SPES | Sustainability performances, evidence & scenarios

D4.1
Report on mapping, harmonising and integrating novel data sources for research purposes

January 2024

Funded by European Union's Horizon Europe Programme under Grant Agreement No. 101094551

Authors

Veronica Ballerini – Researcher of the SPES Project, University of Florence
Davide Beraldo – Researcher of the SPES Project, University of Amsterdam
Chiara Bocci – Researcher of the SPES Project, University of Florence
Lisa Braito – Researcher of the SPES Project, University of Florence
Roberta Milana – Interning researcher of the SPES Project, University of Amsterdam
Emilia Rocco – Researcher of the SPES Project, University of Florence
Martin Trans – Researcher of the SPES Project, University of Amsterdam

Contributors and peer reviewers:
Mario Biggeri, University of Florence; Paolo Brunori, London School of Economics & Political Science; Jeroen de Vos, University of Amsterdam; Andrea Ferrannini, University of Florence; András Gábos, TARKI; Luca Lodi, University of Florence; Fabrizia Mealli, European University Institute; Stefania Milan, University of Amsterdam; Gregor Zens, IIASA.

Acknowledgements:
The authors would like to thank all SPES partners for their inputs provided in several SPES meetings and communications. We would like to thank also all interviewees involved in our data collection activities for having shared with us their opinion and insights on the topic.

Cite as:
Ballerini, V., Beraldo, D., Bocci, C., Braito, L., Milana, R., Trans, M. (2024). Report on mapping, harmonising and integrating novel data sources for research purposes. SPES Report no. 4.1, SPES project – Sustainability Performances, Evidence and Scenarios. Florence: University of Florence. Available at: https://www.sustainabilityperformances.eu/publications-deliverables/

Disclaimer

This Report D4.1 for the project SPES has been prepared by the University of Amsterdam and the University of Florence as part of Task 4.1 “Map complex and novel data sources and methods” of Work Package 4.

This task has allowed SPES research partners to identify and provide an evaluation of complex and novel data sources and associated methods that can be repurposed for the development of innovative measurement framework.

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

Table of contents

PART I

Mapping Innovative Data Sources Across the Many Dimensions of Sustainability Research

Davide Beraldo, Martin Trans, Roberta Milana

1. Introduction

The study of sustainability spans across several academic fields, such as environmental studies, developmental economics, social and demographic sciences, and geography. Consequently, it relies on data pertaining to diverse domains, such as environmental data, demographic data, economic data, geo-spatial data. Traditional sources of data relevant for sustainability research include national as well as transnational, public as well as private bodies such as statistical offices; universities and research institutes; meteorological services; energy agencies; global development institutes. A number of socio-technical transformations have radically changed the research landscapes in the past decade. The inexorable transformation of different aspects of reality into quantified data has significantly impacted research in several fields, including the many fields linked to sustainability studies.

This process of *datafication* (Kitchin, 2022; van Dijk, 2014) presents unprecedented opportunities for researchers, enabling the collection of data at a scale, granularity and temporality unimaginable until recently. This enhances researchers' ability to monitor, understand, model and predict complex systems involved in the study of sustainability transition - such as ecological, economic and social systems. However, this condition of data abundance comes with important shortcomings and potential pitfalls (boyd and Crawford 2012). The sheer volume of big data requires the adaptation of development of sophisticated and often resource-expensive methods and procedures. The quality of data becomes an issue of concern because data repurposed from non-conventional sources often do not obey established methodological principles such as representativeness and coherence. Moreover, new ethical considerations around privacy, ownership and security arise around data generally not generated for research purposes.

The purpose of this part of the report is to provide insights into what type of innovative sources of data can be adopted in order to study the many dimensions of sustainability transition. Defining innovation is a complex task, as novelty is a contextual and relational property. It has a temporal dimension, related to an element of recency, as well as a domain-specific dimension, in that what is innovative in a certain context is not innovative in another. This general consideration is exacerbated in the case of methodological innovation within academic research because of the proliferation of academic fields, each characterised by different methodological approaches and trajectories. Due to its marked interdisciplinarity and wickedness, sustainability as a phenomenon complicates the picture even further.

