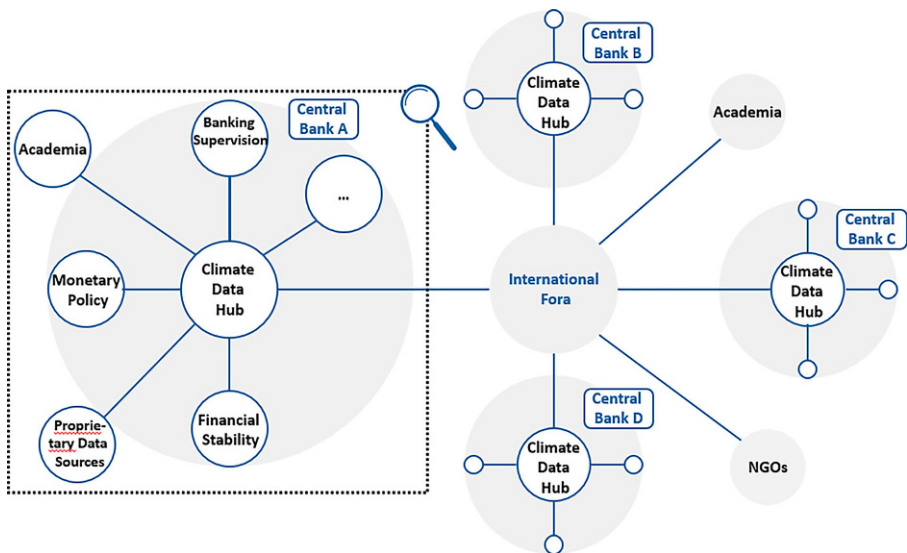


Fig. 2 The Two-Layer Hub and Spoke Model of Collaboration on Climate-Related Innovation Activities

since its launch in 2017, raising attention to data gaps (NGFS 2022b) and working towards improvements within its expert networks.

This multi-layered setup is reflected in the NGFS' self-description as a global focal point for central banks and supervisors to address climate-related and environmental risks with a strong focus on climate-related data (NGFS 2022a). Similarly, the Bank for International Settlements' (BIS) Innovation Hub aspires to be a global network of central bank experts on innovation. It recently stated green finance as one of four focus areas (BIS 2022a). In this existing network, the BIS increasingly launches projects focused on climate-related issues. The BIS Innovation Hub currently has strong a pipeline of several projects that aim to improve climate-related analyses and data availability, such as Project Viridis (BIS 2022c), and to develop digital solutions for trading green bonds, such as Project Genesis (BIS 2022b).

These networks trigger collaborations from which the entire ecosystem benefits, as many developed frameworks and methods are transferrable to other substantive problems. Statistical offices with strong international networks and established collaborations have similar opportunities to share and develop methods and best practices with regards to climate-related data (UNECE 2018).

This organizational structure efficiently integrates domain and data science knowledge, providing an exchange forum to disseminate best practices and drive innovation in climate-related data with internal and external partners. In the next section, we present a detailed discussion with a case study of one such data hub, namely the setup and experiences of the Sustainable Finance Data Hub at the Deutsche Bundesbank.

5 Case study of the Sustainable Finance Data Hub

To support climate-related analysis and foster innovation toward closing existing data gaps, the Deutsche Bundesbank set up the Sustainable Finance Data Hub (SFDH) in 2020 to serve user needs for the whole bank (Fehr et al. 2022). The SFDH is a specialized unit within the Data Service Centre of the Directorate General (DG) Statistics that provides a data central access point and serves as a center of competence for methodological questions. It aims to promote in-house data availability, enable transparency to the public, engage in central bank cooperation, and close data gaps through novel data generation in innovation projects.

As outlined in Fig. 3, the SFDH focuses on the three core pillars to achieve these goals: data, innovation, and exchange. By leveraging innovative methods and novel data sources and joining forces through an increased national and international exchange, it focuses on improving the availability and usability of climate-related data. Fehr et al. (2021) provide a detailed description of this approach. Subsequently, we outline the activities in its three pillars.

5.1 Providing high-quality climate-related data

The SFDH acts as a focal point to provide climate-related data for all business areas and related networks. To ensure user-centric data provision and generation, it reaches

