Irving Fisher Committee on
Central Bank Statistics

BIS

12th biennial IFC Conference: "Statistics and beyond: new data for decision making in central banks"

22-23 August 2024

The climate data iceberg – a depth of information to integrate[1]

Hendrik Christian Doll, Emily Kormanyos, Susanne Walter and Gabriela Alves Werb,
Deutsche Bundesbank

[1] This contribution was prepared for the conference. The views expressed in this publication are those of the authors and do not necessarily represent the official views of the Committee, its members, or the BIS.

The climate data iceberg – A depth of information to integrate

Hendrik Christian Doll[1], Emily Kormanyos[1], Susanne Walter[2], Gabriela Alves Werb[2,3]

Abstract

Central banks need climate-related data to align evidence-based climate change considerations with their core tasks. While structured data from administrative and proprietary sources are limited and contain considerable gaps, a wealth of climate-related information is dispersed and lies below the surface in unstructured form, such as sustainability reports or satellite images. To characterise this situation, we introduce the image of the climate data iceberg. Information from unstructured sources can bridge current data gaps and enhance the usability of existing data by improving its accuracy, extending its scope, and reducing data sharing barriers. In this paper, we discuss the challenges and opportunities central banks and supervisors face in leveraging this unstructured information for climate analysis and research. We further investigate how innovative efforts between central banks and other institutions can help generate actionable and usable climate-related data, exemplified by our own experiences and early-stage learnings from such collaborations.

Keywords: Data gaps, sustainable finance data, climate data, satellite data, sustainability reports, natural language processing, multimodal learning.

JEL classification: E58, Q56, C81.

[1] Deutsche Bundesbank, Sustainable Finance Data Hub

[2] Deutsche Bundesbank, Data Service Centre

[3] Frankfurt University of Applied Sciences

1. Introduction

Given the growing concerns about climate change's impact on the financial system, there is a growing argument for incorporating climate risks into central banks' mandates, particularly in the context of stress tests (Battiston et al., 2017; Fabris, 2020; Monasterolo, 2020; Schellhorn, 2020). However, central banks face challenges due to fragmented climate-related data, which include substantial gaps, such as in their availability, reliability, and comparability (NGFS, 2022; Nightingale et al., 2019; Schmieder et al., 2021).

As a reaction, policymakers are moving to standardize climate-related data. Internationally, the Global Reporting Initiative set guidelines for sustainability reporting (de Villiers et al., 2022), paving the way for the standards proposed by the International Sustainability Standards Board (IFRS, 2023). In the U.S., the Securities and Exchange Commission has introduced climate disclosure rules (SEC, 2024). In the European Union, initiatives include a sustainability taxonomy, guidelines for sustainability-related financial products, and the Corporate Sustainability Reporting Directive (CSRD), which standardizes data reporting in a machine-readable format (EU, 2019, 2020, 2022, 2023).

However, data availability will take time to mature, especially for smaller firms that are for example exempt from reporting under the CSRD until 2028, limiting data depth for comprehensive analyses. To address the current data gaps, central banks increasingly procure proprietary climate data (Deutsche Bundesbank, 2022). However, licensing often restricts data sharing and transparency in decision-making due to the proprietary nature of the proprietary algorithms.

This paper discusses the potential of unstructured, public data sources, and provides an overview of Deutsche Bundesbank's initiatives and collaborative efforts to foster innovation to close existing gaps in climate-related data.

2. Leveraging unstructured data sources to close data gaps

Traditionally, information to support applications in the public sector, industry, and academia stems from structured commercial and administrative data sources, which typically also include data estimated with proprietary algorithms. However, as outlined in Figure 1, these data sources represent only a small fraction of the current climate-related data information available, as the tip of an iceberg.

A vast amount of climate-related information is publicly available (Alonso-Robisco et al., 2024; Nightingale et al., 2019), but often unstructured and spread across sources like corporate sustainability reports, social media posts, newspaper articles, investor presentations, and investment fund legal documents. Aerial, satellite as well as remote-sensing data also represent a potentially rich source of information to assess physical risks and supply-chain risks.

This widely dispersed climate-related data is what we refer to as the climate data iceberg which, though less visible, holds significant potential to close data gaps, especially with advances in text extraction and image recognition technology.

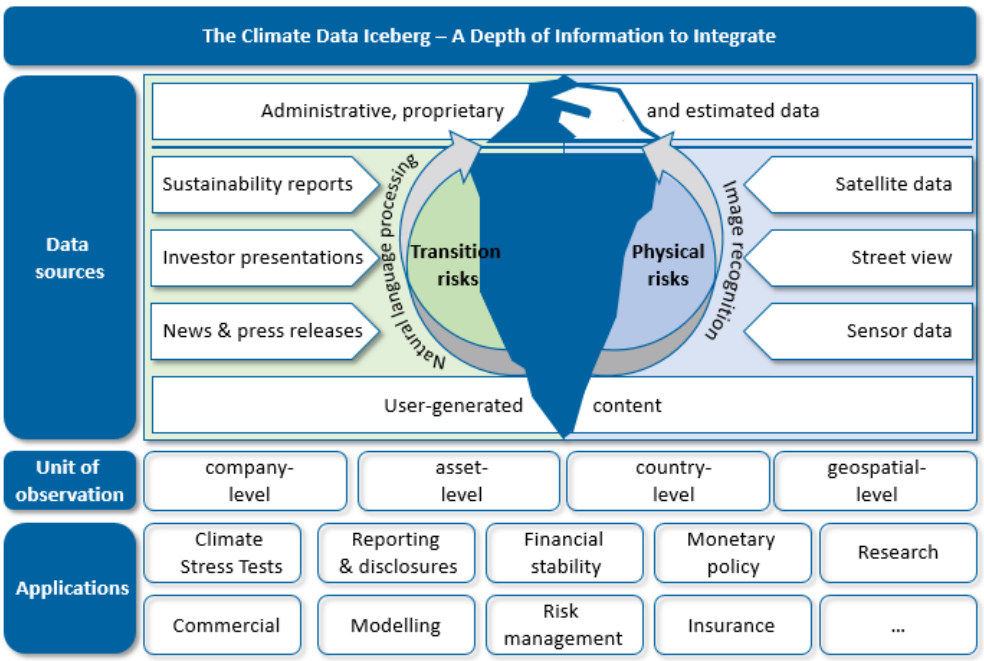


Figure 1: The climate data iceberg – a depth of information to integrate

2.1 Available textual sources of climate-related information

Many jurisdictions require large firms to publish annual and, increasingly, corporate sustainability reports. These reports, often referred to as Economic, Social and Governance (ESG), Corporate Social Responsibility (CSR), or non-financial reports, provide insights into environmental, social, and governance data, while also serving as a marketing tool (Bingler et al., 2022).

The content of sustainability reports varies by jurisdiction, leading to differences in coverage, scope, and standardization. New regulations may improve standardization by defining specific reporting variables, units, and scopes. However, central banks still need to rely on unstructured data sources as a bridge solution to obtain data for climate stress tests, financial stability assessments, and disclosures until structured data becomes widely available. Relying on publicly available reports is advantageous because:

(i) They cover a broader range of firms than any single proprietary database or jurisdiction.

(ii) They capture unique trends, such as changes in emissions goals, not covered by third-party data providers or by most regulations.

(iii) They help validate commercial data, where discrepancies among providers are common (BaFin, 2024).

(iv) Using public reports can reduce licensing restrictions.

The widespread availability of large language models (LLMs) provides a promising path for efficient extraction of information from these reports, with recent literature focusing on textual analyses for climate risk (e.g., Dimmelmeier et al., 2024). Recent studies find that grounding LLMs in domain-related documents produces

more targeted and accurate responses (Martín et al., 2024; Ni et al., 2023; Vaghefi et al., 2023; Zou et al., 2023). Beyond sustainability reports, valuable climate-related information is found in investor presentations, fund documents (Cruciani and Santagiustina, 2023), social media (Liu, Luo, and Lu, 2023), and news articles (Allahdadi, Fretheim, and Vindedal, 2024).

2.2 Available image sources of climate-related information

Images, particularly from satellite data, also represent valuable source for climate-related information with high relevance for central banks. Public satellite data and freemium models from national and international sources provide tailored data for various applications:

(i) Emissions Tracking: Satellite images can capture site-level GHG emissions with greater granularity than standard firm-level data, enabling precise tracking at plant sites.

(ii) Physical Risk Assessment: Satellite data can model climate risks (e.g., floods, wildfires) at a detailed geographic level, crucial for estimating assets at risk.

(iii) Data Quality Checks: Satellite data can validate structured data (e.g., correcting emission allocations incorrectly assigned to headquarters rather than actual sites).

(iv) Environmental Impact Monitoring: Satellite images allow for tracking impacts on local ecosystems, including flora, fauna, and land use changes.

