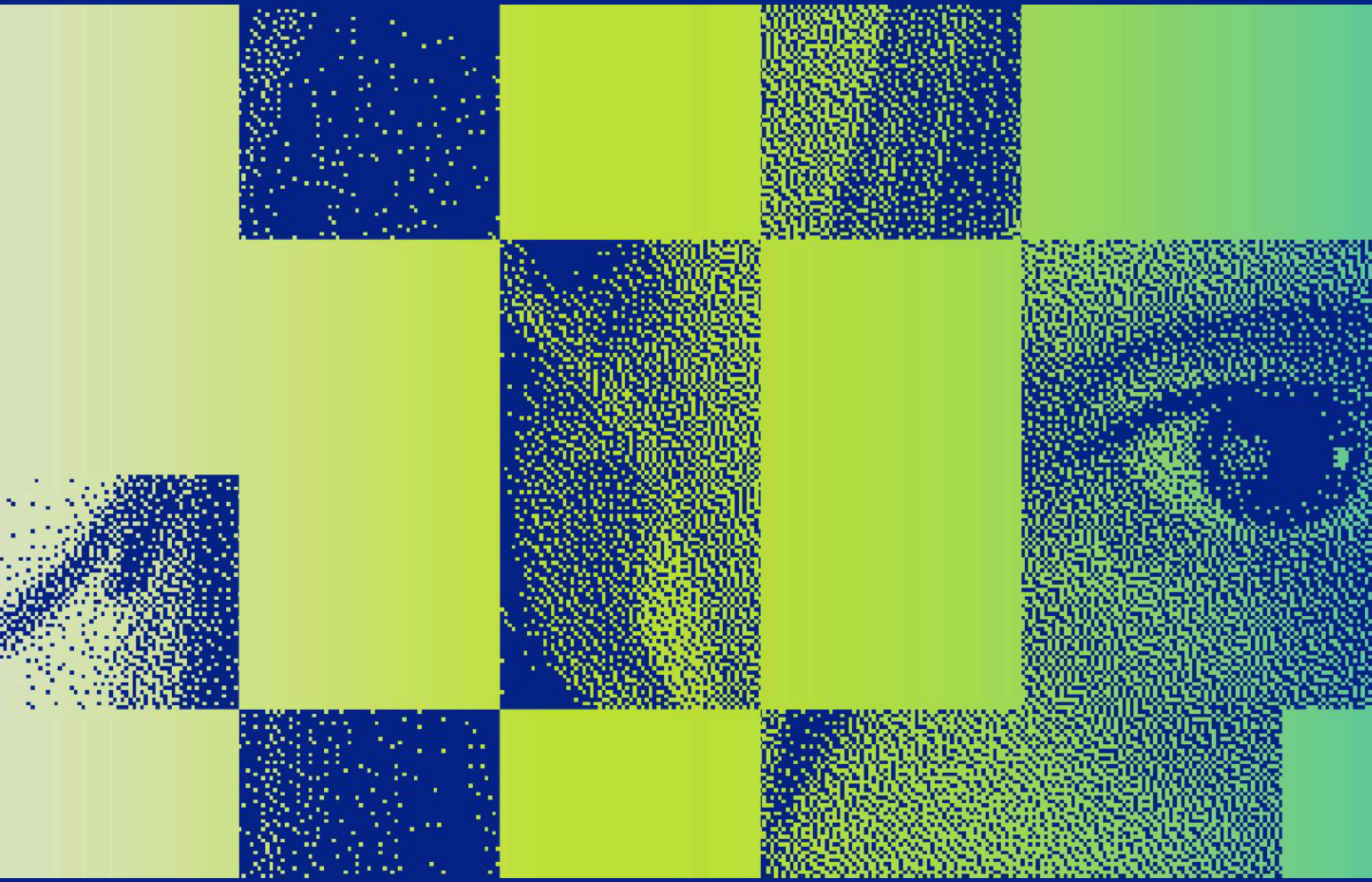


Global Policy Observatory for Data-centric AI

Data-centric AI
ODI Research
ADVANCING TRUST IN DATA
DAI

November 2024

Contents

About

This report was researched and produced by the Open Data Institute (ODI) and published in November 2024. Its lead authors were Thomas Carey-Wilson and Elena Simperl, with help from Arunav Das. If you want to share feedback by email or would like to get in touch, contact the DCAI project lead Elena Simperl at elena.simperl@theodi.org

Global Policy Observatory for Data-centric AI

Governments play a pivotal role in shaping the AI ecosystem. This includes establishing legal frameworks, such as the European Union's AI Act[1] and China's Algorithmic Recommendations Management Provisions[2], which set standards and guidelines for technology development and use.

Beyond legislation, governments influence public discourse by framing AI in terms of its benefits, risks, and broader societal implications. For instance, the UK's Plan for Digital Regulation outlines a pro-innovation approach, seeking to balance growth and safety.[3] Also, public funding can encourage advancements and broad adoption of emerging technologies. For example, the Indian government has launched the $1.25 billion 'IndiaAI Mission' to support computing infrastructure and startups, aiming to position the country as a significant player in the global AI landscape.[4]

Governments contribute to supranational initiatives to establish international standards and guidelines. For instance, the Organisation for Economic Co-operation and Development (OECD) has agreed on principles to promote trustworthy AI that respects human rights and democratic values.[5]

Although data was historically under emphasised as a critical component in AI development, there is now increasing attention paid to its importance in building effective AI systems. Google Brain co-founder Andrew Ng named this new focus 'data-centric AI' to highlight the importance of data engineering in the AI lifecycle[6].

[1] European Union (2024), 'Regulation (EU) 2024/1689 of the European Parliament and of the Council'
[2] Creemers R., Webster, G., Toner, H., Stanford University (2022), 'Internet Information Service Algorithmic Recommendation Management Provisions'
[3] Department for Science, Innovation and Technology (2023), 'Plan for Digital Regulation: developing an Outcomes Monitoring Framework 2022'
[4] INDIAai
[5]' OECD, 'OECD AI Principles overview'
[6] Brown, S., MIT Sloan (2022), 'Why it's time for 'data-centric artificial intelligence''

Historically, data was under emphasised in AI development. However, it is now recognised as a significant bottleneck alongside computational resources and expertise. Recent studies, such as those by Epoch AI, indicate that the availability of high-quality human-generated text data for training AI models could be exhausted between 2026 and 2032. This scarcity has led to increased data acquisition efforts, including agreements with platforms like Reddit and news organisations. The ODI views data-centric AI as a socio-technical approach that emphasises responsible AI practices, fosters public trust and promotes a competitive AI economy.[7] Through our observatory, we aim to assess whether this heightened focus on data-centric AI is reflected in global AI policies over time.