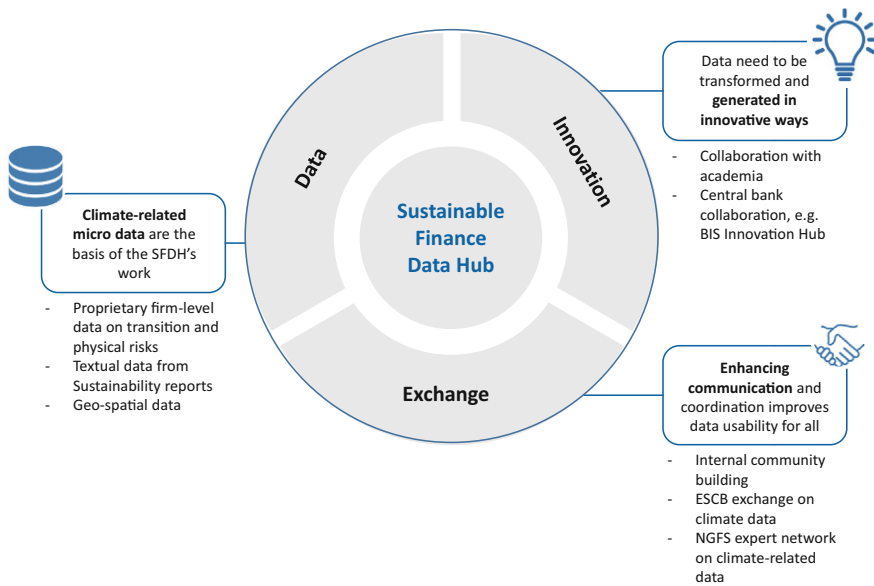


Fig. 3 Core Tasks of the Bundesbank's Sustainable Finance Data Hub

out to users bi-yearly in user group meetings to assess their needs. A regular survey further complements this assessment. Thereby, the SFDH obtains information about data usage and current demands—e.g., which data packages have a higher demand and which information might still be missing. The SFDH further engages with multiple stakeholders to procure data and curate innovative solutions to address the priority needs.

Recently, the Bundesbank led the joint Eurosystem procurement for two climate-related data providers that are now available for all ESCB Central Banks under the same commercial conditions (Deutsche Bundesbank 2022a). As a result, all participating central banks can easily leverage the same granular climate-related data, improving the comparability of their analyses and intensifying collaboration and exchange.

When making these data available, the SFDH is committed to the FAIR data principles. Therefore, after concluding the procurement phase, it is essential to ensure inter-operability by integrating the commercial data with other administrative data, as most analyses rely on merging climate-related data to instrument-, firm-, or bank-level data. Consequently, the SFDH provides integration solutions to almost all internal administrative data by applying record linkage using supervised machine learning (Doll et al. 2021). Applying high-quality string-based record linkage with machine learning algorithms delivers immediate user value by improving the data usability in several contexts. The resulting data are available to all internal users in traditional file structures and in the Bundesbank internal Hadoop cluster.

Furthermore, the established methods and pipelines provide a solid basis for collecting and using novel data from unstructured sources. To this end, innovation activities provide a comprehensive set of published textual corporate sustainability

Table 1 Available and Expected Data Sets for Internal Users in the SFDH (as of 2023)

Name	Description	Status
Firm-level proprietary data	Proprietary data from two providers on transition risks and physical risk at the asset- and firm-level in multiple packages	Available
Textual data	Corporate sustainability reports in textual form of large global corporations for several years	Available
Geo-spatial proprietary data	Proprietary data on physical risks at the geo-spatial level	Ongoing provisioning
New administrative data	Firm-level data that will be available through regulation, such as the CSRD (phased availability from 2025 onwards)	Scheduled changes in regulation
Enhanced administrative data	Enhanced quality checks of available data through data validation using image data, such as from satellites and street view	Foreseen through academic collaboration projects
New structured data from unstructured sources	Indicators extracted from available textual data	Foreseen through innovation in international central bank collaboration projects

reports over multiple years and firms. This data set is available to users interested in exploring analyses based on unstructured data. As the data needs continue evolving, the SFDH has an ongoing pipeline, including the procurement of additional climate-related data packages scheduled for 2023. Furthermore, it aims to continuously provide novel data generated by innovation activities—a list of available and expected data for internal users is provided Table 1.

Finally, to provide easily findable and inter-operable data sets for research and analyses, the SFDH harmonizes the available data in a content-wise manner and documents the data comprehensively following a structured metadata schema from the Data Documentation Initiative (Vardigan et al. 2008). The easy onboarding and data provision processes foster internal data accessibility, whereas novel data generation is expected to provide additional value and enhance data accessibility for external users.

5.2 Enhancing data usability through coordination

To promote public transparency and facilitate the comprehension of ongoing developments, the SFDH makes information available to a broader audience through dashboards and selected data sets on the Bundesbank's statistics website. This is an area in which statistical offices and central banks increasingly provide immediate value to society (IMF 2023; Townsend 2021). Furthermore, the procured granular climate-related data support the Bundesbank's public climate reporting, both at a national and European level, in collaboration with the ECB (e.g., Deutsche Bundesbank 2022b; ECB 2023).

The SFDH also actively reaches out to internal and external stakeholders to foster collaboration within its network. For building internal community within the Bundesbank, the SFDH acts as a central hub providing various fora to connect interested

stakeholders and spread best practices in handling climate-related data. Activities include the “Climate Data Consultation Hours”, organizing the “Sustainable Finance Community”, engaging in the “Sustainable Finance Expert Group” and discussing user needs in the user group, among others. This approach represents the internal hub and spoke layer within the Bundesbank’s network, where the SFDH is the hub and the business areas are the spokes.