Most existing contributions to the issue of methodological innovation in relation to sustainability are strictly linked to the Sustainable Development Goals (SDGs) framework (UN, 2015). In line with the approach outlined in SPES's D2.1 (Biggeri et al., 2023), this report builds upon a different framework revolving around 5 pillars: Productivity; Equity; Environmental Sustainability; Participation & Empowerment; Human Security,

In order to answer the question 'What innovative data are currently leveraged to study the many dimensions of sustainability?', this report adopted a systematic approach, mapping data sources claimed to be innovative within a large sample of academic sources, collected following an expansive operationalisation of the concept of sustainability and leveraging Large Language Models (LLM) for the purpose. It is important to stress that, despite dealing with literature and despite its systematic character, this is by no means a systematic literature review on the topic of innovation in sustainability research, and these results should be taken as indications. The many studies included in our 'fishnet' are used here as entry-points to extract and classify different types of innovative data sources, and to provide an extensive list of illustrations on the issue. The concrete method adopted, that could be considered innovative in itself, is outlined in the following section.

2.Methods

This part of the report is itself an instructive example of the potential and the limitations associated with innovative approaches to data. Academic texts are in themselves a rather traditional secondary source of information. The rise of academic search engines that allows to retrieve a quasi-exhaustive set of literature sources associated with a list of keywords opened up the way for systematic literature reviews, aimed at mapping a certain domain of knowledge as extensively as possible. Moreover, the steady advancement in Natural Language Processing witnessed in the past decade, which recently scale-jumped with the popularisation of Large Language Models (LLMs), allows to transform a large collection of academic texts into a dataset for quali-quantitative analysis.

The main objective of the literature collected is to generate a substantial amount of data sources to be classified and inspected for illustrative purposes. In order to identify a sample of innovative data sources, the present research adopted a pragmatic approach based on two principles: temporal delimitation and researchers' self-definition. First, we used 2019 as a cut-off point to provide a lower bound to our literature search; the year was chosen because it corresponds to the publication of the report conducted by the MAKSWELL project (Van den Brakel, 2019), a Horizon2020-funded research project that elaborated upon the promises and limitations of non-traditional data sources for the assessment of the SDGs. The assumption behind this choice is that, given the highly dynamic character of methodological innovation, it is more useful to focus on the years following this report. Second, in order to identify literature dealing with innovative data sources, we used the researchers' own claims of novelty contained within the texts. Consequently, we searched for texts containing keywords related to the innovative and novel character of the data and measurements adopted. By delegating the definition of what is innovative to the claims made within the texts we collected, we circumvented the otherwise daunting task of assessing the methodological innovativeness of a number of fields of which the authors are not necessarily experts.

A core component of our methodology has been the adoption of LLMs, and OpenAI's GPT more specifically, as a tool to assist in the exploration of a large set of academic texts. This involved information extraction, classification, selection and enrichment tasks. While constituting an exciting source of methodological innovation and exploration, it is important to remark that LLMs are not infallible. Their performance, in particular, is highly dependent on the quality of input data, as well as on effective prompting. Evaluating their performance however can be a daunting task, considering the many intervening factors and the non-deterministic nature of their outcome. The development of reliable (and affordable) research protocols for exploiting the opportunities of LLMs is a current vibrant area of debate and investigation. While pursuing the ambition to automatically map and classify innovative data sources with the assistance of tools such as search engines and LLMs, we encountered a number of obstacles. These limitations included challenges associated to API limitations and costs (e.g., how much text it is possible to process at a time; unpredictability in the output of non-deterministic systems) as well as assumptions related to homogeneity and standardisation of data entries (e.g., the proportion of PDF available for automated download; the proportion of book-length publications and theses).

The data collection and analysis followed a number of phases. First, we needed to design an appropriate set of queries that would allow us to collect a large corpus of texts related to innovative data sources used to study different dimensions of sustainability. Second, we relied on GPT to extract specific snippets from the texts, corresponding to the data sources used and related information. Third, we relied on existing reviews to create a classification scheme. Fourth, we instructed a model to automatically classify each data source into a data category. Fifth, we compiled a list of examples that illustrates how different categories of data sources have been used by recent studies to investigate the different dimensions of sustainability.

2.1.Designing the query

Our queries resulted from the combination of two components: a variable theme-specific component, and a fixed innovation-related component.