We are launching the Global AI Policy Data Observatory to assist policymakers and policy researchers interested in a data-centric approach to AI governance. Drawing from the OECD.AI repository[8], we will analyse **512** policy documents from **64** countries and two supranational organisations, including national strategies and emerging AI-related regulations, to create the evidence base to inform emerging data-centric AI research, regulation and standards. By analysing these documents, we aim to understand how 'data' and data-related topics are addressed in AI policies over time, identify underlying trends, and provide actionable recommendations. The **Annex** includes additional information on our methodology.

Based on this research, we recommend the following:

- Finding: Low and middle-income countries (LMICs) with emerging data governance and open data initiatives have an opportunity to bridge the data divide and strengthen their participation in global AI data governance →
  Recommendation: **Supranational bodies developing global AI data governance frameworks should design guidance on building digital infrastructure and accessing high quality data resources and make available to LMICs.**

- Finding: While interest in responsible AI continues to grow, essential data-centric concerns such as data provenance, lineage, transparency, licences and standards remain under emphasised in AI policies →
  Recommendation: **Policymakers and policy researchers should invest in data-centric AI toolkits and robust data documentation practices, supported by machine-readable metadata, thereby enabling regulatory oversight and establishing best practices.**

- Finding: There is a declining focus on open data in AI policies globally, which is concerning given its critical role to redress imbalances in the AI ecosystem →
  Recommendation: **As more and more data holders build walls around their data, policymakers need to promote safe and equitable data sharing, and think about data access remedies for AI contexts.**

[7] ODI (2024), 'Building a better future with data and AI: a white paper'
[8] OECD, 'OECD.AI repository'

1. Who is talking about data in their AI policies?

a. A minority of countries are emphasising data as an AI policy topic

To understand how government policies discuss ‘data’ about AI, we scanned policy documents for mentions of this keyword.

Overall, **37.5%** of the documents mention ‘data’, with a broad coverage of **92.2%** of countries analysed. Drawing from these policy documents, we calculated a score for each country that reflects the number of times data is mentioned in their policy documents, weighted to take into account the number of documents per country. Based on these scores, the highest score is **111.8** and the lowest score is **0.2.** We visualised these scores as a map and included a bar chart for the supranational organisations, which are difficult to represent on the world map, in **Figure 1.**

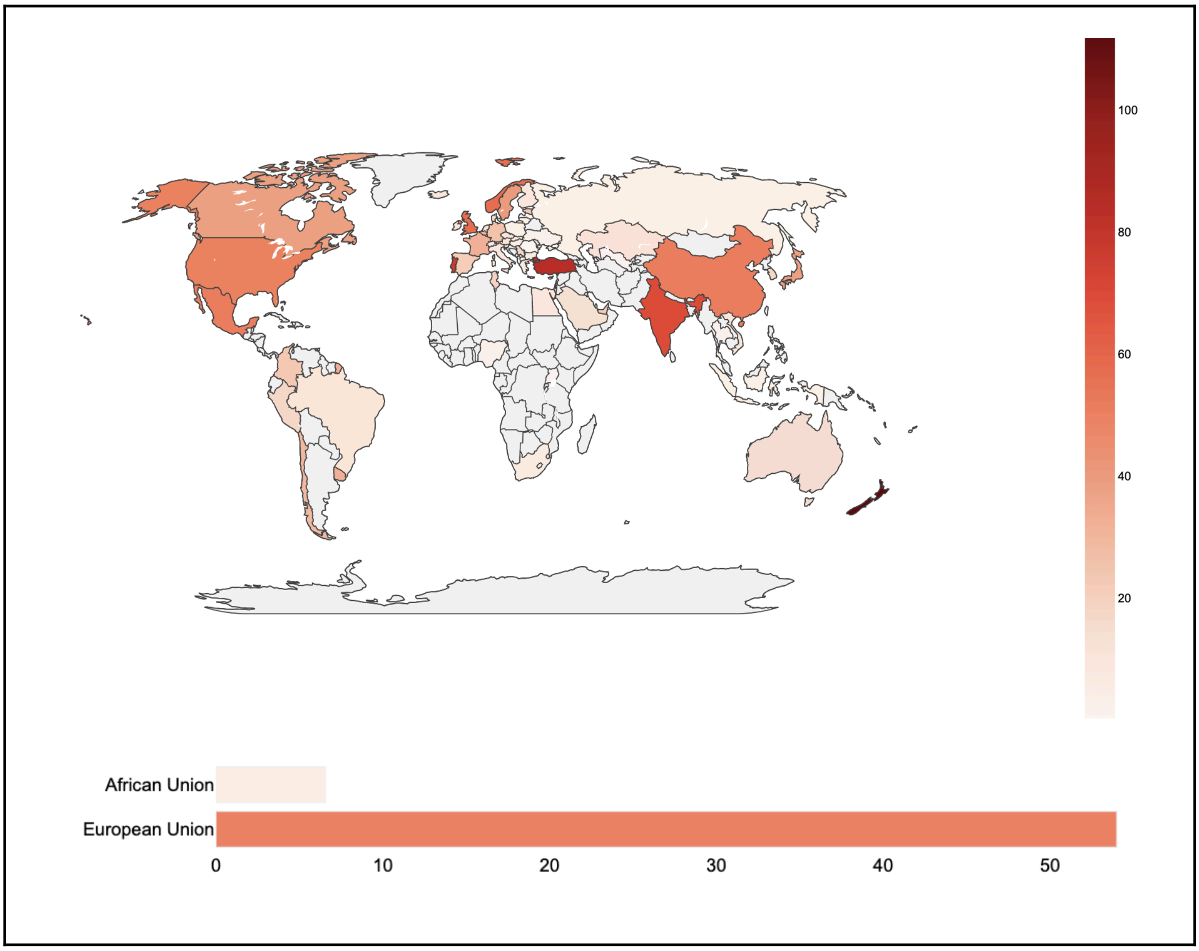


Figure 1. Emphasis on ‘data’ in national AI policies

Across the world, European governments are prominent in mentioning data, making up **41.2%** of the upper quartile (top **25%**), followed by East Asian governments at **23.5%.** North America and the Middle East each contribute **11.8%**, while Latin America and the Pacific represent **5.9%**. In this upper quartile, policies consistently underscore the critical role of data in AI development and deployment, highlighting several key themes:

1. **Access to high-quality data** is seen as a foundational priority, with many policies emphasising the need for robust data-sharing infrastructure.
2. **Governance and trust** emerge as central concerns, particularly regarding privacy, security and ethical use of data.
3. Data is also widely framed as **a strategic asset**, seen as essential for driving innovation, supporting economic growth and addressing societal challenges.
4. **International collaboration** is another prominent theme, with efforts to establish global standards and foster cooperation in data sharing and governance.
5. Finally, many policies illustrate **sector-specific applications of data**, such as its use in healthcare, fraud prevention and other domains to enhance public services and improve societal outcomes.