The “Climate Data Consultation Hours” are a bi-weekly platform hosted by the SFDH where stakeholders from all business areas can obtain information about climate-related data access. It is a low-threshold format to raise data-related questions, connect with other users that may have faced similar questions, and provide feedback. This format fosters ongoing user engagement and enables the SFDH to tailor its activities in a user-centric manner. Furthermore, it catalyzes concrete data quality improvements, such as correcting inconsistencies and providing novel ideas to improve existing data products.

This forum is complemented by the “Sustainable Finance Community”, a monthly forum organized by the SFDH in which climate data experts from the DG Statistics connect, present their work and discuss new developments. Similarly, the SFDH organizes “Sustainable Finance Expert Group” to connect stakeholders across all DGs and ensure efficient information flows among data providers and users.

These internal community-building activities aim to ensure a user-centric focus in all SFDH’s developments (feedback from users to the hub), to share information about new developments (hub to users), and among users about ongoing projects (users to users). In an area with abundant data gaps and fast momentum, such as climate-related data, this format streamlines information flows and efficiently allocates scarce expert resources.

Externally, the SFDH contributes to and organizes exchange fora at the European and international levels to share and benefit from best practices. For example, it acts as the current secretariat of an exchange network for ESCB central banks to collaborate and enhance the analytical value of the jointly procured climate-related data. Other examples are international networks such as the NGFS or the BIS Innovation Hub.

The SFDH actively contributes to NGFS working groups and currently chairs the NGFS expert’s network on climate-related data. The NGFS expert’s network on climate-related data aims to consolidate the NGFS community of statisticians and data scientists to foster a regular dialogue on climate data-related topics. The work plan also envisions exchanging experiences and best practices for setting up climate-related data hubs. This organizational engagement is complemented by a Bundesbank board member who is the current vice chair of the entire NGFS and is scheduled to be the chair in 2024 (Deutsche Bundesbank 2022c).

5.3 Leveraging innovation to enhance climate-related data

In the BIS Innovation Hub, central banks collaborate to develop joint initiatives. In joint work with the Banco de España and the ECB, the SFDH successfully pitched the Project Gaia to the new Eurosystem Centre of the BIS Innovation Hub (BIS 2023). This project aims to increase the transparency of climate-related disclosures

by exploring natural language processing, optical character recognition, and machine learning to extract relevant metrics from publicly available sustainability reports. The project aspires to provide an open-source database of firm sustainability reports with a full-text search engine to allow analysts to identify sustainability-related disclosures in an intuitive user interface to facilitate climate risk assessments.

The project establishes a sandbox to apply novel methods for improving the quality of climate-related data by building an integrated pipeline to web scrape reports, extracting relevant indicators from text, tables, and charts, and providing a user interface for analysis. Consequently, Project Gaia represents a concrete innovation step toward our call for increasing data coverage by leveraging unstructured data and improving their usability.

The internal and external networks and projects build the basis to foster collaborative innovation, thereby supporting the SFDH's third pillar, innovation. This setup represents the external hub and spoke layer, where the international fora are the hubs and the SFDH participates as the spoke. This additional layer enables the SFDH to share the Bundesbank's best practices across the network and to act as a disseminator, circulating international best practices to internal stakeholders.

University collaborations and internal coding challenges further complement the innovation activities at the SFDH. These joint projects aim to explore advanced methods to extract relevant information from unstructured data sources. In a planned project with the Technical University of Darmstadt, the SFDH seeks to validate firm-level structured master data with image data from satellites and street view. The goal is to apply multimodal learning methods to validate contextual information in administrative data, such as sector classification or addresses. While this is a more general use case, there is a close collaboration with the SFDH to transfer the knowledge and experiences with multimodal learning to improve climate-related data.

Furthermore, in a project with the Frankfurt University of Applied Sciences, the SFDH explores novel methods to extract information about firms' environmental, social, and governance (ESG) performance from text and tables in corporate sustainability reports. To foster transparency and promote data democratization to a broader group of interested stakeholders, this project also includes a searchable repository of reports in PDF format and a publicly available dashboard to visualize the extracted indicators.

Overall, the call from central bank users for high-quality, easy-to-use climate data shapes the SFDH's role in promoting innovation within the different layers of its network. The following section discusses how these experiences and best practices may transfer to other organizations that leverage climate-related and other administrative data.

6 Lessons learned for central banks and official statistics

The setup of the Bundesbank's SFDH provided several lessons learned. First, it underlined climate-related data's potential to serve as a sandbox for using alternative data sources and machine learning methods, showcasing the possibilities to leverage these innovations to improve several dimensions of data quality. Second, it

outlined the potential to transfer novel methods and practices to other administrative data collections. Third, it demonstrated that combining data domain knowledge and innovation in the institutional setup is vital to curate high-quality data and enable knowledge sharing. In this section, we outline these lessons in greater detail.

Climate-related data is an excellent catalyst for collaboration because many institutions strive to acquire the same global data, given its global public good nature. To counter the challenges the current climate-related data landscape presents, the SFDH has made substantial progress by combining existing administrative data sources with commercial data and fostering innovation projects that leverage alternative data sources and novel methods. Nevertheless, we see this undertaking as a continuous process that involves assimilating best practices and lessons learned to tackle current and future challenges.