The fixed innovation-related component of the query is composed of different formulations aimed at distinguishing established from new types of data sources. This consisted in including a wide series of fixed phrases and combinations of literal keywords formatted in a series of OR logical statements:

> "novel data source" OR "new data source" OR "innovative data source" OR "nonconventional data source" OR "non-conventional data source" OR "novel source of data" OR "new source of data" OR "innovative source of data" OR "nonconventional source of data" OR "non-conventional source of data" OR "novel data sources" OR "new data sources" OR "innovative data sources" OR "nonconventional data sources" OR "non-conventional data sources" OR "novel sources of data" OR "new sources of data" OR "innovative sources of data" OR "nonconventional sources of data" OR "non-conventional sources of data" OR "novel measure" OR "new measure" OR "innovative measure" OR "nonconventional measure" OR "non conventional measure" OR "novel measure" OR "novel measures" OR "new measures" OR "innovative measures" OR "nonconventional measures" OR "non conventional measures" OR "novel measures" OR "novel measurement" OR "new measurement" OR "innovative measurement" OR "nonconventional measurement" OR "non-conventional measurement" OR "novel measurement" OR "novel measurements" OR "new measurements" OR "innovative measurements" OR "nonconventional measurements" OR "non-conventional measurements" OR "novel measurements"

The theme specific components emerged from a conceptual exercise in which the goal was to map a sufficient number of dimensions of sustainability, relevant to the project. The starting point for designing this component of the query were the above-mentioned 'five pillars' that constitute the main reference of SPES's theoretical framework. As a pre-emptive measure to address the semantic overlap between the pillars, we introduced a cross-pillar element taking the epistemological status of a pillar itself, in our querying. This ensured that important dimensions of economy, energy, research and innovation were not left unaddressed seeing as they are a part of a more intersectional dimension.

Each pillar represents a broad, albeit essential component of sustainability. For this reason, for each pillar we agreed upon three dimensions that best represent areas of investigation relevant for the SPES project. The name of the pillar itself has also been included together with these additional three dimensions per pillar. Each dimension has been further decomposed into three more specific sub-dimensions, each associated to one or more keywords to account for lexical, orthographic or stylistic variations. These keywords (88 in total) represented the theme-specific component of our queries (see Figure 1). It is important to stress that neither the relation between pillars and dimensions nor the relation between dimensions and phenomena is to be understood in terms of exhaustiveness. As already explained, the scope of this literature search was extensive rather than exhaustive. Nonetheless, grounding our literature search on this hierarchical unfolding of the concept of sustainability allowed us to craft an effective ‘fishnet’ that ensured to cover a broad and diverse semantic field.