Meanwhile, the regional distribution of these countries aligns with the EU's[9] established leadership in regulatory areas like data protection and privacy[10]. Furthermore, East Asian countries have taken the lead in establishing relevant regulatory norms. For example, Japan's[11] presidency of the G7 at the Hiroshima Summit sought to establish common ground for responsible AI development and use, including governance of training data[12]. Our research shows that only certain governments emphasise data in their AI policies, but their location alone does not fully explain which countries are leading in this focus.

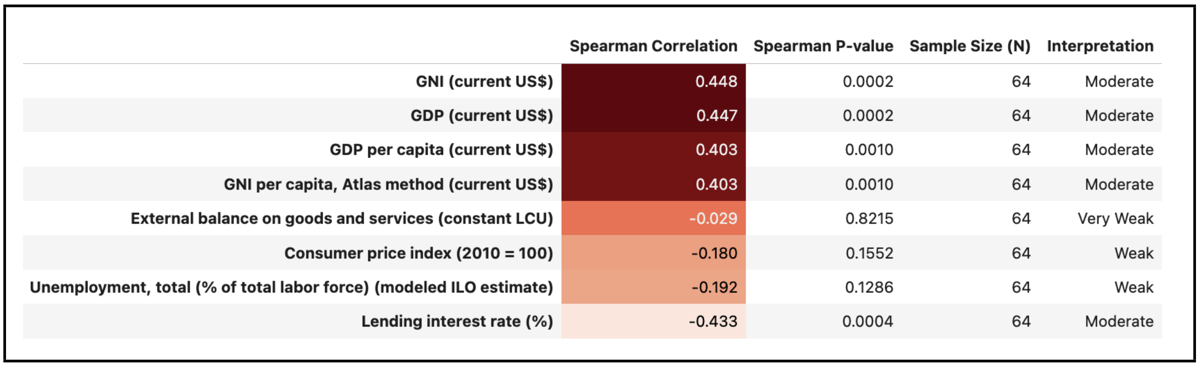
b. Specific types of countries emphasise data as an AI policy topic

To further explore the trends around which countries emphasise data in their AI policies, we used a statistical method called Spearman's rank correlation coefficient. This helped us identify which factors are associated with a greater focus on data in AI policies.

We consider the correlation between data-centric AI policy and the country's development stage, degree of globalisation, and digital readiness.

[9] European Union, alongside many European countries, is present in the upper quartile, ranking 13th.
[10] Bradford, A., Oxford Academic (2019), 'The Brussels Effect: How the European Union Rules the World'
[11] Japan is present in the upper quartile, ranking 8th.
[12] OECD Publishin (2023), 'G7 Hiroshima Process on Generative Artificial Intelligence (AI) Towards a G7 Common Understanding on Generative AI'

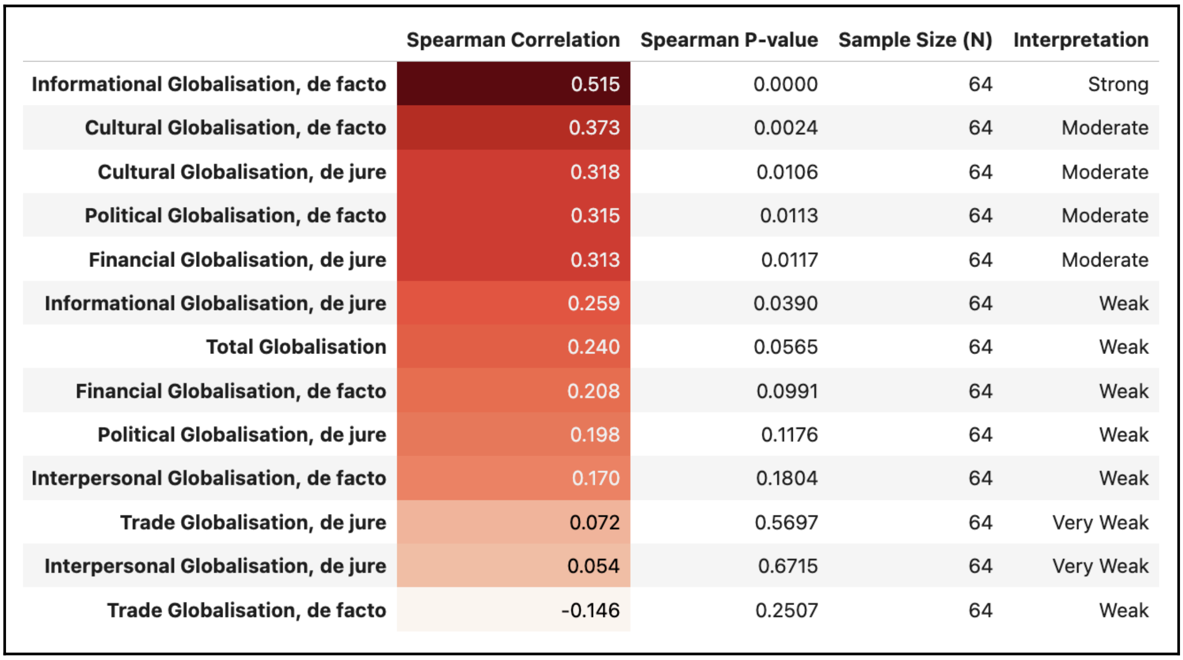
Firstly, using ***World Development Indicators***[13] from the World Bank's DataBank, we looked at the relationship between our scores and various macroeconomic indicators. **Figure 2** reveals a significant positive relationship between Gross National Income (GNI) and Gross Domestic Product (GDP) and an inverse relationship with the lending interest rate. This suggests that the countries most interested in data as an AI policy topic are typically wealthier nations that potentially have a higher degree of economic stability.