Satellite data also enable insights into firm property sustainability features (e.g., solar panels, green roofs) and surrounding environmental features, allowing for assessments of assets at climate risk (Alonso-Robisco et al., 2024). Combining satellite data with administrative data as building usage from local agencies provides a more comprehensive view of the risks at the asset level. Furthermore, recent advances in multimodal methods (e.g., combining text and images in the same model) improve classification outcomes in several domains (e.g., Bernardi et al., 2016; Hu and Flaxman, 2018; Kiela et al., 2019; Pradeep et al., 2021).

This climate data iceberg holds a promising and largely untapped potential to incorporate climate-related information into decision-making for central banks, fintechs, investors, and data providers. Table 1 in the appendix lists publicly available sources, sorted by information type, to help stakeholders efficiently access and use these data.

3. Innovative statistical initiatives with Bundesbank's Data Service Centre to fill data gaps

To bridge climate-related data gaps, many central banks are recognizing the need for skills that go beyond traditional economics and finance, such as IT and natural sciences. The emerging, interdisciplinary demands in this domain motivate central banks to partner with each other and with academic experts to foster innovative approaches.

This case study presents projects staffed by teams with central bank and academia experts, integrating domain and technical knowledge while sharing resources and expertise. This setup promotes joint ownership of project success, facilitates knowledge transfer, and fosters integration into production systems. The Bundesbank's Data Service Centre collaborates in four major projects to address key aspects of unstructured climate data:

(i) Generative AI for Climate-Risk Analysis: In collaboration with central banking partners through the BIS Innovation Hub project Gaia, generative AI supports user-friendly climate-risk analysis (BIS, 2024).

(v) Efficient Extraction Algorithms: Partnering with Ludwig-Maximilian University Munich, this project focuses on algorithms for extracting data from sustainability reports (Dimmelmeier et al., 2024).

(vi) Mapping Earth Observation Potential: Collaborating with Banco de España, this project assesses Earth Observation imagery applications (Alonso-Robisco et al., 2024).

(vii) Enhanced Data with Satellite Information: With the Technical University Darmstadt, this project uses satellite data to improve structured data sets (Alves Werb et al., 2024).

Sections 4.1 and 4.2 describe projects using textual and multimodal data, respectively. Together, these initiatives form crucial steps towards a robust climate-related data infrastructure that integrates structured, textual, and visual data sources.

4.1 Selected projects leveraging textual data

Project Gaia is an initiative by the Bank for International Settlements (BIS) Innovation Hub Eurosystem Centre, with collaboration from the European Central Bank (ECB), Banco de España, and Deutsche Bundesbank. The project explores AI and machine learning applications to improve climate risk analysis through generative AI, specifically using GPT-4 to extract climate-related information from textual data (BIS, 2024). Designed to enhance central bank access to climate data in sustainability reports, Project Gaia aims to simplify information extraction for central bank analysts. A proof-of-concept demonstrates the feasibility of an AI tool that extracts key climate indicators from text, tables, and figures within unstructured PDF documents with approximately 80% accuracy. Nevertheless, though complex table structures still present challenges.

Building on Project Gaia's findings, Bundesbank's Data Service Centre, in collaboration with researchers at Ludwig Maximilian University of Munich (LMU), continues to explore automated extraction methods for climate-relevant indicators. This partnership evaluates different extraction methods, analyses results, and integrates various data types to create a functional, firm-level database of climate data accessible to financial analysts. The research agenda focuses on refining extraction strategies, addressing challenges in data consistency, and enhancing LLM performance for better climate data retrieval (Dimmelmeier et al., 2024). Figure 2 summarises the project motivation and goal.

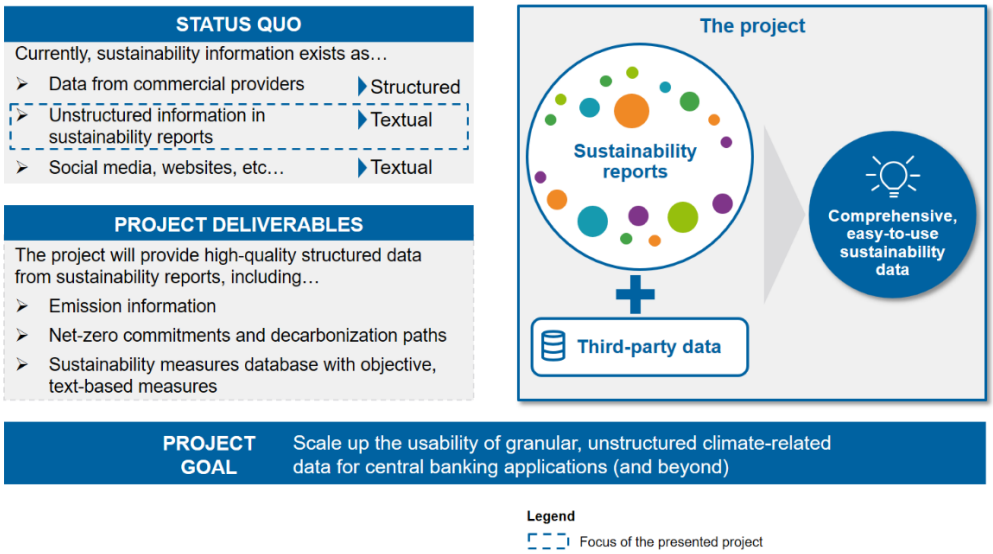


Figure 2: The Greenhouse gas insights and sustainability tracking (GIST) project between Deutsche Bundesbank and Ludwig-Maximilian-University Munich extracts and provides data from unstructured sustainability reports.

Dimmelmeier et al. (2024) tested three methods for extracting emission data from sustainability reports. Approach E1 successfully extracted data from eight reports but encountered unit harmonization issues. Building on these results, approach E2 broadened the search to reduce retrieval failures, yet still had limited accuracy due to frequent incorrect extractions by the LLM. Finally, approach E3 focused solely on tables but produced poor results, highlighting the limitations of table-only extraction. The results of this study emphasize the need to refine retrieval strategies, optimize LLM functionality, and standardize units for accurate data extraction. A high-level overview of the pipeline is shown in Figure 3.

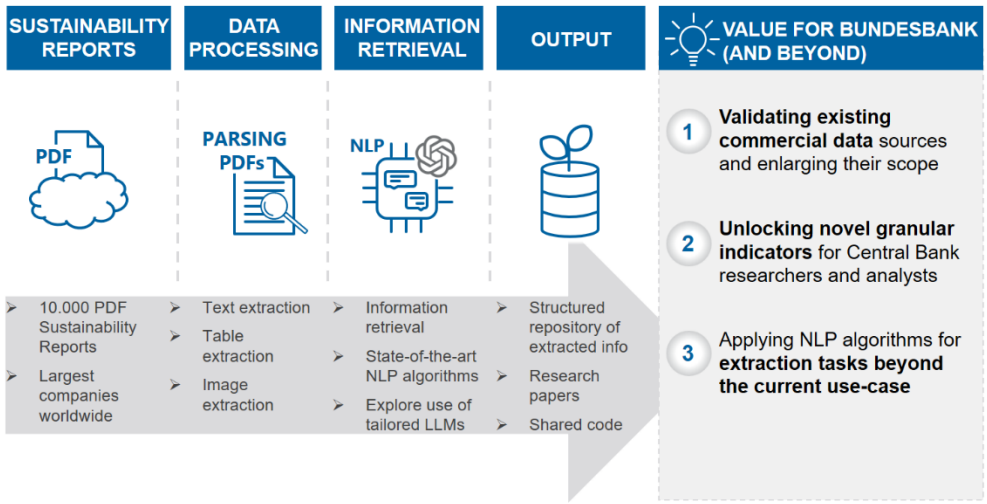


Figure 3: Recent advancements in Natural Language Processing fuel the Greenhouse gas insights and sustainability tracking (GIST) project extraction pipeline. OpenAI Logo © OpenAI, 2024.

Once concluded, the project will deliver value by enhancing the quality of existing data sources for central bank analysis, unlocking novel granular indicators and having created knowledge for extraction task beyond climate-related use cases. In

summary, this project paves the way for leveraging textual climate-related information and integrating results with structured data in the institution. In the next section, we describe two projects that purse similar goals analysing multimodal data.

4.2 Selected projects leveraging multimodal data

Unlike emissions, which arguably disperse globally, further environmental impacts are often stronger close to firms' facilities, affecting biodiversity through by-products like wastewater, toxins, and light pollution (e.g., Aska et al., 2024; Camilleri-Fenech et al., 2018; Néstor and Mariana, 2019). To assess these impacts, it is crucial to link firms' economic activities to their geographical sites.

Therefore, in collaboration with Banco de España, the Bundesbank's Data Service Centre explored the literature on satellite and remote sensing applications, highlighting uses in nowcasting, equity and commodities trading, and insurance. The study finds that despite established uses in other domains, satellite data's potential for climate finance remains underexplored. Through bibliometric analysis, the study identified five promising areas: physical risk, deforestation, energy, agriculture, and land use.