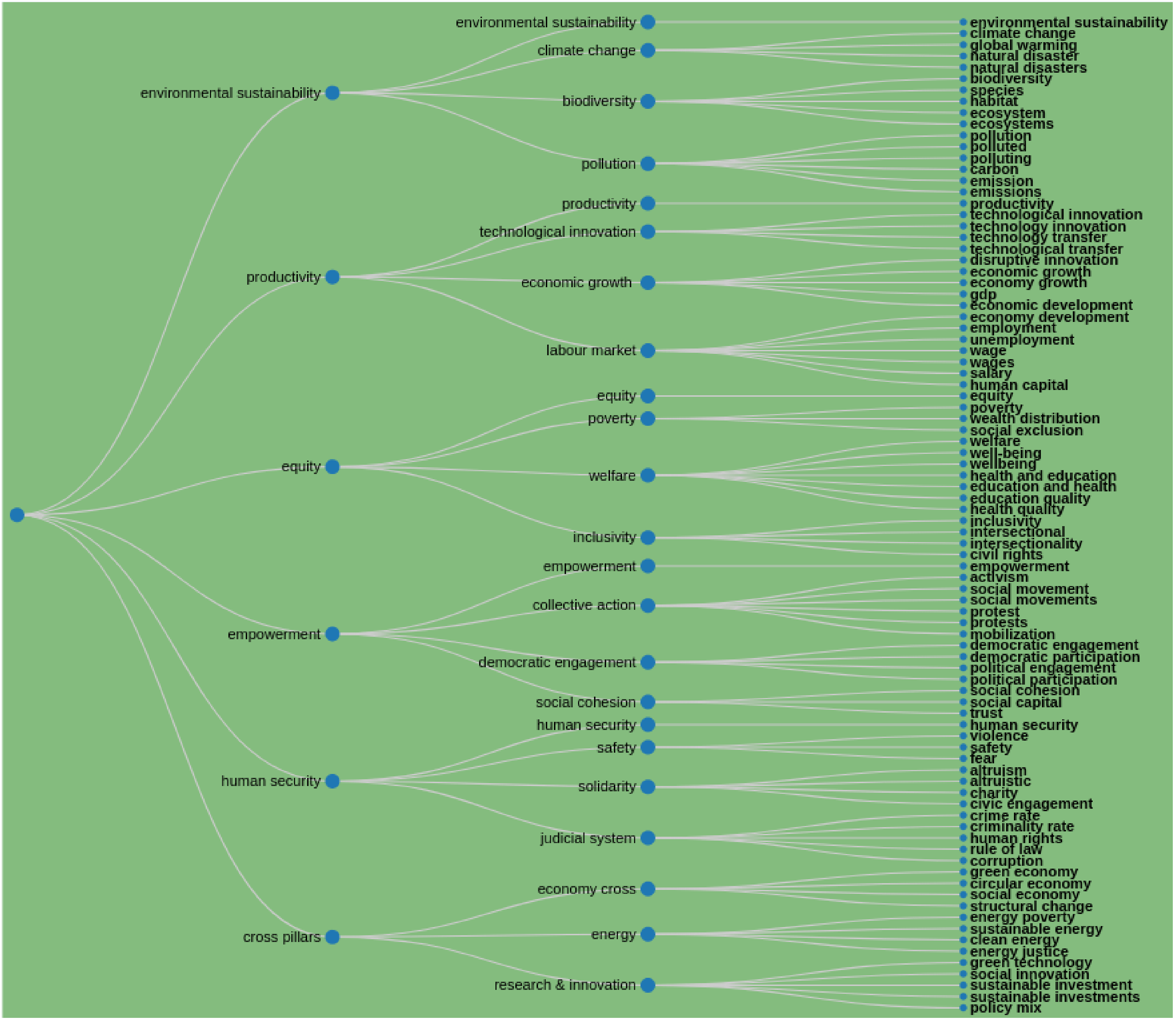

Figure 1: Hierarchical Operationalisation of Pillars, Dimensions and Keywords

Our queries resulted from the combination of the 88 theme-specific component and the innovation-related component with an AND operator, hence resulting in a total of 88 queries. Before proceeding with launching the queries and collecting the resulting literature sources, we manually tested several queries and evaluated their results for relevance.

Another important operative choice taken in the querying process: that of adopting a literal search approach in Google Scholar, by enclosing the search terms into double quotes. This implies that any text that does not literally match the necessarily specific formulation of our query is not included in our corpus. Such operative choice is likely to have left out relevant literature sources, but it was deemed necessary after realising that a semantic search (i.e., a search that attempts to match the contextual meaning of a query) was generating too many false positives. Since our goal, as mentioned, was not that of producing an accurate representation of methodological innovation, but rather to generate as many useful examples as possible, false positives are more of a problem than false negatives.

2.2.Collecting a corpus of texts

For simplicity, we decided to use Google Scholar as a search engine, since it is widely used by the scholarly community and its logic allowed us to easily scale query execution. The choice to focus on a single search engine has likely left out some relevant source from our search; however, including other academic search engines (such as Scopus or WebOfScience) would have significantly increased the overall effort, since it would have required developing additional *ad hoc* parsers.

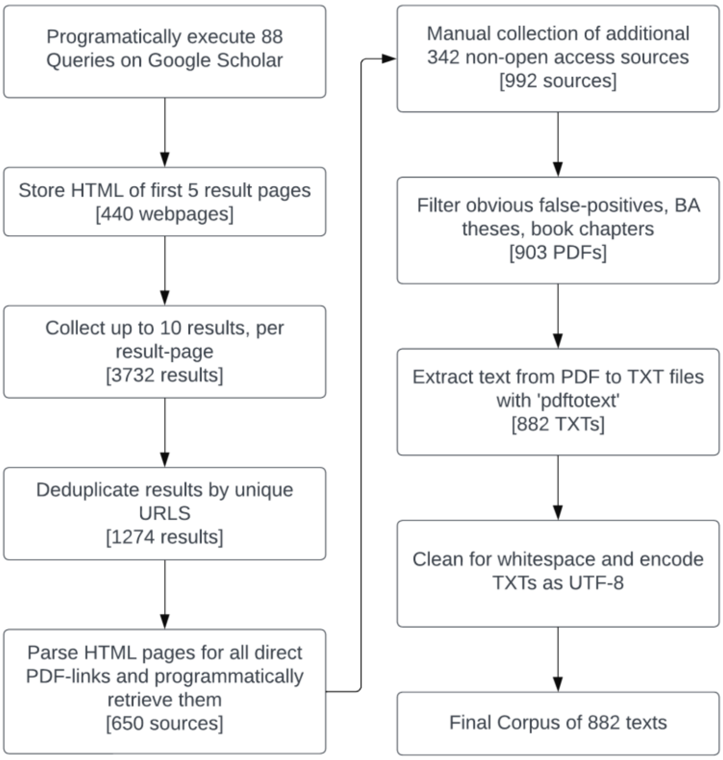

Figure 2: Flowchart of steps taken to collect corpus

The 88 queries reported above have been launched programmatically through a custom script. For each query, we stored the HTML code of the first 5 result pages, each containing up to 10 results (see Figure 2). Hence, we collected up to 50 literature results for each query / keyword, albeit some queries returned less than the maximum. The queries returned a total of 3,732 literature sources, which included a substantial number of duplicates. After deduplicating, we were left with 1,274.