	Spearman Correlation	Spearman P-value	Sample Size (N)	Interpretation
GNI (current US$)	0.448	0.0002	64	Moderate
GDP (current US$)	0.447	0.0002	64	Moderate
GDP per capita (current US$)	0.403	0.0010	64	Moderate
GNI per capita, Atlas method (current US$)	0.403	0.0010	64	Moderate
External balance on goods and services (constant LCU)	-0.029	0.8215	64	Very Weak
Consumer price index (2010 = 100)	-0.180	0.1552	64	Weak
Unemployment, total (% of total labor force) (modeled ILO estimate)	-0.192	0.1286	64	Weak
Lending interest rate (%)	-0.433	0.0004	64	Moderate

Figure 2. Relationship between macroeconomic indicators and emphasis on 'data'

Secondly, given our focus on global trends, we used the ***KOF Globalisation Index*** (KOFGI)[14] to determine whether different forms of globalisation made a difference. Unsurprisingly, in **Figure 3**, 'informational globalisation' by far holds the most positive relationship to a focus on data-centric AI policy. This indicator is measured in terms of 'patent applications, nonresidents', 'high-technology exports' and 'internet bandwidth'. Together, these factors point to knowledge-based, innovation-driven economies that are competitive in high-technology sectors.

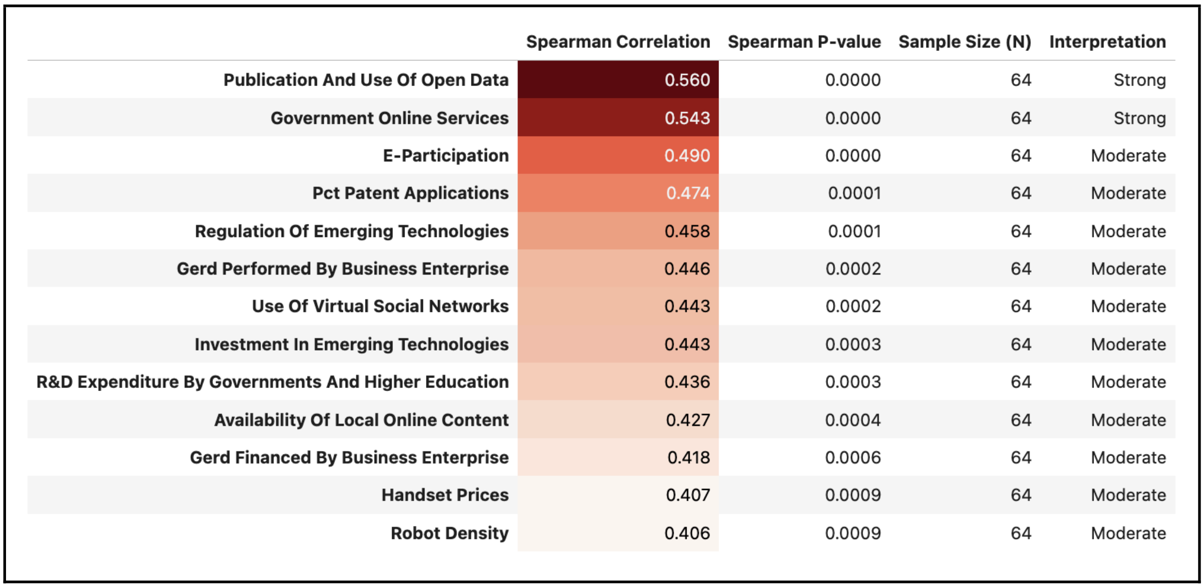


	Spearman Correlation	Spearman P-value	Sample Size (N)	Interpretation
Informational Globalisation, de facto	0.515	0.0000	64	Strong
Cultural Globalisation, de facto	0.373	0.0024	64	Moderate
Cultural Globalisation, de jure	0.318	0.0106	64	Moderate
Political Globalisation, de facto	0.315	0.0113	64	Moderate
Financial Globalisation, de jure	0.313	0.0117	64	Moderate
Informational Globalisation, de jure	0.259	0.0390	64	Weak
Total Globalisation	0.240	0.0565	64	Weak
Financial Globalisation, de facto	0.208	0.0991	64	Weak
Political Globalisation, de jure	0.198	0.1176	64	Weak
Interpersonal Globalisation, de facto	0.170	0.1804	64	Weak
Trade Globalisation, de jure	0.072	0.5697	64	Very Weak
Interpersonal Globalisation, de jure	0.054	0.6715	64	Very Weak
Trade Globalisation, de facto	-0.146	0.2507	64	Weak

Figure 3. Relationship between forms of globalisation and emphasis on 'data'

[13] World Bank Group, 'World Development Indicators'
[14] KOF Swiss Economic Institute, 'KOF Globalisation Index'

Lastly, using the ***Network Readiness Index*** (NRI)[15], **Figure 4** shows a strong relationship between countries mentioning these topics in national AI policy documents and certain digital readiness indicators. In particular, countries that actively publish and promote the use of open data and invest in government online services are the most likely to focus on data-centric topics in their AI policies.



	Spearman Correlation	Spearman P-value	Sample Size (N)	Interpretation
Publication And Use Of Open Data	0.560	0.0000	64	Strong
Government Online Services	0.543	0.0000	64	Strong
E-Participation	0.490	0.0000	64	Moderate
Pct Patent Applications	0.474	0.0001	64	Moderate
Regulation Of Emerging Technologies	0.458	0.0001	64	Moderate
Gerd Performed By Business Enterprise	0.446	0.0002	64	Moderate
Use Of Virtual Social Networks	0.443	0.0002	64	Moderate
Investment In Emerging Technologies	0.443	0.0003	64	Moderate
R&D Expenditure By Governments And Higher Education	0.436	0.0003	64	Moderate
Availability Of Local Online Content	0.427	0.0004	64	Moderate
Gerd Financed By Business Enterprise	0.418	0.0006	64	Moderate
Handset Prices	0.407	0.0009	64	Moderate
Robot Density	0.406	0.0009	64	Moderate

Figure 4. Relationship between digital readiness indicators and emphasis on ‘data’

Therefore, statistically, the country most likely to emphasise data in its AI policies is one that is wealthy, economically stable, knowledge-based, innovation-driven, and publishes and uses open data. The United Nations Conference on Trade and Development (UNCTAD)’s Digital Economy Report 2021 highlights a ‘data-related divide’. It notes that many developing countries are becoming mere providers of raw data to global digital platforms, while having to pay for the digital intelligence generated from their data.[16] Wealthier nations benefit from the active publication and use of open data, while lower-income countries that lack such practices may struggle to engage equally in global data initiatives.[17]

Our findings identify significant disparities in government policies concerning data use in AI frameworks. Notably, some governments prioritise data integration in AI policy, while others do not. This discrepancy is crucial for policymakers advocating for a global AI data governance framework[18], as it suggests that lower and middle-income countries (LMICs) that lack robust open data publication and digital government services may face challenges in engaging fully in global AI data governance discussions.