Its often relatively low confidentiality facilitates knowledge sharing for climate-related data at the granular level, which is not always possible with other types of administrative data. However, the collaboration experience and trust developed within the networks established to promote climate-related data provide a solid foundation for future collaboration in other use cases.

Despite the different constraints imposed by other settings, transferring these lessons learned and further innovations to other types of administrative data is crucial, as their quality can highly benefit from incorporating alternative data sources. This opportunity is especially present when one or more of the following conditions hold: (i) alternative data sources exist to improve data coverage, (ii) alternative data sources exist to validate existing indicators, or (iii) alternative data sources exist to obtain higher frequency or more granular data.

This potential has also been recognized by the central banks' statistics community (e.g., Rosolia et al. 2021) and by statistical offices (Salgado and Oancea 2020; UNECE 2021b). Furthermore, there is also an excellent, partly unexplored potential to transfer the application of machine learning to improve the quality of many statistical data collections.

In the context of official statistics, the United Nations Economic Commission for Europe (UNECE) provides an illustrative example of collaboration within statistical offices to leverage alternative data sources for migration statistics, given the limitations of available surveys and existing administrative data. They provide examples of projects integrating mobile phone, credit card, and social network data to enhance the existing data and reduce issues of coverage and accuracy (UNECE 2022).

Furthermore, while setting up the SFDH, we learned that combining data and innovation knowledge in a central hub is vital for efficient knowledge sharing. Consequently, the choice of organizational setup is an important success factor. This organizational setup is conceptionally known as a hub and spoke model. In its inner layer, the hub is the SFDH and the spokes are the internal business areas working with climate-related data. In the outer layer, international fora and working groups are the hubs and the SFDH acts as the spoke.

This configuration also fosters innovation in other domains, especially in more extensive networks, when the number of involved institutions and stakeholders becomes large. In the context of central banks, the European SSM, with its many

national members, provides a natural setting to apply the hub and spoke setup for other innovation initiatives.

Statistical offices have long established similarly large multi-layered networks for close collaboration, where efficient information flows are paramount. In the case of Germany, its 14 state statistical offices with one additional federal statistical office represent one collaboration layer. This layer is complemented by a layer of international collaboration, such as within the European Statistical System (ESS), the United Nations Statistics Division (UNSD), or the UNECE.

As part of this hub and spoke approach, user-centric innovation remains an ongoing priority. Based on the SFDH's experiences, it is paramount to reach out and leverage users' domain knowledge and ideas in all business areas. To this end, the SFDH runs regular surveys and user workshops to collect data needs, usage profiles, and novel ideas for further innovation. All innovation projects start with a design thinking workshop to collect user pain points and suggestions. This approach stimulates stakeholder engagement and generalizes to innovation activities in other contexts.

These lessons apply to other innovation activities across central banks and statistical offices. A broad stakeholder buy-in fosters idea generation, user-centric development, and timely incorporation of the results into production. Ensuring efficient information flows enhances the speed at which participants disseminate best practices throughout the network, within and across organizations. In our experience, a hub and spoke setup provides a fruitful and successful setting to do so.

Moreover, internal management support is a crucial factor in the setup of the SFDH that arguably extends to other settings. Broad support is critical to rely on resources from a wide range of stakeholders and to promote collaboration with various business areas. This support is also important in setting up international cooperation layers. In the case of Bundesbank, the close involvement of a board member in networks as the NGFS and its president pointing to the need to devote more attention to climate-related matters at his inaugural speech paved the way for the related initiatives (Deutsche Bundesbank 2022d). The urgency to increase the availability of climate-related data has been a crucial driver for a fast setup of hubs, innovation projects, and collaborations over the past years.

7 Conclusion

This paper provides a comprehensive review of the challenges the current climate-related data landscape brings to central banks and statistical offices. We describe potential avenues to leverage alternative data sources and novel methods to close gaps in selected dimensions of climate-related data quality. We describe a framework for developing collaborative climate-related data networks and proceed with a case study of the setup of the Sustainable Finance Data Hub at the Deutsche Bundesbank. Finally, we outline lessons learned and best practices to enable innovation and foster quality for climate-related and other types of administrative data.

By doing so, we aim to provide a foundation for central banks and statistical offices to respond by forming robust collaboration networks, apply suitable orga-

nizational changes, and optimally leverage innovation initiatives to close existing data gaps. This step is crucial to enable multiple stakeholders to capture the range of climate risks and estimate their consequences throughout the economy, fostering informed decisions.

Given the pressing need for high-quality data to foster informed analyses and the existing challenges in the climate-related data landscape, there is significant momentum in innovation and collaboration activities. Several institutions and stakeholders are joining efforts to leverage alternative data sources and novel methods to close the gaps in climate-related data, especially data coverage, timeliness, accessibility, reliability, and integration. Central banks have a unique perspective on user needs since they produce and consume climate-related data.