Furthermore, the combined potential of visual, textual, and structured information presents a novel and promising path to jointly learn from multiple representations of a firm and its assets. Partnering with computer scientists at TU Darmstadt, the Bundesbank aims to create a multimodal framework to enhance firm master data accuracy, benefiting central banks, supervisors, and statistical offices. A key use case involves linking climate-related data to firm sites, which includes geolocating sites, identifying economic activity, and accurately attributing subsidiaries' activities to parent firms (Alves Werb et al., 2024).This attribution is also crucial under new regulations requiring site-specific reporting.

Therefore, this project involves enhancing and validating the German subset of the Register of Institutions and Affiliates Data (RIAD) (Gábor-Tóth, Schild, & Walter, 2023), which contains master data for over 1.3 million firms and 293 variables for Germany. RIAD acts as a "single source of truth" for Bundesbank divisions and researchers. This project uses multimodal learning to validate firms' activities by combining satellite images, textual data, and visual information from sources like Google Maps, the National Aeronautics and Space Administration (NASA), the European Space Agency (ESA), and the Federal Agency for Cartography and Geodesy (BKG). High-resolution images and textual descriptions help classify buildings and verify economic sector classifications. For instance, if a firm claims to produce goods but has only a residential address, it raises data quality concerns, as outlined in Figure 4.

This project aims to develop an automated approach to flag inconsistencies between reported activities and physical sites, reducing manual checks and enhancing data accuracy. However, sector classification from images alone can be challenging, as firms may list separate administrative and production sites. Additional data like employee numbers help clarify the coherence firm size and its activities, making this project a valuable proof of concept for handling complex data validation scenarios.

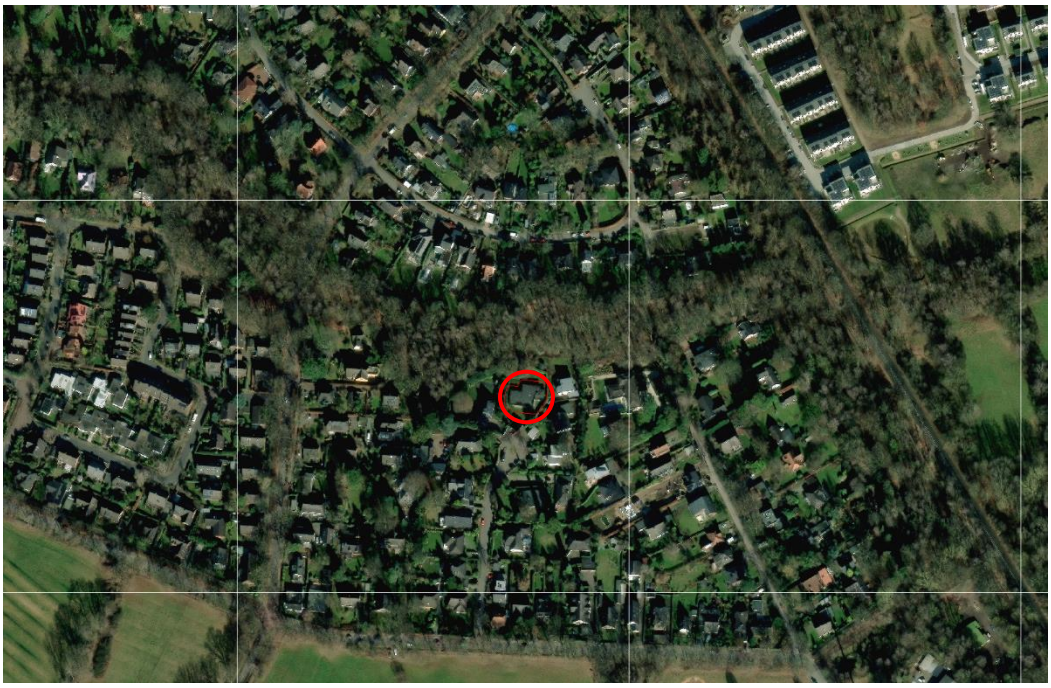


Figure 4: A manufacturing company of large products in a residential building as an example of an economic sector classification to be enhanced using unstructured data.-DE-BKG2024 (2024)

4. Lessons learned and the road forward

From completed and ongoing projects, several future pathways and lessons emerge. One key trend for central banks is incorporating biodiversity considerations, spotlighted by the NGFS Taskforce on Biodiversity-loss and Nature-related Risks (NGFS, 2024a). Biodiversity loss metrics, often derived from satellite imagery, are a growing area of interest (Skidmore et al., 2021). However, less attention has been given to the social and governance (S and G) components of ESG due to limited structured data (e.g., Rajesh, 2020). With advances in textual analysis, "S" and "G" insights can now be extracted directly from reports, social media, and other sources, providing valuable unstructured ESG information for risk analysis.

On the technical side, integrating unstructured data with existing datasets remains crucial. Even as regulatory frameworks push for harmonized disclosures, data alignment issues persist, particularly for historical information central banks require in time-series format. As climate data is a global public good, central banks have incentives to collaborate that go beyond classic data collection and standard setting collaboration, since each other's climate-related data (e.g., on GHG emissions) constitutes crucial information for decision-making.

A significant lesson from these projects is that the unstructured data skills developed in-house can benefit broader central banking functions, such as improving product data accuracy or monitoring risks through social media. Previous studies have outlined the potential of unstructured data sources to provide novel insights in several areas (Rosolia, Stapel-Weber, and Tissot, 2021). With methods for handling unstructured data becoming more accessible, we expect that the pressure to incorporate these data into central bank's decision-making will increase.

Beyond data extraction, central banks face challenges in data standardisation and repository management, which are crucial for data-driven decision-making. High-quality technical infrastructure must allow for efficient data discovery, retrieval, and processing. Data quality issues, especially for textual and satellite imagery, pose specific challenges; initial extraction must ensure completeness, granularity, and accuracy.

Satellite and textual data introduce complexities, such as mapping geolocated images to administrative data and transforming text into structured insights. While AI can simplify some of these tasks, data analysts still need to understand their specific data requirements and underlying assumptions to ensure unbiased results. Collaboration with external partners introduces trust and dependency considerations, highlighting the need for data protection due diligence.

Success factors for academia–central bank collaborations identified in the case study include having a clear task division by methods and domains, ensuring that each partner bringing strategic resources to the table, and sharing a common, attainable goal (e.g., developing novel methods, advancing the academic field, and prototyping practical applications).

5. Conclusions

Central banks and supervisors are increasingly utilizing climate-related data, which extends beyond structured data to a wealth of unstructured data, such as texts and images. We introduce the "climate data iceberg" to represent this landscape: the visible tip comprises structured, accessible data, while below lies extensive body of unstructured climate data. We then present a selection of projects conducted at the Deutsche Bundesbank's Data Service Centre that aim at harnessing novel methods to bring the information from the bottom of the iceberg to the surface. Such information can complement, enhance, or even replace existing structured data collections.

However, multiple challenges remain, including merging diverse data sources and integrating prototypes into central banks' complex infrastructures. The resources and technical support needed for this transition are substantial, but successful generalization of these approaches could expand their applicability across institutions and beyond climate-related data. Moving forward, central banks can benefit by incorporating unstructured information from various sources, such as investor presentations, balance sheets, social media, and online price and product data to foster insights in multiple domains.

References

Allahdadi, M. R., Fretheim, T., & Vindedal, K. (2024). *The value of climate change news: A textual analysis.* Available at SSRN 4868998.

Alonso-Robisco, A., Bas, J., Carbo, J. M., de Juan, A., & Marques, J. M. (2024). Where and how machine learning plays a role in climate finance research. *Journal of Sustainable Finance & Investment*, 1-42. Advance online publication.

Alves Werb, G., Reichenbach, L., Yalcin-Roder, E., & Walter, S. (2024). *A picture is worth a thousand definitions: validating company data with satellite images and textual data*. Deutsche Bundesbank Technical Report (forthcoming).

Aska, B., Franks, D. M., Stringer, M., & Sonter, L. J. (2024). Biodiversity conservation threatened by global mining wastes. *Nature Sustainability, 7*(1), 23-30.

BaFin (2024). Durchführung einer Marktstudie zur Erhebung von und Umgang mit ESG-Daten und ESG-Ratingverfahren durch Kapitalverwaltungsgesellschaften. *Bundesanstalt für Finanzdienstleistungsaufsicht. Marktstudie*. Retrieved 2024-07-16, from https://www.bafin.de/SharedDocs/Downloads/DE/dl_ESG-Studie_PDF_20240214.html

Battiston, S., Mandel, A., Monasterolo, I., Schütze, F., & Visentin, G. (2017). A climate stress-test of the financial system. *Nature Climate Change, 7*(4), 283-288.

Bernardi, R., Cakici, R., Elliott, D., Erdem, A., Erdem, E., Ikizler-Cinbis, N., Plank, B. (2016). Automatic description generation from images: A survey of models, datasets, and evaluation measures. *Journal of Artificial Intelligence Research, 55*(1), 409-442.