The HTML pages resulting from the Google Scholar queries have been parsed in order to partially automate the literature sources collection. Roughly half of the literature sources retrieved were available in open access and the corresponding entry displayed a direct PDF link. This allowed us to automatically download such entries programmatically. The remaining sources included both non-open access sources, requiring institutional authentication, and open access sources that did not provide a direct PDF link. For these sources we proceeded with manual collection.

Several other factors reduced our sample of literature sources further. This includes filtering obvious false-positives, BA theses and chapters from books that we could not retrieve. Some publications provided unforeseen technical difficulties in relation to formatting or did not include text-elements that allowed us to extract the text from the PDF file. This reduced our number of sources in our corpus from 903 PDF files to a final of 882 text files.

2.3.Extracting information

In order to extract relevant information from the corpus of texts we relied on OpenAI's APIs, through which it is possible to interrogate their Large Language Models (GPT3.5-turbo and GPT4-turbo). We developed a script interrogating OpenAI APIs, tasked to extract from each literature source snippets of text, as literal as possible, related to: each data source mentioned in the text; the phenomenon for which such data has been adopted; the method used to analyse such data; a one-liner elaborating on the use of the data source. Moreover, for each text, the model also extracted a number of metadata: authors; year of publication; title; journal; field; a brief summary.

After comparing the outcome of both GPT4-turbo and GPT3.5-turbo, we concluded that they performed very similarly on our extraction tasks, hence we opted for the more cost-effective GPT3.5-turbo. This however comes with an important limitation related to the necessity to process smaller chunks of text, since this model has a smaller maximum length of text that it can process. Consequently, we restricted each query to the model to 13,000 characters of text. In order to minimise the risk that the section of the text mentioning data and methods would be excluded from the analysis, while keeping costs and time for data collection under control, we decided to limit the information extraction to the two first chunks of 13,000 characters from each text. It is necessary to mention that a certain amount of texts consisted of entire books / theses, hence the chunking might have left important aspects of these sources out of the scope of the analysis.

The extraction task resulted in 2,945 entries, corresponding to each individual combination of data sources, phenomena and methods identified by the model within the texts.

2.4.Creating a classification scheme

For the creation of a classification scheme, we also relied on GPT. In this case, we adopted the dialogue version ChatGPT, and more in particular we created a CustomGTP (based on the most performant GPT4), which is a version of ChatGPT that is tailored to a specific task and topic through a customisation of the background knowledge that the model builds upon in order to provide its answers. We trained our CustomGPT with recent publications on the topic of innovative and non-traditional data for the study of sustainability collected through a cursory literature search (Del Río Castro et al., 2021; Del Vecchio et al., 2018; ElMassah & Mohieldin, 2020; Fritz et al., 2019; Ilieva & McPhearson, 2018; Kharrazi et al., 2016; van den Brakel et al., 2019; Weber et al., 2021).

Initial attempts to generate a purely inductive classification scheme, one emerging from the bottom-up based on the list of data sources extracted by the model, did not yield satisfactory results. Consequently, we relied on the texts that we included in the CustomGPT, and asked the model to consolidate the type of data sources mentioned as innovative in these texts into a codebook, that in turn could be used to classify an extensive list of data sources. After a few iterations, we settled for the following coding scheme:

1. *Social Media Data*: User-generated data from social media platforms which can be analysed for trends, public opinion, and behaviour related to sustainability (e.g., Twitter feeds on environmental issues, Instagram posts on recycling practices)
2. *Citizen Science Data*: Data collected by the general public through participation in scientific research (e.g., Environmental monitoring, biodiversity tracking, community-led surveys)
3. *Urban Big Data*: Large volumes of data generated within urban settings, public transportation systems, urban sensors, etc. (e.g., Traffic flow data, public transportation data, urban infrastructure usage)
4. *Remote Sensing and Satellite Data*: Data collected via satellites and aerial imaging, useful for monitoring environmental changes, land use, and urban development (e.g., Deforestation tracking, urban heat islands, coastline changes)
5. *Mobile Phone and Telecommunication Data*: Data generated from mobile phone usage and telecommunication networks, useful for understanding human mobility and communication patterns (e.g., Movement patterns during natural disasters, communication trends in response to public awareness campaigns)
6. *Internet of Things Data*: Data from interconnected devices and sensors, often used for monitoring and optimising resource use (e.g., Smart energy metres, water quality sensors in rivers and lakes)
7. *Administrative and Governmental Data*: Data collected by government bodies and agencies, often as part of their administrative functions (e.g., Census data, municipal waste management records, public transport usage statistics)