From our perspective, several key organizational factors are crucial for success in this context. Internal collaboration across business areas and external reach across peer organizations, institutional backing, and efficient information flow are essential in fostering innovation and disseminating best practices.

Given their influence and potential to bridge multiple stakeholders' needs and contributions, we believe that central banks can substantively influence the climate-related data landscape and innovation in data extraction, integration, and dissemination. We hope that the discussion in this article can contribute to fostering the exchange between researchers and practitioners on avenues to promote access to high-quality climate-related data and can inform the setup of dedicated organizational units. Furthermore, we foresee a great potential to apply the methods and lessons learned in the context of climate-related data to other types of administrative data.

Acknowledgements We thank Stefan Bender and Christine Schlitzer for valuable comments on an early version of this paper.

Funding Open Access funding enabled and organized by Projekt DEAL.

Conflict of interest H.C. Doll and G. Alves Werb declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- (2021) Corporate Sustainability Reporting Directive. Procedure (EU) 2021/0104/COD. <https://eur-lex.europa.eu/legal-content/EN/HIS/?uri=CELEX%3A52021PC0189>. Accessed 2023-03-18
- Albinet C, Whitehurst AS, Jewell LA, Bugbee K, Laur H, Murphy KJ, Duncanson L (2019) A joint ESA-NASA multi-mission algorithm and analysis platform (MAAP) for biomass, NISAR, and GEDI. *Surv Geophys* 40(4):1017–1027

- Alderighi M, Cento A, Nijkamp P, Rietveld P (2007) Assessment of new hub-and-spoke and point-to-point airline network configurations. *Transp Rev* 27(5):529–549
- Alves Werb G, Doll HC (2023) Climate Data Integration for Central Banking and Supervision. Working Paper
- Artman M, Quang BP, Demski J, Esham N, Izzati N, Kling L, Yang H (2023) Sustainable finance statistics: progress, challenges and leveraging digital tools. *J Digit Bank* 7(3):198–214
- Auffhammer M (2018) Quantifying economic damages from climate change. *J Econ Perspect* 32(4):33–52
- Bakker BFM, van Rooijen J, van Toor L (2014) The system of social statistical datasets of statistics Netherlands: an integral approach to the production of register-based social statistics. *Stat J IAOS* 30:411–424
- Baltrušaitis T, Ahuja C, Morency L-P (2018) Challenges and applications in multimodal machine learning. In: *The handbook of multimodal-multisensor interfaces: signal processing, architectures, and detection of emotion and cognition*, vol 2. Association for Computing Machinery and Morgan & Claypool, pp 17–48
- Banca d'Italia (2022) The bank of Italy Steps up its action on climate risks and sustainable finance. Press release. <https://www.bancaditalia.it/media/notizia/the-bank-of-italy-steps-up-its-action-on-climate-risks-and-sustainable-finance/?dotcache=refresh>. Accessed 2023-02-10
- Barlaug N, Gulla JA (2021) Neural networks for entity matching: a survey. *ACM Trans Knowl Discov Data* 15(3):Article 52
- Baum JAC, Cowan R, Jonard N (2010) Network-independent partner selection and the evolution of innovation networks. *Manage Sci* 56(11):2094–2110
- Beręsewicz M (2017) A two-step procedure to measure representativeness of Internet data sources. *Int Stat Rev* 85(3):473–493
- BIS (2021) Climate-related financial risks—measurement methodologies. <https://www.bis.org/bcbis/publ/d518.pdf>. Accessed 2023-02-15
- BIS (2022a) BIS innovation hub to focus on CBDC, payments, deFi and green finance in 2022 work programme. Press release. <https://www.bis.org/press/p220125.htm>. Accessed 2023-03-05
- BIS (2022b) Genesis 2.0: smart contract-based carbon credits attached to green bonds. https://www.bis.org/about/bisih/topics/green_finance/genesis_2.htm. Accessed 2023-02-18
- BIS (2022c) Project Viridis: a climate risk platform for financial authorities. https://www.bis.org/about/bisih/topics/green_finance/viridis.htm. Accessed 2023-02-18
- BIS (2023) BIS innovation hub centre—Eurosysteem. <https://www.bis.org/about/bisih/locations/eurosystem.htm>. Accessed 2023-03-20
- Cai L, Zhu Y (2015) The challenges of data quality and data quality assessment in the big data era. *Data Sci J* 14(0):2
- Campiglio E, Dafermos Y, Monnin P, Ryan-Collins J, Schotten G, Tanaka M (2018) Climate change challenges for central banks and financial regulators. *Nature Clim Change* 8(6):462–468
- Chambers R, Diniz da Silva A (2020) Improved secondary analysis of linked data: a framework and an illustration. *J Royal Stat Soc Ser A Stat Soc* 183(1):37–59
- Chen JC, Dunn A, Hood K, Driessen A, Batch A (2022) Off to the races: a comparison of machine learning and alternative data for predicting economic indicators. In: Abraham KG, Jarmin RS, Moyer B, Shapiro MD (eds) *Big data for twenty first century economic statistics*. University of Chicago Press, pp 373–402
- Christen P (2012) *Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection*. Springer, Berlin, Heidelberg
- Daas PJH, Roos M, De Blois C, Hoekstra R, Ten Bosch O, Ma Y (2011) New data sources for statistics: experiences at statistics Netherlands. aper presented at the 2011 European New Technique and Technologies for Statistics Conference, Brussels
- Damanpour F (1996) Organizational complexity and innovation: developing and testing multiple contingency models. *Manage Sci* 42(5):693–716
- Destatis (2022) Potential greenhouse effect of fluorinated greenhouse gases used in 2021 down 9 % on 2020. Press release. https://www.destatis.de/EN/Press/2022/12/PE22_524_32421.html. Accessed 2023-03-21
- Deutsche Bundesbank (2022a) Climate-Related Data Successfully Procured. Press Release. <https://www.bundesbank.de/en/press/press-releases/climate-related-data-successfully-secured-869246>. Accessed 2022-12-09
- Deutsche Bundesbank (2022b) Climate-Related Disclosures by the Deutsche Bundesbank 2022. <https://www.bundesbank.de/resource/blob/838206/df8916fd80f723164101dd53d2768324/mL/2022-klima-bericht-data.pdf>. Accessed 2023-03-21