Bingler, J. A., Kraus, M., Leippold, M., & Webersinke, N. (2022). Cheap talk and cherry-picking: What ClimateBert has to say on corporate climate risk disclosures. *Finance Research Letters, 47*, 102776.

BIS (2024). Project Gaia – Enabling climate risk analysis using generative AI. *Bank for International Settlements, Basel.* . Retrieved 2024-07-08, from https://www.bis.org/publ/othp84.pdf

Camilleri-Fenech, M., Oliver-Solà, J., Farreny, R., & Gabarrell, X. (2018). Where do islands put their waste? – A material flow and carbon footprint analysis of municipal waste management in the Maltese Islands. *Journal of Cleaner Production, 195*(2018), 1609-1619.

Cruciani, C., & Santagiustina, C. R. M. A. (2023). The present and future of sustainability disclosure in equity investment funds' pre-contractual documents: Mapping ESG discourse through STM. *Finance Research Letters, 58*, 104033.

de Villiers, C., La Torre, M., & Molinari, M. (2022). The Global Reporting Initiative's (GRI) past, present and future: critical reflections and a research agenda on sustainability reporting (standard-setting). *Pacific Accounting Review, 34*(5), 728-747.

Deutsche Bundesbank (2022). Climate-related data successfully procured. Press release. Retrieved 2022-12-09, from https://www.bundesbank.de/en/press/press-releases/climate-related-data-successfully-procured-869246

Dimmelmeier, A., Doll, H. C., Schierholz, M., Kormanyos, E., Fehr, M., Ma, B., Kreuter, F. (2024). *Informing climate risk analysis using textual information – A research agenda*. Proceedings of the 1st Workshop on Natural Language Processing

meets Climate Change. Association for Computational Linguistics (forthcoming).

EU (2019). Regulation (EU) 2019/2088 of the European Parliament and of the Council of 27 November 2019 on sustainability-related disclosures in the financial services sector. Official Journal of the European Union. Retrieved 2024/03/21, from https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32019R2088

EU (2020). Regulation (EU) 2020/852 of the European Parliament and of the Council of 18 June 2020 on the establishment of a framework to facilitate sustainable investment, and amending Regulation (EU) 2019/2088. Official Journal of the European Union. Retrieved 2024/03/21, from https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32020R0852

EU (2022). Directive (EU) 2022/2464 of the European Parliament and of the Council of 14 December 2022 amending Regulation (EU) No 537/2014, Directive 2004/109/EC, Directive 2006/43/EC and Directive 2013/34/EU, as regards corporate sustainability reporting. Official Journal of the European Union. Retrieved 2024/03/21, from https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32022L2464.

EU (2023). Commission Delegated Regulation (EU) 2023/2772 of 31 July 2023 supplementing Directive 2013/34/EU of the European Parliament and of the Council as regards sustainability reporting standards. Official Journal of the European Union. Retrieved 2024/03/21, from https://eur-lex.europa.eu/eli/reg_del/2023/2772/oj

Fabris, N. (2020). Financial stability and climate change. *Journal of Central Banking Theory and Practice, 9*(3), 27-43.

Gábor-Tóth, E., Schild, C., and Walter, S. (2023). Understanding Overlaps between Different Company Data. Technical Report 2023-06. Deutsche Bundesbank, Research Data and Service Centre, https://www.bundesbank.de/resource/blob/846050/e6d1cdec6f14ea1ae19211522f9dd831/mL/2023-06-company-data.pdf.

GeoBasis-DE-BKG2024. (2024). Terms of use: Https://sg.geodatenzentrum.de/web_public/nutzungsbedingungen.pdf

Hu, A., & Flaxman, S. (2018). *Multimodal sentiment analysis to explore the structure of emotions*. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 350–358. .

IFRS (2023). Proposed IFRS taxonomy: IFRS sustainability disclosure taxonomy. Retrieved 2024-06-15, from https://www.ifrs.org/content/dam/ifrs/project/ifrs-sustainability-disclosure-taxonomy/proposed-taxonomy/pt-cd-issb-2023-1-sustainability-taxonomy.pdf

Kiela, D., Bhooshan, S., Firooz, H., Perez, E., & Testuggine, D. (2019). *Supervised multimodal bitransformers for classifying images and text*. arXiv preprint arXiv:1909.02950.

Liu, M., Luo, X., & Lu, W.-Z. (2023). Public perceptions of environmental, social, and governance (ESG) based on social media data: Evidence from China. *Journal of Cleaner Production, 387*, 135840.

Martín, R., Ranger, N., Schimanski, T., & Leippold, M. (2024). *Harnessing AI to assess corporate adaptation plans on alignment with climate adaptation and resilience goals*. Available at SSRN 4878341.

Monasterolo, I. (2020). Climate change and the financial system. *Annual Review of Resource Economics, 12*(2020), 299-320.

Néstor, M. C., & Mariana, C. (2019). Impact of Pharmaceutical Waste on Biodiversity. In L. M. Gómez-Oliván (Ed.), *Ecopharmacovigilance: Multidisciplinary Approaches to Environmental Safety of Medicines* (pp. 235-253). Cham: Springer International Publishing.

NGFS (2022). Final report on bridging data gaps Retrieved 2024-05-15, from https://www.ngfs.net/sites/default/files/medias/documents/final_report_on_bridging_data_gaps.pdf

NGFS (2024b). NGFS publishes Conceptual Framework for Nature-related Financial Risks at launch event in Paris. *Network for Greening the Financial System. Press release*. Retrieved 2024-07-11 from https://www.ngfs.net/en/communique-de-presse/ngfs-publishes-two-complementary-reports-nature-related-risks

Ni, J., Bingler, J., Colesanti-Senni, C., Kraus, M., Gostlow, G., Schimanski, T., Leippold, M. (2023). *Chatreport: Democratizing sustainability disclosure analysis through llm-based tools*. arXiv preprint arXiv:2307.15770.

Nightingale, J., Mittaz, J. P. D., Douglas, S., Dee, D., Ryder, J., Taylor, M., Merchant, C. (2019). Ten priority science gaps in assessing climate data record quality. *Remote Sensing, 11*(8), 986.

Pradeep, R., Ma, X., Nogueira, R., & Lin, J. (2021). *Scientific claim verification with VerT5erini*. Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis, 94–103.

Rajesh, R. (2020). Exploring the sustainability performances of firms using environmental, social, and governance scores. *Journal of Cleaner Production, 247*, 119600.

Rosolia, A., Stapel-Weber, S., & Tissot, B. (2021). New developments in central bank statistics around the world. *Statistical Journal of the IAOS, 37*(4), 1055-1060.

Schellhorn, C. (2020). Financial system stability, the timing of climate change action and the federal reserve. *Journal of Central Banking Theory and Practice, 9*(3), 45-59.

Schmieder, C., Tissot, B., Peronaci, R., Quang, P. B., Triebskorn, E., Izzati, N., & Artman, M. (2021). *Sustainable finance data for central banks*. IFC Report No. 14. Bank for International Settlements.

SEC (2024). SEC adopts rules to enhance and standardize climate-related disclosures for investors. Press release. Retrieved 2024-04-11, from https://www.sec.gov/news/press-release/2024-31

Skidmore, A. K., Coops, N. C., Neinavaz, E., Ali, A., Schaepman, M. E., Paganini, M., Wingate, V. (2021). Priority list of biodiversity metrics to observe from space. *Nature Ecology & Evolution, 5*(7), 896-906.

Vaghefi, S. A., Stammbach, D., Muccione, V., Bingler, J., Ni, J., Kraus, M., Leippold, M. (2023). ChatClimate: Grounding conversational AI in climate science. *Communications Earth & Environment, 4*(1), 480.

Zou, Y., Shi, M., Chen, Z., Deng, Z., Lei, Z., Zeng, Z., Zhou, W. (2023). ESGReveal: An LLM-based approach for extracting structured data from ESG reports. arXiv preprint arXiv:2312.17264.