[15] Portulans Institute (2024), ‘Network Readiness Index 2024’
[16] Unctad (2021), ‘Digital Economy Report 2021’
[17] Ada Lovelace Institute (2021), ‘The data divide’
[18] Slotin J. & McLaren, J., Global Partnership for Sustainable Data (2024), ‘The Global Digital Compact is an opportunity to set a baseline for data governance’

The absence of strong data governance, and limited access to openly published data, in these nations may hinder their ability to leverage AI effectively, potentially exacerbating the existing data divide. Therefore, **global AI data governance frameworks, such as those considered by the UN**[19] **and the Global Partnership on AI (GPAI)**[20], **should incorporate support mechanisms for LMICs**. These mechanisms could include guidance on developing digital infrastructure and facilitating access to high-quality, AI-ready data resources.

Additionally, countries identified at the lower end of this analysis often have limited public data available for AI applications. This presents an opportunity to reinvigorate open data initiatives in these regions, ensuring that new datasets are prepared for AI integration from the outset. Such efforts would enhance the capacity of LMICs to participate in global AI governance and promote equitable access to AI advancements.

[19] United Nations (2024), 'Governing AI for Humanity'
[20] GPAI

2. Which aspects of data are being discussed?

a. Key areas of good AI data governance are overlooked

To gain greater insight into which aspects of data are being mentioned in these policy documents, we mapped ten topics associated with data-centric AI. We found that several of these are highly popular in AI policies: 1) data protection, 2) open data, 3) data ethics, 4) data governance, and 5) data sharing. Others, despite often being mentioned in responsible AI discourse in recent years, are still overlooked: 1) data lineage, 2) data licensing, 3) data transparency, 4) data provenance, and 5) responsible data.

Figure 5 visualises that many individual countries emphasise these topics. Out of **66** countries that mention data, **52** mentioned at least one of these topics and therefore are depicted here. In this diagram, lines connect countries to keywords, and the thickness of each line reflects the total score for that country's mentions of that keyword—the thicker the line, the greater the emphasis.

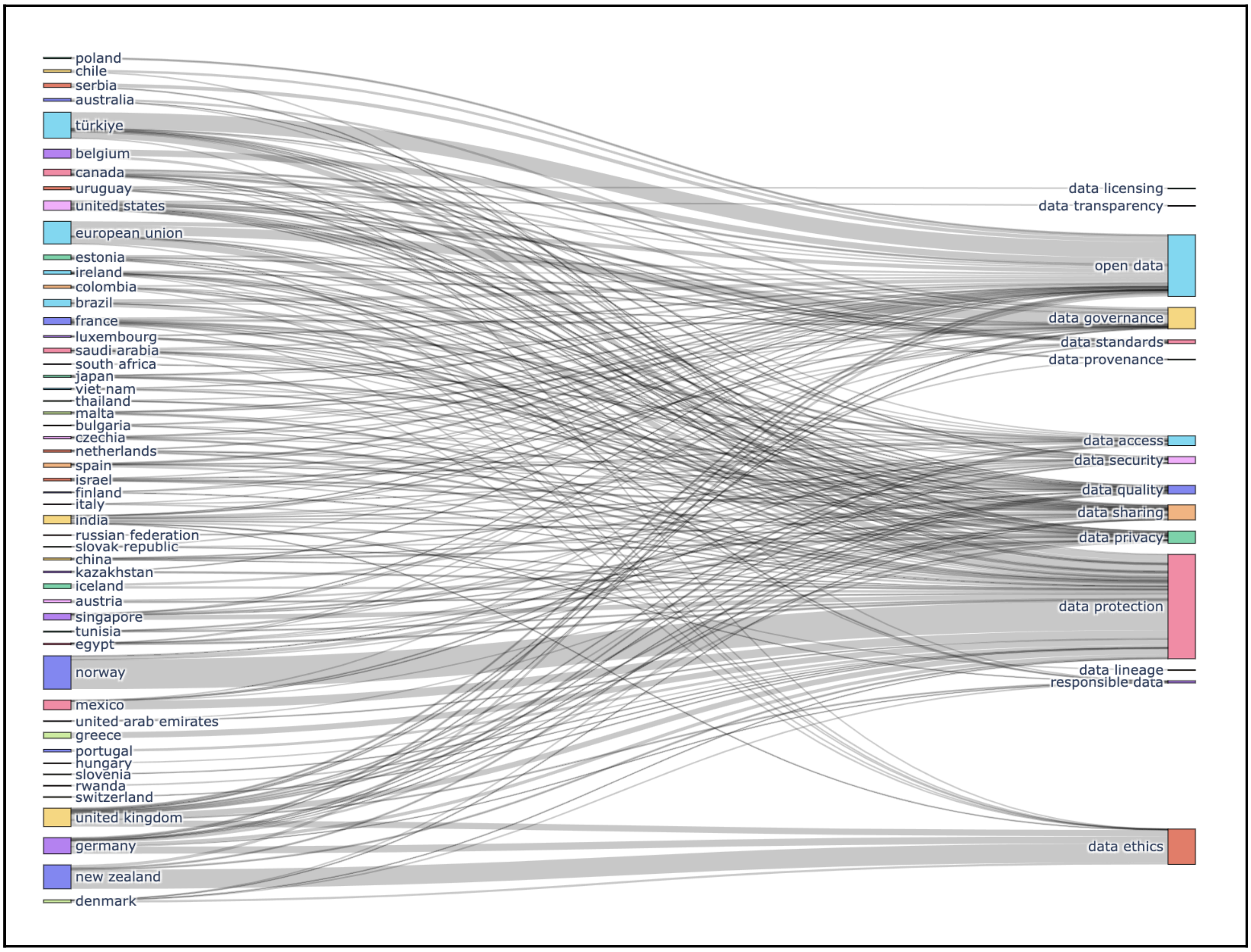


Figure 5. Countries' emphasis on each keyword

Several of these overlooked areas are vital to effective AI policymaking. Understanding data provenance—tracing the origins of data—allows for assessing its quality, reliability and potential biases, which are essential for building trustworthy and fair AI systems.[21] Similarly, data lineage tracks how data has been transformed over time through various processing stages in the AI data lifecycle. When documented as metadata (data about data) and structured using standards like Croissant[22], data provenance and lineage information both enhance transparency in the data used to train, evaluate and deploy AI systems.[23]

Despite their critical importance, AI data standards are under emphasised in policy discussions, with 'data standards' ranking sixth in popularity. Standardisation is essential for ensuring consistency, interoperability and quality across AI systems. Establishing common standards for data formats, metadata schemas, documentation and compliance reporting reduces overheads and allows more organisations to make use of AI advancements.