- Deutsche Bundesbank (2022c) Mauderer Appointed as Vice-Chair of the NGFS. Press Release. <https://www.bundesbank.de/en/tasks/topics/mauderer-appointed-as-vice-chair-of-the-ngfs-883878>. Accessed 2023-02-18
- Deutsche Bundesbank (2022d) Speech at the Ceremony to Mark the Inauguration of the New President of the Deutsche Bundesbank. Press Release. <https://www.bundesbank.de/en/press/speeches/speech-883330>. Accessed 2023-02-15
- Doll H, Gábor-Tóth E, Schild CJ (2021) Linking Deutsche Bundesbank Company Data. Deutsche Bundesbank Technical Report 2021-05. <https://www.bundesbank.de/resource/blob/624432/207c774d468e82d76ec19ef6bfa1c8a7/mL/2021-05-company-data.pdf>. Accessed 2023-01-15
- Ducoulombier F (2021) Understanding the importance of scope 3 emissions and the implications of data limitations. *J Impact ESG Invest* 1(4):63
- EBA (2022) Final Draft Implementing Technical Standards on Prudential Disclosures on ESG Risks in Accordance with Article 449a CRR. https://www.eba.europa.eu/sites/default/documents/files/document_library/Publications/Draft%20Technical%20Standards/2022/1026171/EBA%20draft%20ITS%20on%20Pillar%203%20disclosures%20on%20ESG%20risks.pdf. Accessed 2023-03-10
- Ebraheem M, Thirumuruganathan S, Joty S, Ouzzani M, Tang N (2018) Distributed representations of tuples for entity resolution. *Proc VLDB Endow* 11(11):1454–1467
- ECB (2022a) 2022 climate risk stress test. https://www.bankingsupervision.europa.eu/ecb/pub/pdf/ssm.climate_stress_test_report.20220708~2e3cc0999f.en.pdf. Accessed 2022-12-29
- ECB (2022b) ECB takes further steps to incorporate climate change into its monetary policy operations. <https://www.ecb.europa.eu/press/pr/date/2022/html/ecb.pr220704~4f48a72462.en.html>. Accessed 2023-03-21
- ECB (2023) Towards Climate-Related Statistical Indicators. https://www.ecb.europa.eu/pub/pdf/other/ecb.climate_change_indicators202301~47c4bbbc92.en.pdf. Accessed 2023-03-21
- United Nations Economic and Social Council (2022) Climate Change Statistics. Report of the Secretary General. <https://unstats.un.org/unsd/statcom/53rd-session/documents/2022-17-ClimateChangeStats-E.pdf>. Accessed 2023-03-21
- Eurostat (2021) Sustainable development in the European Union. Monitoring report on progress towards the SDGs in an EU context. <https://ec.europa.eu/eurostat/documents/15234730/15241861/KS-03-21-096-EN-N.pdf/e8413539-6c6a-131f-45d7-6af7178a0b67?t=1667395592318>. Accessed 2023-03-21
- Eurostat (2022) Empowering Society by Reusing Privately Held Data for Official Statistics—A European Approach. Final Report Prepared by the High-Level Expert Group on Facilitating the Use of New Data Sources for Official Statistics. <https://ec.europa.eu/eurostat/documents/7870049/14803739/KS-FT-22-004-EN-N.pdf/052b4357-bf8e-9ce4-c063-7e806c045dac?t=1656335798606>. Accessed 2023-03-15
- FED (2021) Climate Change and Financial Stability. <https://www.federalreserve.gov/econres/notes/feds-notes/climate-change-and-financial-stability-20210319.html>. Accessed 2023-03-21
- Fehr M, Gies K, Schlitzer C, Triebkorn E, Wengerek S (2021) Short and medium-term approaches to closing climate data gaps: German experiences and conclusions on the way forward. Paper presented at the 9th IMF Statistical Forum: Measuring Climate Change: The Economic and Financial Dimensions.
- Fehr M, Schlitzer C, Triebkorn E (2022) The Bundesbank's Sustainable Finance Data Hub. IFC Bulletins Chapters, 56. <https://www.bis.org/ifc/publ/ifcb56.pdf>. Accessed 2023-02-10
- Froidevaux A, Macalos J, Khalfoun I, Defrasnes M, d'Orsetti S, Salez N, Sciberras A (2022) Leveraging alternative data sources for socio-economic Nowcasting. Paper presented at the Proceedings of the 2022 ACM Conference on Information Technology for Social Good., Limassol. <https://doi.org/10.1145/3524458.3547253>
- Gessendorfer J, Beste J, Drechsler J, Sakshaug JW (2018) Statistical matching as a supplement to record linkage: a valuable method to tackle nonconsent bias? *J Off Stat* 34(4):909–933
- Greenhouse Gas Protocol (2011) Corporate Value Chain (Scope 3) Accounting and Reporting Standard: Supplement to the GHG Protocol Corporate Accounting and Reporting Standard. World Resources Institute and World Business Council for Sustainable Development. https://ghgprotocol.org/sites/default/files/standards/Corporate-Value-Chain-Accounting-Reporting-Standard_041613_2.pdf. Accessed 2023-02-10
- Hai R, Geisler S, Quix C (2016) Constance: an intelligent data lake system. Paper presented at the Proceedings of the 2016 International Conference on Management of Data, San Francisco
- Han Y, Lahiri P (2019) Statistical analysis with linked data. *Int Stat Rev* 87(S1):S139–S157
- Hosseini K, Nanni F, Ardanuy CM (2020) DeezyMatch: a flexible deep learning approach to fuzzy string matching