8. *Financial Transaction Data*: Data related to economic activities, including transactions, investments, and consumption patterns (e.g., spending patterns on sustainable goods, investment trends in green technologies)
9. *Crowdsourced Data*: Data collected through voluntary contributions from a large number of people, typically via the internet (e.g., crowdsourced environmental monitoring, community reporting platforms)
10. *Web Data*: Data collected from the world wide web, either available for download or obtained through HTML scraping. (e.g., sustainable companies websites, open data repositories on climate data, blogs on social justice activism.)
11. *Other*: Data that does not fit the descriptions of codes from 1 to 10.

2.5.Classifying information

For the classification task, we also relied on OpenAI API's and GPT3.5-turbo model. We followed a few-shot approach, one where the model is trained on a small number of examples - those provided with the codebook. Since LLMs perform inherently better on smaller chunks of text, provided that the instructions are sufficiently clear, we queried OpenAI API for each extracted data source individually. Previous attempts at classifying the entire list of data sources at once through ChatGPT's interface with GPT4, indeed, provided unsatisfactory results. In particular, the model tended to focus mostly on the head and tail of the list, hence incorrectly classifying as 'Other' a substantial number of entries in the central part of the list, even for cases where the classification task seemed very straightforward.

The inclusion of the residual code 'Other' as an explicitly provided classification option turned out to be a critical decision. We ran the classification with both versions of the codebook, including or excluding the option to classify poorly fitting content as 'Other'. Running the classification that included 'Other' as an explicit code resulted in a substantial number of entries being classified as 'Other' (1,291), as well as a few entries being classified with codes not included in the codebook (69). A cursory inspection of the entries classified as 'Other' confirmed the low incidence of false negatives in the classification task (i.e., entries that should have been classified with other codes), and made it clear that their high number was largely due to some of the literature sources being a weak match to our search criteria (e.g., the adoption of the keywords 'novel' and 'new' in relation to recency and not to novelty; the query matching the reference list rather than the body of the text). This confirmed that the coding scheme including the category 'Other' was a better choice than the one which did not, since it would have included a large proportion of false positives (i.e., entries that should have not been classified with a certain code). An inspection of a sample of the classified entries confirmed the low incidence of such false positives, which would have significantly altered the distributions of data categories and would have constituted an important source of noise for the selection of illustrative examples.

2.6.Selecting illustrative examples

We developed a dataset comprising 1,585 entries, each representing a data source identified as innovative in academic texts. These sources were categorised based on their use in studying various dimensions of sustainability. To illustrate the diversity of these data sources, we organised examples according to the dimensions that we used to operationalise sustainability, according to the pillars laid out by Biggeri et al. (2023) in the SPES framework. For each sustainability dimension, we selected one example publication, corresponding to one of the three most frequently occurring data categories within that dimension. Given that each combination of dimension and data category could consist of several dozen entries, we employed GPT to assist in making an informed selection of the most innovative and informative examples for each combination of data category and sustainability dimension. This process involved iterating over the 23 identified dimensions of sustainability and the top three data categories within each. We prompted the GPT API to extract what it deemed the most innovative example in each case. Initially, our classified dataset was based solely on excerpts that included the data source, the studied phenomenon, the methodology, and an overview of the use of sources. However, we found that the resulting list of examples lacked the depth and detail necessary for our purposes. To address this, we enriched and refined each example using ChatGPT. We input the entire text corresponding to each example into the model and prompted it to elaborate on how the data source was used innovatively to study the specific phenomenon. This step provided the necessary polish to our examples, making them more suitable for demonstrating innovative uses of data sources in sustainability research. Each extracted elaboration was subsequently reviewed and adjusted for accuracy.

3.Results

3.1.Overview of data categories

The following table (Table 1) lists the occurrences of data sources for each data category included in the classification scheme. The count of data sources per each data source provides an overview of the distribution. The data categories are additionally utilised to differentiate and structure the illustrative examples provided in the subsequent section.