Standards can emerge from industry consortia, become de facto through widespread use, or, with government participation, be formally established by international organisations like ISO, IEC and ITU. Governments also develop influential standards directly, such as NIST's Artificial Intelligence Risk Management Framework.[24] International efforts like ISO/IEC 8183 provide globally recognised guidelines for managing data throughout its lifecycle in AI applications.[25]

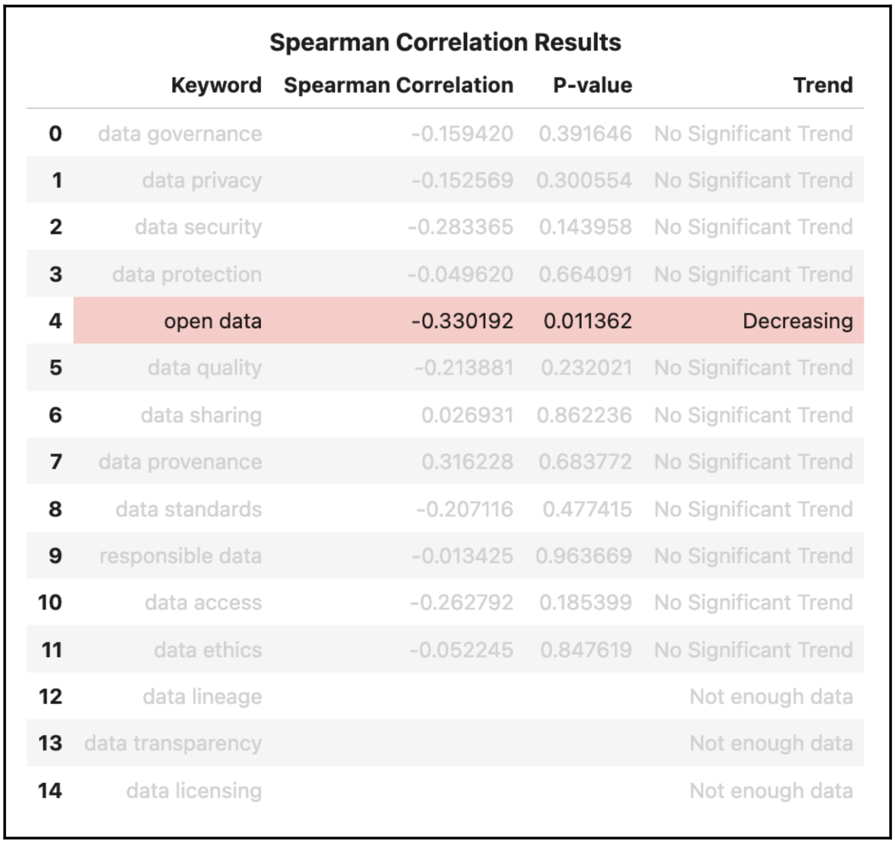
Standardised documentation can address critical aspects of AI datasets that are often underrepresented. Existing tools like data cards[26] and datasheets for datasets[27] can support compliance by detailing datasets' provenance, collection methodologies, intended use cases, and potential biases, among other areas. This documentation, in turn, supports the implementation of regulatory frameworks like the EU's AI Act, which mandates transparency and explainability in high-risk AI systems, making data lineage and provenance information crucial for legal compliance.[28]

Despite the growing focus on responsible AI, our analysis shows that crucial data-related issues—like provenance, lineage, ensuring transparency, proper licensing and setting standards—are still not getting enough attention in AI policies. **To bridge these gaps, policymakers and researchers should invest in data-focused AI tools and adopt strong data documentation practices supported by machine-readable metadata**. This will enhance regulatory oversight and help establish best practices, enabling countries to address these overlooked areas and lead the way in effective AI data governance.

[21] Longpre S. et al, Cornell University (2024), 'Data Authenticity, Consent, & Provenance for AI are all broken: what will it take to fix them?'
[22] Benjelloun, O. et al (2024), 'Croissant Format Specification'
[23] Carey-Wilson, T. et al, ODI (2024), 'Understanding data governance in AI: Mapping governance'
[24] National Institute of Standards and Technology (2023), 'Artificial Intelligence Risk Framework'
[25] ISO (2023), 'ISO/IEC 8183:2023'
[26] Pushkarna, M,, Zaldiva, A., Kjartansson, O., ACM Conference on Fairness, Accountability, and Transparency (2022), 'Data Carts: Purposeful and Transparent Dataset Documentation for Responsible AI'
[27] Gebru, T. et al, CACM (2021), 'Datasheets for Datasets'
[28] EU AI Act, 'Key Issues'

b. Open data is significantly declining in AI policies

Finally, we look to the future by asking whether these topics are increasing or decreasing over time (from 2010 to 2024) using Spearman rank correlation to identify a trend over time. Most of these topics show no evidence of a significant trend over time. However, as shown in **Figure 6**, we see that 'open data'—despite being a popular topic in AI policy—is significantly decreasing over time at the global level. Given recent concerns about the decline of open data sources for AI, this aligns with several alarming trends.