- IFC (2021) Sustainable finance data for central banks. IFC report no 14. https://www.bis.org/ifc/publ/ifc_report_14.pdf. Accessed 2023-01-23
- IFRS (2022) Exposure draft—IFRS S2 climate-related disclosures. <https://www.ifrs.org/content/dam/ifrs/project/climate-related-disclosures/issb-exposure-draft-2022-2-climate-related-disclosures.pdf>. Accessed 2023-03-27
- Ilyas IF, Rekatsinas T (2022) Machine learning and data cleaning: which serves the other? *J Data Inf Qual* 14(3):Article 13
- IMF (2022) G20 leaders welcome new data gaps initiative to address climate change, inclusion and financial innovation. Press release. <https://www.imf.org/en/News/Articles/2022/11/28/pr22410-g20-leaders-welcome-ndgi-to-address-climate-change-inclusion-financial-innovation>. Accessed 2023-03-17
- IMF (2023) Climate change indicators dashboard. <https://climatedata.imf.org/>. Accessed 2023-03-01
- IPCC (2022) Climate Change 2022: Impacts, Adaptation, and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. [Pörtner H-O, Roberts DC, Poloczanska ES, Mintenbeck K, Tignor M, Alegría A, Craig M, Langsdorf S, Löschke S, Möller V, Okem A (eds)]. Cambridge University Press
- Kapetanios G, Papailias F (2018) Big data & macroeconomic nowcasting: methodological review. Discussion Papers from Economic Statistics Centre of Excellence, ESCoE., DP-2018-2012
- Kasai J, Qian K, Gurajada S, Li Y, Popa L (2019) Low-resource deep entity resolution with transfer and active learning. arXiv preprint arXiv:1906.08042
- Kitchell S (1995) Corporate culture, environmental adaptation, and innovation adoption: a qualitative/quantitative approach. *J Acad Mark Sci* 23(3):195–205
- Krivoshchev E, Atzeni M, Mirylenka K, Scotton P, Miksovic C, Zorin A (2021) Business entity matching with siamese graph convolutional networks. *Proc AAAI Conf Artif Intell* 35(18):16054–16056
- Kumar P, Zaheer A (2018) Ego-network stability and innovation in alliances. *Acad Manag J* 62(3):691–716
- Lagarde C (2021) Climate change and central banking. In: Dombret A, Kenadjian PS (eds) *Green banking and green central banking*. De Gruyter, pp 151–158
- Longo MC, Giaccone SC, Garraffo F (2013) Applying the hub-and-spoke model to virtual communities: the IBM innovation approach. *Int J Technol Mark* 8(2):142–158
- MAS (2022) Monetary Authority of Singapore Sustainability Report 2021 / 2022. https://www.mas.gov.sg/-/media/mas-media-library/publications/sustainability-report/2022/mas-sustainability-report-2021_2022.pdf. Accessed 2023-01-22
- McCauley E (2022) The Impact of Suptech on European Banking Supervision. Speech at the Supervision Innovators Conference 2022. <https://www.bankingsupervision.europa.eu/press/speeches/date/2022/html/ssm.sp220914~d0201e42a9.en.html>. Accessed 2023-01-18
- Meyer BD, Mittag N (2021) Combining administrative and survey data to improve income measurement. In: *Administrative records for survey methodology*, pp 297–322
- NGFS (2020) Guide to Climate Scenario Analysis for Central Banks and Supervisors. NGFS Technical Document. <https://www.bundesbank.de/resource/blob/835286/e54fbbd96ecfc052505eeda7ba5f243/mL/guide-to-climate-scenario-analysis-data.pdf>. Accessed 2023-01-23
- NGFS (2022a) Annual Report 2021. https://www.ngfs.net/sites/default/files/medias/documents/ngfs_annual_report_2021.pdf. Accessed 2023-01-20
- NGFS (2022b) Final Report on Bridging Data Gaps. https://www.ngfs.net/sites/default/files/medias/documents/final_report_on_bridging_data_gaps.pdf. Accessed 2022-12-29
- Owen-Smith J, Riccaboni M, Pammolli F, Powell WW (2002) A comparison of U.S. and European university-industry relations in the life sciences. *Manage Sci* 48(1):24–43
- Pindyck RS (2017) The use and misuse of models for climate policy. *Rev Environ Econ Policy* 11(1):100–114
- Rashkovetsky D, Mauracher F, Langer M, Schmitt M (2021) Wildfire detection from multisensor satellite imagery using deep semantic segmentation. *IEEE J Sel Top Appl Earth Observations Remote Sensing* 14:7001–7016
- Rosolia A, Stapel-Weber S, Tissot B (2021) New developments in central bank statistics around the world. *Stat J IAOS* 37:1055–1060
- Rybníček R, Königsguber R (2019) What makes industry—university collaboration succeed? A systematic review of the literature. *J Bus Econ* 89(2):221–250
- Salgado D, Oancea B (2020) On new data sources for the production of official statistics. Statistics Spain Working Papers 01/2020. <https://doi.org/10.48550/arXiv.2003.06797>