Data Categories	Count
Administrative and Governmental Data	362
Social Media Data	288
Remote Sensing and Satellite Data	248
Financial Transaction Data	131
Citizen Science Data	122
Web Data	120
Urban Big Data	115
Internet of Things Data	82
Mobile Phone and Telecommunication Data	77
Crowdsourced Data	40

Table 1: Count of Data Sources across Data Categories

The following network graph (Figure 3) summarises the relation patterns between data categories (in blue) and the sustainability pillars (in green), based on which study used which type of data. The connections (links) between a data category and a sustainability pillar are based on the frequency with which data sources within that category have been used to study phenomena related to the respective pillar. Essentially, the graph visualises which types of data are most commonly employed in research pertaining to each pillar of sustainability.

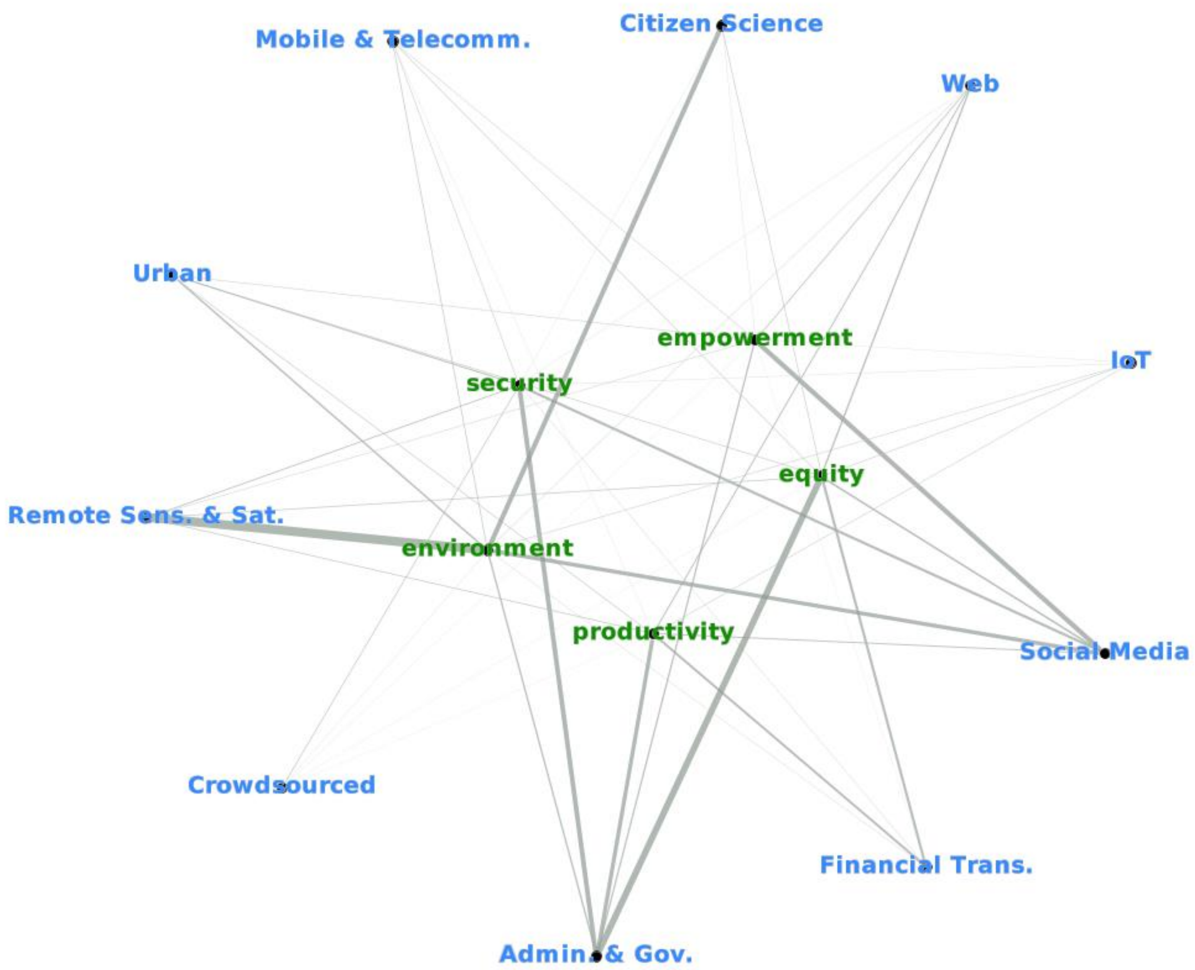

Figure 3: Network graph of relational patterns between data categories and sustainability pillars

Administrative and government data is relied upon within the pillars of equity, productivity and security, while Remote Sensing and Satellite data are strongly tied to the environmental pillar. Social media data is relied comparatively more upon within the pillars of empowerment, environment and security. Citizen Science data is strongly connected with the environmental pillar. Crowdsourced data is generally linked with the security pillar.