Spearman Correlation Results

	Keyword	Spearman Correlation	P-value	Trend
0	data governance	-0.159420	0.391646	No Significant Trend
1	data privacy	-0.152569	0.300554	No Significant Trend
2	data security	-0.283365	0.143958	No Significant Trend
3	data protection	-0.049620	0.664091	No Significant Trend
4	open data	-0.330192	0.011362	Decreasing
5	data quality	-0.213881	0.232021	No Significant Trend
6	data sharing	0.026931	0.862236	No Significant Trend
7	data provenance	0.316228	0.683772	No Significant Trend
8	data standards	-0.207116	0.477415	No Significant Trend
9	responsible data	-0.013425	0.963669	No Significant Trend
10	data access	-0.262792	0.185399	No Significant Trend
11	data ethics	-0.052245	0.847619	No Significant Trend
12	data lineage			Not enough data
13	data transparency			Not enough data
14	data licensing			Not enough data

Figure 6. Trend in keyword mentions over time

Open data is critical for AI development, enabling affordable training on diverse datasets, enhancing AI model's accuracy and reliability.[29] For instance, open-source AI relies on accessible datasets to develop models that are transparent, reproducible and adaptable to various applications.[30] According to supranational bodies like UNESCO and the EU, open data policies can help to ensure that AI models are built on ethical and transparent sources of data, fostering trust and safety while maximising AI's potential with high-quality, cross-sector information.[31] Through access to public open data, AI systems can

[29] Snaith, B. et al, ODI (2024), 'Policy intervention 4: Ensuring broad access to data for training AI models'
[30] White, M. et al (2024), 'The Model Openness Framework: Promoting Completeness and Openness for Reproducibility, Transparency, and Usability in Artificial Intelligence'
[31] UNESCO (2023), 'Recommendation on the Ethics of Artificial Intelligence'

learn from comprehensive, real-world examples without violating individual privacy or data security.[32]

Yet, amidst growing restrictions on open web data—evident through stricter terms of service and the widespread use of robots.txt to block AI crawlers from accessing high-quality content—the essential data commons for AI development are diminishing.[33] Our results align with similar alarming trends related to diminishing open data resources from governments.

Stefaan Verhulst highlights that stalled open data policies have led to a more restrictive data ecosystem, hindering access to crucial data for public interest uses like research and AI development.[34] For instance, challenges with the EU Open Data Directive—including its narrow focus on economic reuse over broader open data goals[35] and widespread implementation issues[36]—have restricted access to essential public data for AI development, reflecting the alarming decline in open data resources.

The latest Global Data Barometer report (2022) confirms this trend, indicating that while open data agendas persist, their growth has slowed compared to earlier years. The report notes that since 2016, there have been very minimal increases in datasets that are fully machine-readable, openly licensed, freely available, and bulk-accessible.[37] Studies have repeatedly indicated that this stagnation is driven by a combination of insufficient political commitment, fragmented data management, limited data literacy, and a lack of standardised data interoperability frameworks.[38] [39] [40]

Given these observations, it is concerning that open data is receiving less attention in AI policies, especially since it plays a crucial role in addressing imbalances in the AI ecosystem. As more data holders build walls around their data, we are seeing a reduction in the shared resources essential for AI development, leading to a more restrictive environment for AI datasets. **Policymakers need to promote safe and equitable data sharing and explore the use of data access remedies[41] specific to the AI context.**

[32] European Data (2024), ‘The role of Artificial Intelligence in open data: Legal and policy challenges’
[33] Longpre, S. et al (2024), ‘Consent in Crisis: The Rapid Decline of the AI Data Commons’
[34] Velhurst, S. (2024), ‘Are we entering a “Data Winter”?’
[35] Pugh, S A., Open Knowledge (2019), ‘Missed opportunities in the EU’s revised open data and re-use of public sector information directive’
[36] Broomfield, H., Information Polity (2023), ‘Where is open data in the Open Data Directive?’
[37] Davies, T,, Fumega, S., Gray, J. (2022), ‘Global Data Barometer’
[38] World Wide Web Foundation (nd), ‘Open Data Barometer Global Report’
[39] Davies, T,, Fumega, S., Gray, J. (2022), ‘Global Data Barometer’
[40] Adams, R. et al, GPAI (2023), ‘The Role of Government as a Provider of Data for Artificial Intelligence: Interim Report’
[41] Schnurr, D., Bentham House Conference (2023). ‘Data Access Remedies: Regulatory Approaches, Economic Trade-Offs and Information Technology Design’

Concluding remarks

Our analysis highlights significant disparities in how countries emphasise data within their AI policies. Wealthier, innovation-driven nations with strong data governance and open data initiatives are leading the way, while many LMICs risk falling behind. Since AI relies on good data governance, there is a risk that limited compute and talent resources and access to data will create a new digital divide in AI. Additionally, essential areas such as data lineage, provenance, transparency, licensing and responsible data are frequently overlooked. The global decline in focus on open data within AI policies is particularly concerning, as it is critical for developing ethical and effective AI systems.

To address these challenges, we recommend that global AI data governance frameworks support LMICs by enhancing digital infrastructure and expanding access to high-quality data resources for all. Policymakers should prioritise underrepresented areas to bolster regulatory oversight of AI datasets, using tools like the forthcoming AI Data Transparency Index (AIDTI) and other data tools to understand pain points and identify areas for improvement. Developed by the ODI, the AIDTI is a framework designed to assess the transparency of public AI models concerning the data used in their development. Moreover, policymakers and policy researchers must encourage broader data accessibility by reassessing open data policies to ensure government datasets remain available for AI development, thus fostering innovation and promoting responsible AI practices worldwide.

Annex. Methodology

This observatory aims to analyse how ''data' and data-related topics are addressed in global AI policies by examining policy documents from various countries. The methodology comprises data collection, text extraction and preprocessing, keyword identification, TF-IDF score calculation, data normalisation statistical analyses, including Spearman's rank correlation coefficient, and trend analysis. The code and datasets used for each step can be found via this GitHub repository.[42]

Data collection

We sourced **512** AI policy documents from **64** countries and **two** supranational organisations using the OECD.AI Policy repository. These documents include national strategies and emerging AI-related regulations available in PDF and HTML formats. Each document's metadata, such as country of origin, start date, policy initiative ID and policy instrument type, was also collected.

Text extraction and preprocessing

For each document, we extracted textual content using the following steps:

1. **PDF documents:** We utilised the PyPDF2 library to read and extract text from PDF files. To handle various languages, we detected the language of each document using the langdetect library.

2. **HTML documents:** We used the requests library to fetch HTML content and BeautifulSoup for parsing and extracting text from HTML pages.

We ensured that non-textual elements, such as images and scripts, were excluded from the analysis. Additionally, we handled encoding issues and removed any extraneous whitespace or punctuation from the extracted text.

[42] Carey-Wilson, T. (nd), 'Global AI Policy Data Observatory'

Keyword identification

We focused on the keyword ‘data’ and a set of data-related keywords relevant to data-centric AI topics. These keywords were decided upon by the combined input of experts working in the ODI’s data-centric AI program.[43] To account for multilingual documents, we developed a translation dictionary covering the languages present in the dataset. The keywords were translated into each language using this dictionary, enabling consistent keyword matching across different languages.