- Schnell R (2010) Record linkage from a technical point of view. In: German Data Forum (RatSWD) (ed) Building on progress: expanding the research infrastructure for the social, economic, and behavioral sciences, 1st edn. Barbara Budrich, pp 531–546
- Tavakkoli Piralilou S, Einali G, Ghorbanzadeh O, Nachappa TG, Gholamnia K, Blaschke T, Ghamisi P (2022) A Google earth engine approach for wildfire susceptibility prediction fusion with remote sensing data of different spatial resolutions. *Remote Sens* 14(3):672
- Tellis GJ, Prabhu JC, Chandy RK (2009) Radical innovation across nations: the preeminence of corporate culture. *J Mark* 73(1):3–23
- Townsend I (2021) New UK Climate Change Portal to Inform Understanding and Decision-Making. <https://blog.ons.gov.uk/2021/10/28/new-uk-climate-change-portal-to-inform-understanding-and-decision-making/>. Accessed 2023-01-20
- UN (2017) Recommendations for Access to Proprietary Data. Global Working Group on Big Data for Official Statistics. <https://unstats.un.org/unsd/statcom/48th-session/documents/BG-3d-recommendations-for-access-to-proprietary-data-E.pdf>. Accessed 2023-03-20
- UNECE (2018) Recommendations for Promoting, Measuring and Communicating the Value of Official Statistics. <https://unece.org/fileadmin/DAM/stats/publications/2018/ECECESSTAT20182.pdf>. Accessed 2023-03-14
- UNECE (2021a) Conference of European Statisticians' Set of Core Climate Change-related Indicators and Statistics Using the System of Environmental-Economic Accounting. https://unece.org/sites/default/files/2021-08/CES_Set_Core_CCR_Indicators-Report.pdf. Accessed 2023-03-17
- UNECE (2021b) Machine Learning for Official Statistics. <https://unece.org/sites/default/files/2022-01/ECECESSTAT20216.pdf>. Accessed 2023-03-17
- UNECE (2022) Use of New Data Sources for Measuring International Migration. https://unece.org/sites/default/files/2023-01/2225454_E_Final%20version%20Rev1.pdf. Accessed 2023-03-20
- Vardigan M, Heus P, Thomas W (2008) Data documentation initiative: toward a standard for the social sciences. *Int J Digit Curation*. <https://doi.org/10.2218/ijdc.v3i1.45>
- Wang RY, Strong DM (1996) Beyond accuracy: what data quality means to data consumers. *J Manag Inf Syst* 12(4):5–33
- Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Mons B (2016) The FAIR guiding principles for scientific data management and stewardship. *Sci Data* 3(1):160018
- World Bank (2021) The Next Wave of Suptech Innovation—Suptech Solutions for Market Conduct Supervision. Technical Note. <https://documents1.worldbank.org/curated/en/735871616428497205/pdf/The-Next-Wave-of-Suptech-Innovation-Suptech-Solutions-for-Market-Conduct-Supervision.pdf>. Accessed 2023-01-10
- Wu R, Chaba S, Sawlani S, Chu X, Thirumuruganathan S (2020) ZeroER: Entity Resolution using Zero Labeled Examples. Paper presented at the Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data, Portland, OR, USA

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.