Appendix

Overview of selected potential unstructured data sources

Where to find sources, information types and usage potential — Table 1

Information type	Data source	Potential use	Publicly available
Firm installations	Renewable Energies Act (EEG) with address	Run the addresses through a geocoder to obtain the coordinates or use the Google API to directly obtain the satellite or Street View images directly.	Yes
	Emissions Trading Scheme (ETS) with address	Run the addresses through a geocoder to obtain the coordinates or use the Google API to directly obtain the satellite or Street View images directly.	Yes
	Firm data with legal requirements	Use the number of employees and the minimum legal working space to approximate the size of real estate (e.g., building or factory).	Partly
	GeoAsset by Spatial Finance Initiative	Open asset-level data on firm asset geolocations, currently available for beef, cement, iron and steel, petrochemical, paper and pulp, and waste management industries.	Yes
Satellite data (imagery, remote sensing) and overflight data	Open Buildings	Interesting for a preliminary training step to identify buildings and differentiate them from other natural features.	Yes
	German Federal Agency for Cartography and Geodesy (BKG)	Official data from administrative sources: • 3D building models (can be used as training data for building classifications, including building parts and roof types, number of floor, among others) • Satellite imagery from various missions • Remote sensing data • Overflight data in high resolution	Partly
	NASA	Satellite images with geocoordinates of firm facilities from the NASA API.	Yes
	ESA	Satellite images with geocoordinates of firm facilities from the ESA API.	Yes
	Copernicus	Satellite images with geocoordinates of firm facilities from the European Copernicus program.	Yes
	Open Street Maps API (Nominatim)	Use the API to collect metadata (annotations) of any buildings associated with the respective coordinates, such as building type, number of floors, roof type, among others.	Yes
	Google API	Satellite images with geocoordinates of firm facilities. Access is partly free with Developers Account.	Partly
	USGS	Earth Explorer of the US Geological Survey	Partly
	Sentinel Hub	Statistical, imaging, batch, process APIs and website-based Dashboards/GUI. Data can be downloaded in image form or processed as statistical indicators. Time-series extractions are possible. Free access is limited, pricing depends the type of use (e.g., commercial or research).	Partly
	EUMETSAT Data Store and EUMETView	Online viewing and analysis of satellite imagery, download of prepared datasets from Data Store. 'Core' datasets are free, 'Recommended' are not.	Partly

Information type	Data source	Potential use	Publicly available
Textual information	Corporate sustainability reports	Sustainability reports offer a wide array of qualitative, quantitative and context information. Repositories and dispersed reports exist publicly available online.	Yes
	Social media	Social media posts can contain information about the public's perception of companies, such as for transition risks. APIs (free and paid) exist for multiple platforms.	Partly
	Newspaper reports and press releases	Newspaper articles and press releases can contain textual sustainability information on companies. Information is often available online, repositories exist, dispersed information is partly available for free, partly behind paywalls.	Partly
	Fund documents	Investment fund documents are published by the issuers and depending on the jurisdiction must or can contain sustainability related information.	Yes
Validation of test data	Crowdsourcing	Map relevant images that are potentially difficult to classify and organise an event to crowdsource efforts to tag these data.	Partly

Source: Authors

DEUTSCHE BUNDESBANK
EUROSYSTEM

The climate data iceberg – A depth of information to integrate

12th biennial IFC Conference, BIS Basel, August 2024

Hendrik Christian Doll, Emily Kormanyos, Susanne Walter, Gabriela Alves Werb

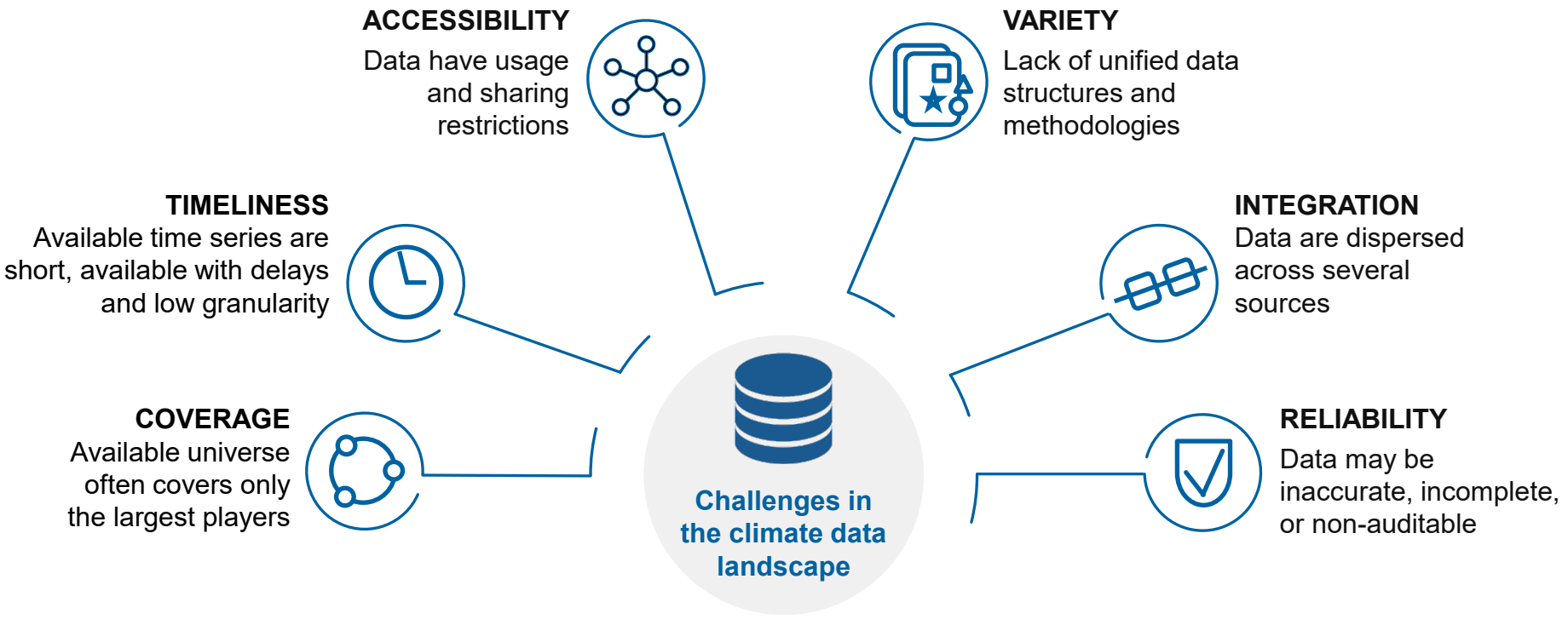
“Climate risks are a source of considerable financial risks.”*

*“Climate change has consequences for us as a central bank pursuing our primary mandate of price stability, […] financial stability and banking supervision.**”*

*“Climate change and climate policy also affect inflation and growth. [...] This will require, amongst other things, better data, which we should also demand.***”*

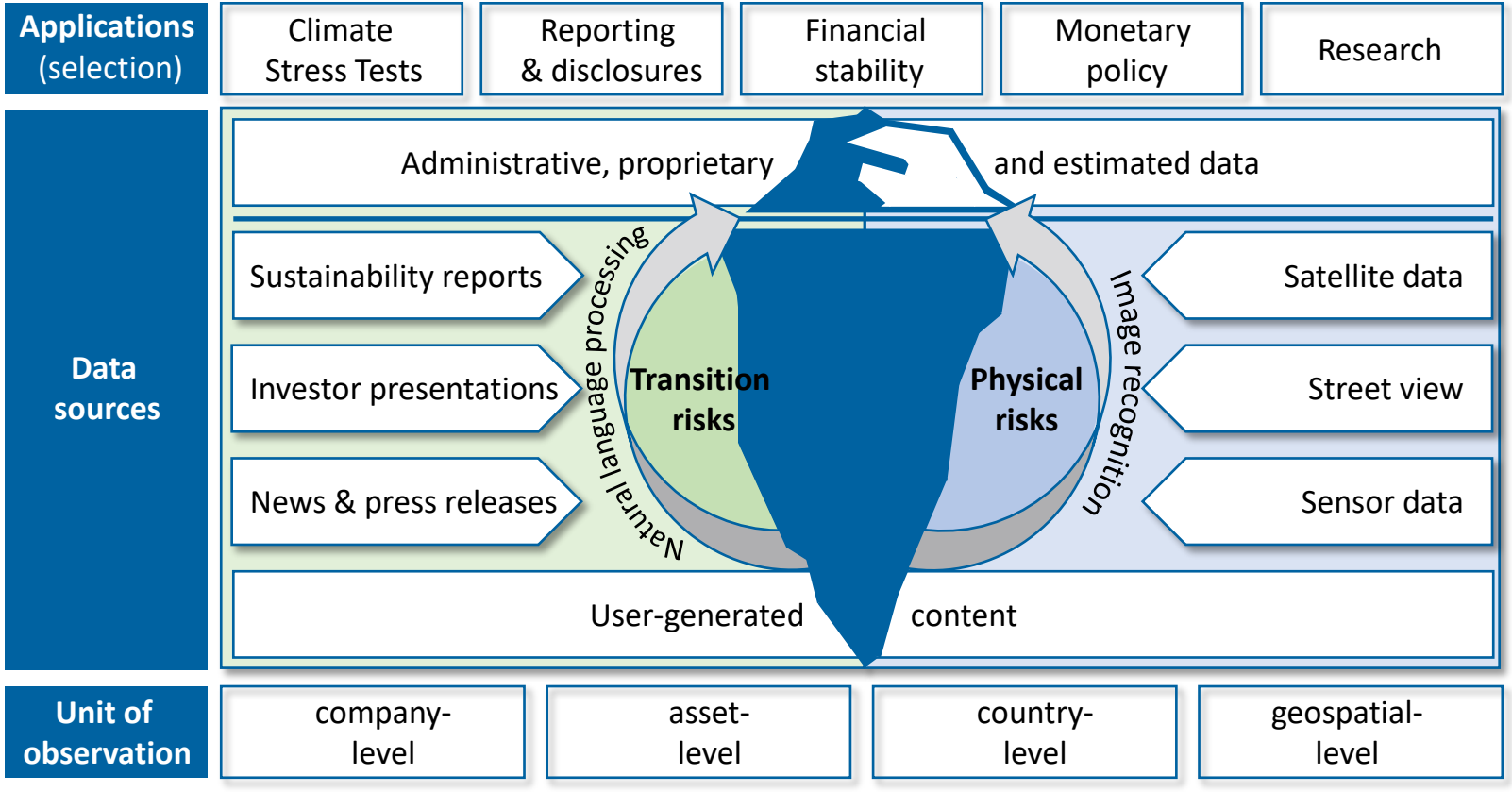
* Mauderer (2019)
** Lagarde (2021)
*** Nagel (2022)