TF-IDF score calculation

To quantify the emphasis on 'data' and data-related topics in each document, we calculated the Term Frequency-Inverse Document Frequency (TF-IDF) score for each keyword within each document. The TF-IDF score highlights the importance of a keyword in a document relative to its occurrence across all documents.

The TF-IDF score for a keyword k in a document d is calculated as:

$$TF - IDF_{k,d} = TF_{k,d} \times IDF_{k}$$

where:

- **Term Frequency (TF)**: The normalised frequency of the keyword in the document

$$TF_{k,d} = \frac{Number\ of\ times\ keyword\ k\ appears\ in\ document\ d}{Total\ number\ of\ words\ in\ document\ d}$$

- **Inverse Document Frequency (IDF)**: Measures how common or rare the keyword is across all documents

$$IDF_{k} = log(\frac{N+1}{n_{k}})\ +\ 1$$

  - N is the total number of documents
  - n is the number of documents containing keyword k

[43] ODI (2023), ‘Data-centric AI’

Data cleaning and normalisation

After computing the TF-IDF scores, we combined the results from both PDF and HTML documents. We performed the following steps to prepare the data for analysis:

1. **Deduplication:** Removed duplicate entries based on the ‘Policy initiative ID’ column to ensure each policy was represented only once.

2. **Scaling for incomplete years:** For documents from the year 2024, we adjusted the TF-IDF scores to account for incomplete data, using a scaling factor:

$$Scaling\ Factor = \frac{Total\ days\ in\ year}{Days\ elapsed\ in\ 2024}$$

   The scaled TF-IDF score for 2024 documents is:

$$TF - IDF_{scaled} = TF - IDF \times Scaling\ Factor$$

3. **Normalisation per country:** To compare countries equitably, we normalised the TF-IDF scores for each country using the formula:

$$Normalised\ TF - IDF_{c} = \frac{\Sigma TF - IDF_{c}}{D_{c}} \times log\ (D_{c} + 1) \times 1000$$

   - $\Sigma TF - IDF_{c}$ is the sum of TF-IDF scores for country c
   - D_{c} is the number of documents from country c

This normalisation process adjusts the aggregate TF-IDF scores to account for the differing number of policy documents per country. By dividing the total TF-IDF score by the number of documents D_c we obtain the average TF-IDF score per document for each country. The logarithmic scaling factor $log\ (D_c + 1)$ adjusts this average to reflect the document count more appropriately. This ensures that countries with a larger number of documents have their emphasis on data-related topics proportionally represented, without disproportionately dominating the normalised scores. Finally, the normalised TF-IDF scores are multiplied by 1000 to enhance their readability, as the raw scores were predominantly small decimal values.

Statistical analyses and visualisations

1a. Mapping

We visualised the normalised TF-IDF scores per country using choropleth maps. Each country's score reflects its emphasis on 'data' and data-related topics in AI policies. The mapping involved:

- Assigning ISO country codes to match countries with geographical data.
- Generating maps to display the distribution of normalised TF-IDF scores globally.

1b. Spearman correlation analyses

We used Spearman's rank correlation coefficient to examine relationships between country characteristics and keyword emphasis. We chose this over Pearson and linear regression for its robustness to non-linear, monotonic relationships and reduced sensitivity to outliers. This approach captures the strength and direction of associations between normalised TF-IDF scores and the following national indicators, helping to explore how countries emphasise data in their AI policies:

- **Macroeconomic indicators:** Gross National Income (GNI), Gross Domestic Product (GDP) and lending interest rates.
- **Globalisation measures:** Scores from the KOF Globalisation Index (KOFGI), including informational globalisation metrics like patent applications and high-technology exports.
- **Digitalisation indicators:** Scores from the Network Readiness Index (NRI), focusing on open data publication and government online services.

The Spearman correlation coefficient r_s assesses the monotonic relationship between two variables and is calculated as:

$$r_s = 1 - \frac{6\Sigma d_i^2}{n(n^2 - 1)}$$

- d_i is the difference between the ranks of the paired variables for observation i
- n is the number of observations

We interpreted the strength of the correlation based on the absolute value of r_s :

- Very weak: $\left|r_s\right| < 0.1$
- Weak: $0.1 \leq r_s < 0.3$
- Moderate: $0.3 \leq r_s < 0.5$
- Strong: $0.5 \leq r_s < 0.7$
- Very strong: $r_s \geq 0.7$

2a. Sankey diagram

We created a Sankey diagram to visualise the flow of TF-IDF scores from countries to specific data-related keywords. The width of the flow represents the magnitude of the normalised TF-IDF scores for each country-keyword pair.

2b. Trend analysis

Spearman's rank correlation coefficient is well suited for analysing trends over time in time series data, as it effectively captures monotonic relationships while remaining robust to outliers.[44] [45] This makes it a reliable method for identifying coherent trends across variables without requiring linear assumptions, which is essential in contexts like ours, with non-linear or autocorrelated data. To identify trends in the emphasis on data-related keywords over time, we performed a time-series analysis:

- **Aggregation:** For each keyword, we aggregated the normalised TF-IDF scores by year.
- **Trend calculation:** We calculated Spearman's correlation between the sequential years and the aggregated TF-IDF scores for each keyword.

A significant positive r_s indicates an increasing trend, while a significant negative r_s indicates a decreasing trend. We generated a table to visualise the temporal trends of keyword emphasis across the years studied (2010 to 2024).

[44] Lun, D. et al, Journal of Applied Statistics (2023), 'Significance testing of rank cross-correlations between autocorrelated time series with short-range dependence'
[45] Ye, J., Cloud Computing and Big Data conference (2015), 'Time Series Similarity Evaluation Based on Spearman's Correlated Coefficents and Distance Measures'

Limitations

This research has some limitations. First, for part 1b, the correlation analyses in each category were confined to an average of scores from 2013 to 2021 as KOFGI data was available only across this period. This ensured consistency across all analyses. Second, although we developed and tested a translation dictionary to account for language variations, some nuances may have been lost, since the authors are primarily English speakers and the study encompassed **71** countries with a wide range of languages.

Lastly, while analysing these **71** countries provides a considerable sample, there are at most **138** additional countries whose AI policies (if they exist) were not included in the OECD.AI repository. At the time of publication, this is the largest publicly available collection of AI policies, but future research should incorporate a larger number of countries to validate whether these findings hold across a more comprehensive sample.