Challenges in the climate-related data landscape*



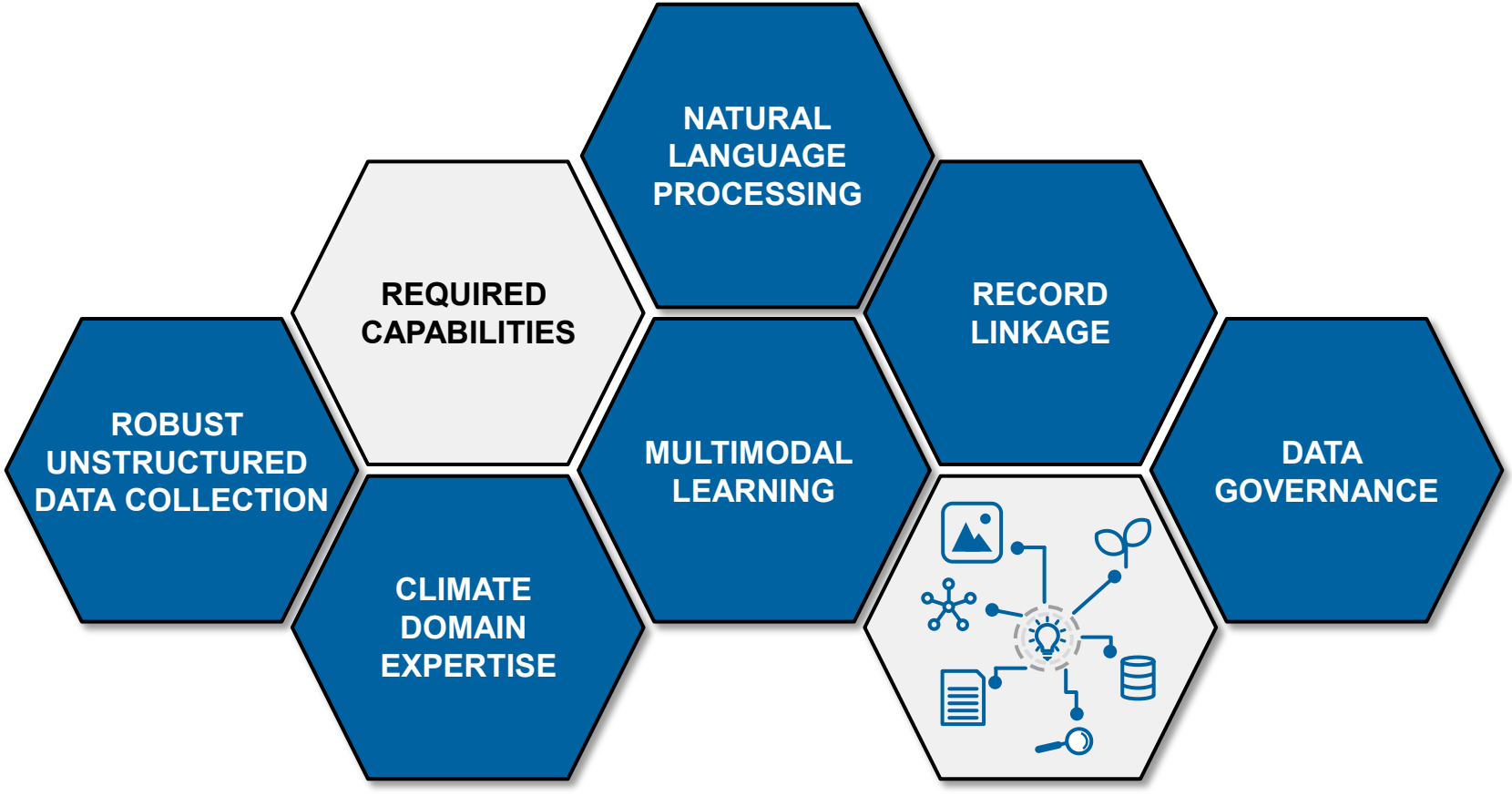
* Source: Doll, Alves Werb (2023)

The climate data iceberg – A depth of information to integrate*



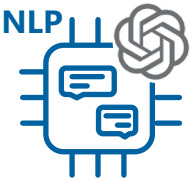
* Source: Doll, Kormanyos, Walter, Alves Werb (forthcoming)

Required capabilities to assess climate-related risks with novel data sources*



* Source: Doll, Kormanyos, Walter, Alves Werb (forthcoming)

Structured sustainability data through machine learning*

SUSTAINABILITY REPORTS	DATA PROCESSING	INFORMATION RETRIEVAL	OUTPUT
			
- 10.000 PDF Sustainability Reports - Largest companies worldwide	- Text extraction - Table extraction - Image extraction	- Information retrieval - State-of-the-art LLMs - Evaluation and integration	- Structured repository of extracted info - Research papers - Shared code

DELIVERABLES

High-quality structured data, such as:

- Emission information
- Net-zero commitments and decarbonization paths

VALUE

- ➢ **Validate existing data** sources and enlarging their scope
- ➢ **Unlock novel granular indicators** for central bank researchers and analysts
- ➢ Leverage lessons learned and methods for other use-cases

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

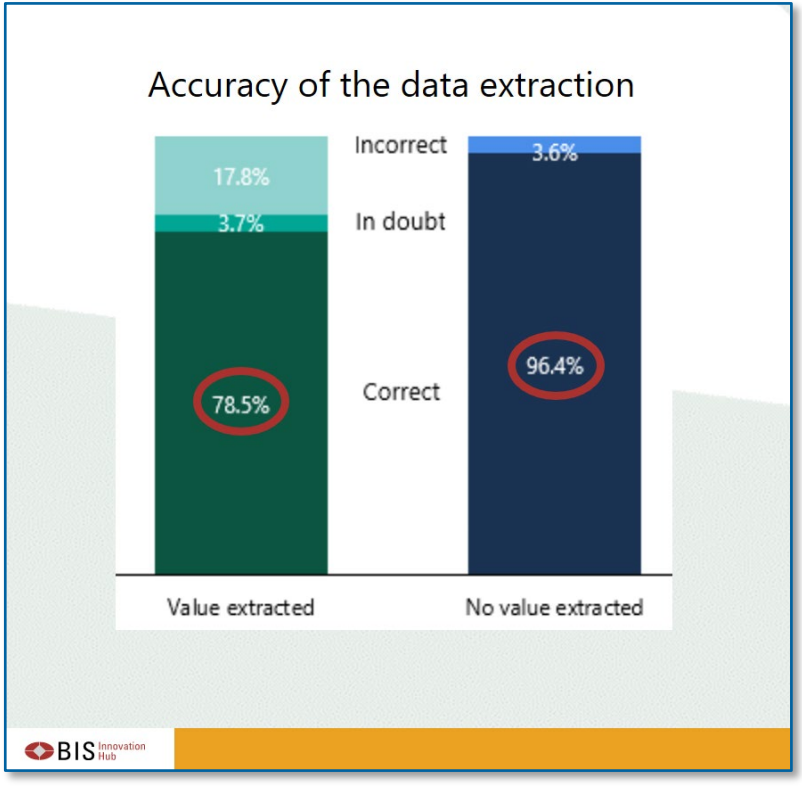
DEUTSCHE BUNDESBANK EUROSYSTEM

* Joint project LMU Munich University and Deutsche Bundesbank. For details cp. Dimmelmeier, Doll, Schierholz, Kormanyos, Fehr, Ma, Beck, Fraser & Kreuter (2024).

Project Gaia – Enabling climate risk analysis using generative AI*

Project Gaia makes assessing climate risk more transparent and efficient, as it uses generative AI to decipher vast unstructured data sets. If realised, Gaia has the potential to be a powerful tool for central banks in their comprehensive approach to assessing economic reality and risks.

Christine Lagarde



*BIS Innovation Hub (2024)

A picture is worth a thousand definitions*

Target

Validate structured firm data

ID	Firm	Street	City	Postal Code	Economic Sector	Employ-ees	Parent Firm
1	Firm 1	Street X	City A	1234	Car manufacturing	500	X
2	Firm 2	Street Y	City B	5678	Car manufacturing	400	X

Application

- **Multimodal deep learning** to validate secondary firm data (not collected for statistical purposes)
- Leverage **publicly available data** (e.g., satellite images, street view, data from companies' websites)

Advantages

- **Reduce effort** with manual validations and quality checks (millions of entities)

Challenges

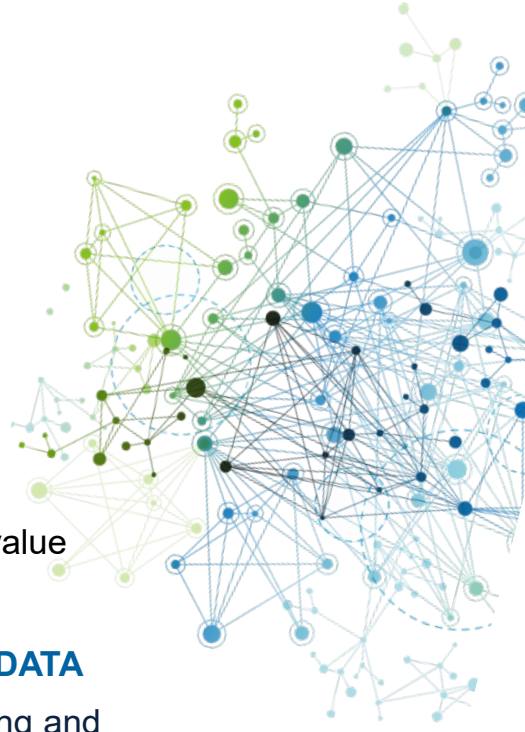
- High effort to generate annotate **training data**
- Handling **special cases** for companies with multiple offices or activities

TECHNISCHE UNIVERSITÄT DARMSTADT

DEUTSCHE BUNDESBANK EUROSYSTEM

* Joint project Technical University of Darmstadt and Deutsche Bundesbank. For details cp. Alves Werb, Reichenbach, Yalcin-Roder & Walter (forthcoming).

Takeaways



1 NOVEL DATA FOR CENTRAL BANKS

Recent methodological advances allow leveraging information from unstructured data sources

2 COLLABORATION TO FOSTER INNOVATION

Building a network of institutional and academic partners to join efforts and leverage interdisciplinary expertise provides value

3 APPLICABILITY BEYOND CLIMATE-RELATED DATA

Knowledge gained can be applied in central banking and supervision for enriching a wide range of structured data sources

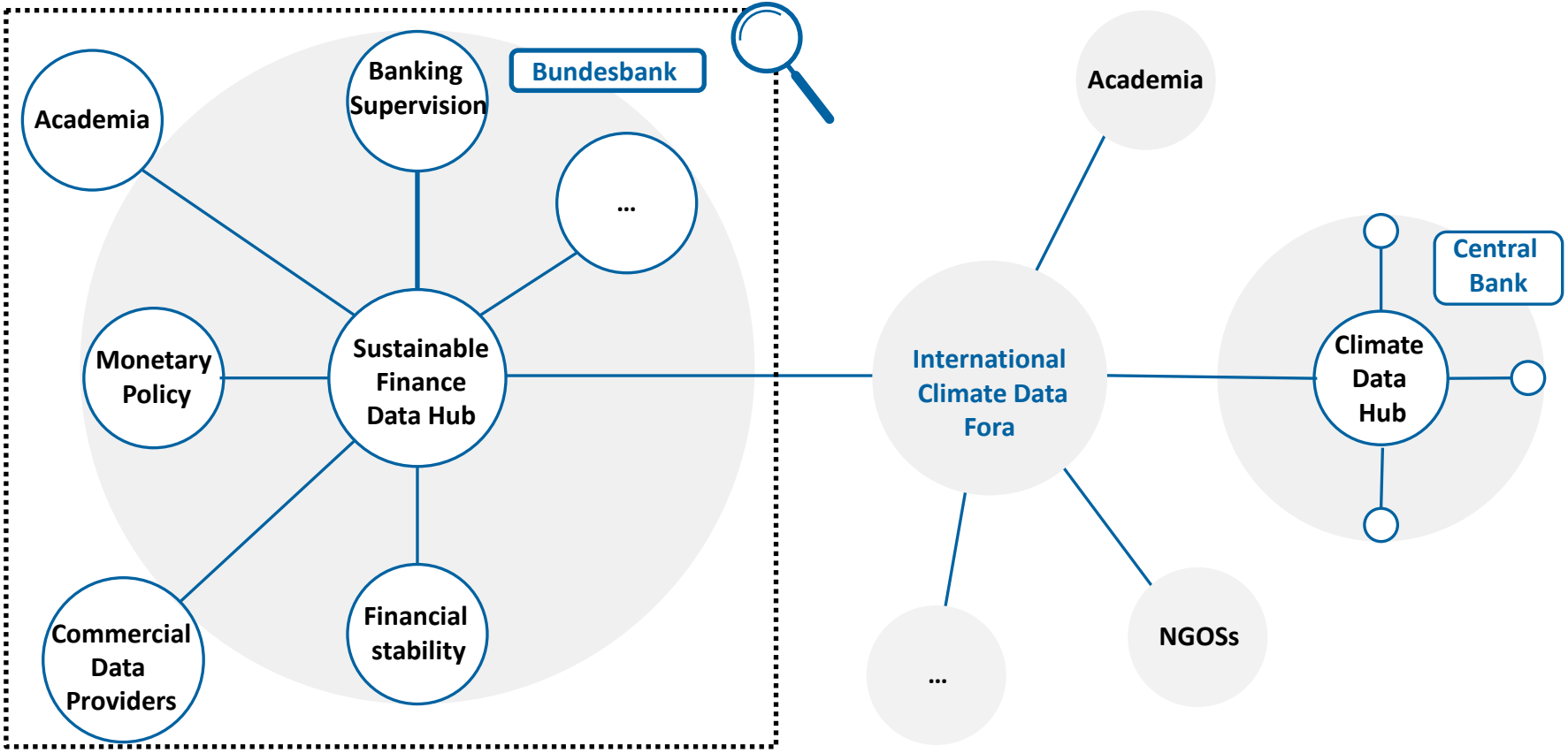
References

- Alves Werb, G., Reichenbach, L., Yalcin-Roder, E., & Walter, S. (2024). A picture is worth a thousand definitions: validating company data with satellite images and street view. Deutsche Bundesbank Technical Report (forthcoming).
- BIS Innovation Hub (2024), Project Gaia: enabling climate risk analysis. Press release, 19 March 2024, Link.
- Dimmelmeier, A., Doll, H.C., Schierholz, M., Kormanyos, E., Fehr, M., Ma, B., Beck, J,. Fraser, A. & Kreuter, F. (2024). Informing climate risk analysis using textual data – A research agenda. Deutsche Bundesbank Technical Report, 2024-01, Link.
- Doll, H. C. & G. A. Werb (2023). Innovation for improving climate-related data – Lessons learned from setting up a data hub. AStA Advances in Statistical Analysis , 17(3), 355-380, Link.
- Doll, H. C., Kormanyos, E., Walter, S. & G. A. Werb (2024). The climate data iceberg – A depth of information to integrate. IFC Bulletin (forthcoming).
- Lagarde, C. (2021). Climate change and central banks: analysing, advising and acting. Speech at the international climate change conference, Venice, Link.
- Mauderer, S. (2019). Central banks – a crisis manager for the climate? Speech at the second financial markets conference 29.10.2019 Frankfurt am Main, Link.
- Nagel, J. (2022). Speech at the ceremony to mark the inauguration of the new President of the Deutsche Bundesbank. 11.01.2022 Frankfurt am Main, Link.

hendrik.doll@bundesbank.de
financial-market-data@bundesbank.de
https://www.bundesbank.de/en/bundesbank/research/rdsc

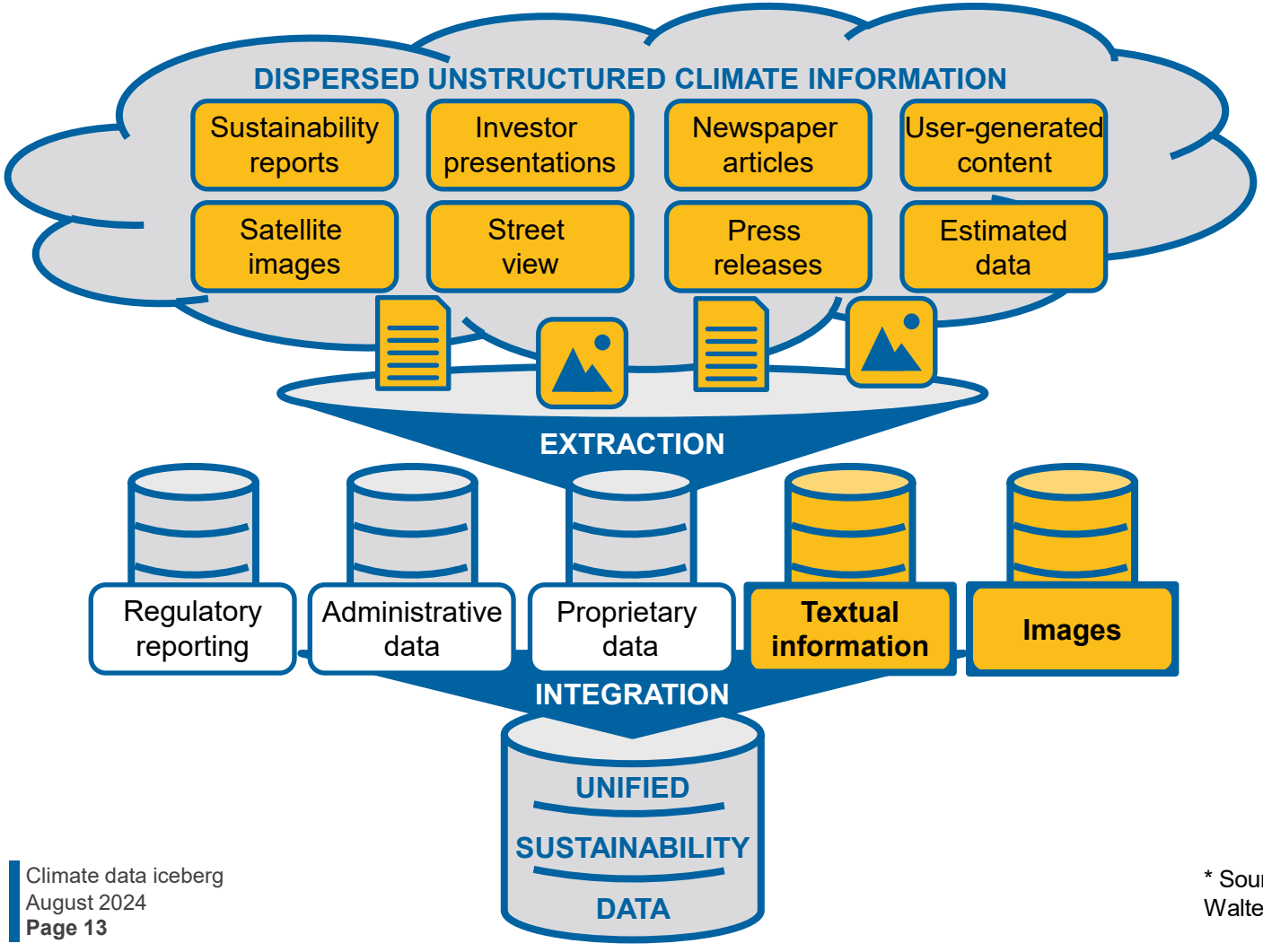
Backup

Building expertise and fostering exchange for innovation*



* Source: Doll, Alves Werb (2023)

Building a comprehensive data infrastructure with unstructured data*

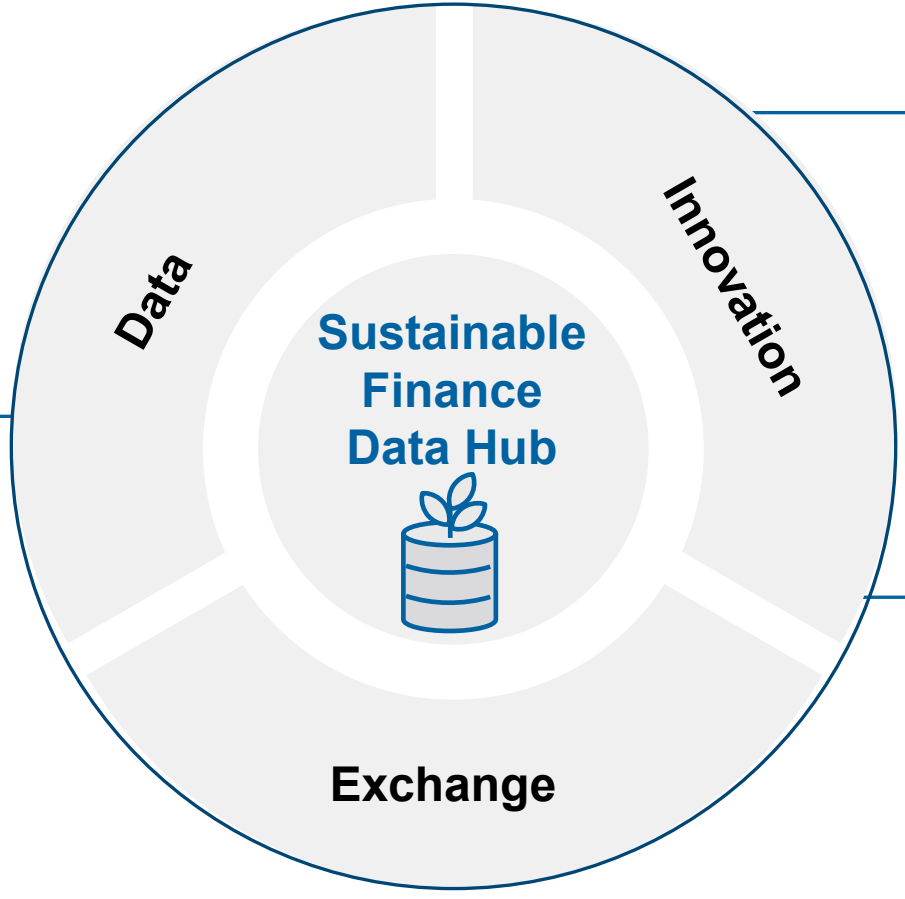


CHALLENGES

- Parsing
- Text extraction
- Image extraction
- Data linkage
- Evaluation

* Source: Doll, Kormanyos, Walter, Alves Werb (forthcoming)

The Sustainable Finance Data Hub's focus areas*



Climate-related micro data

- Common data procurement (ESCB)
- Link with administrative data and facilitate internal data provision
- Center of competence for methodological questions

Generate and transform data in **innovative** ways

- Collaboration with academia
- Central bank cooperation (e.g., BIS Innovation Hub)

Enhance institutional **communication and coordination**

- ESCB exchange on climate data
- Chairing the NGFS data expert network
- Exchange with statistical offices and data producers

* Source: Doll, Alves Werb (2023)

The *Greenhouse gas insights and sustainability tracking* (GIST) project extracts and provides data from unstructured sustainability reports

STATUS QUO

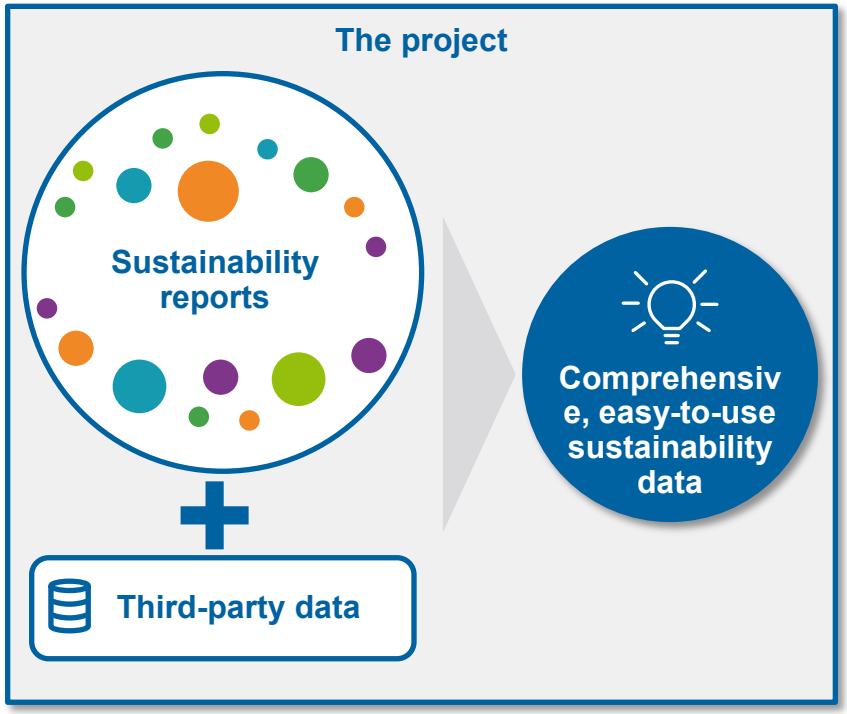
Currently, sustainability information exists as…

- Data from commercial providers ▸ Structured
- Unstructured information in sustainability reports ▸ Textual
- Social media, websites, etc… ▸ Textual

PROJECT DELIVERABLES

The project will provide high-quality structured data from sustainability reports, including…

- Emission information
- Net-zero commitments and decarbonization paths
- Sustainability measures database with objective, text-based measures



PROJECT GOAL Scale up the usability of granular, unstructured climate-related data for central banking applications (and beyond)

Legend Focus of the presented project

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

DEUTSCHE BUNDESBANK EUROSYSTEM

* Joint project LMU Munich University and Deutsche Bundesbank. For details cp. Dimmelmeier, Doll, Schierholz, Kormanyos, Fehr, Ma, Beck, Fraser & Kreuter (2024).

Innovation spillovers beyond climate-related data