The following heatmap (Figure 4) plots the co-occurrences of the data categories and dimensions of sustainability. The darker the colour of the cell, the more often a data source classified within a certain category has been employed to study the corresponding dimension.

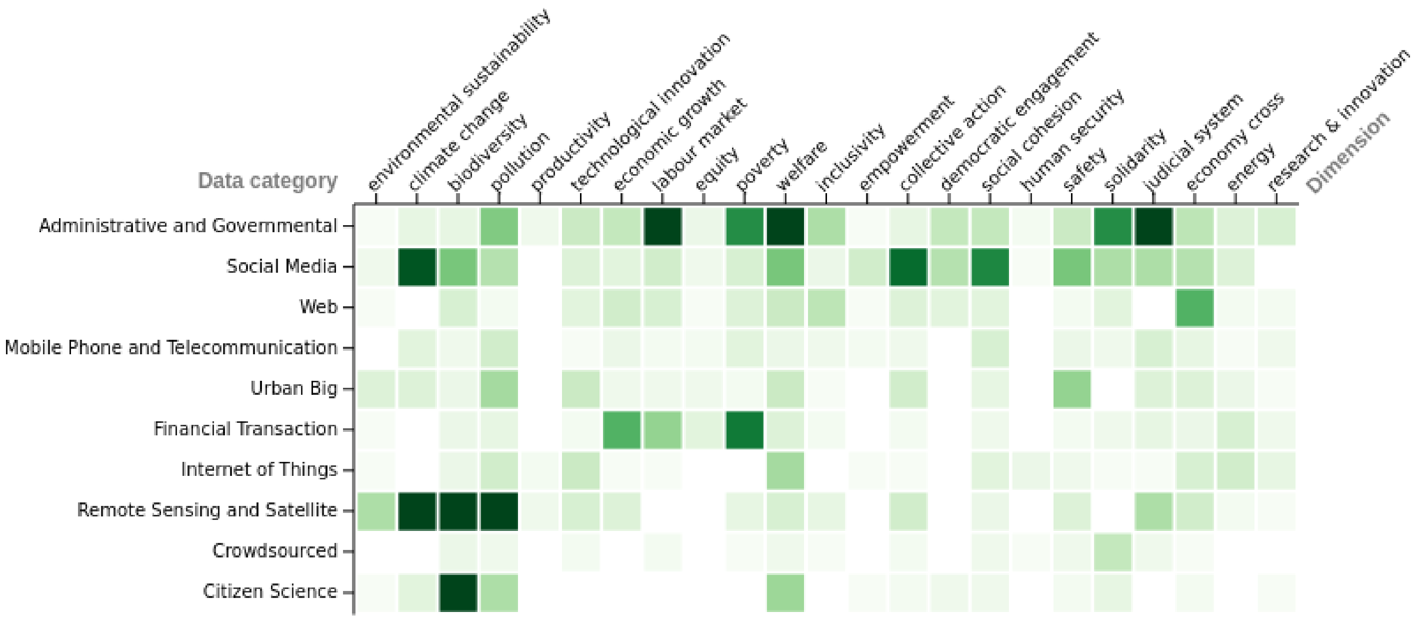

Figure 4: Heatmap of co-occurrence of data categories and dimensions of sustainability

Within our innovative corpus, Administrative and Governmental data have been mostly used to study the labour market, welfare and the judicial system. Social Media data are framed as particularly useful to study climate change, collective action and social cohesion. Financial Transactions data are closely tied to studies of economic growth and poverty. Remote Sensing and Satellite data are used to study predominantly sustainability dimensions such as environment, climate change, biodiversity and pollution. Citizen Science data is generally used to study biodiversity, welfare and poverty.

3.2.Illustrative examples

The goal of this section is to provide a list of examples that illustrate how different types of innovative data sources are used to study several dimensions of sustainability. This section is intended as a navigational map that can be traversed by the reader according to their own domain-specific research interests. The examples are organised by dimension, and each is grouped according to the corresponding pillar. For each dimension, each example illustrates one of the three most frequently adopted categories of data. Table 2 is offered as a readers-guide, providing an overview of the illustrative examples[1].

Within the pillar of environmental sustainability, the chosen studies utilise remote sensing and satellite data (Guo et al., 2020), big data from urban infrastructures (Chaudhuri & Bose, 2020), and social media data (Mouttaki et al., 2022) to address challenges in land cover accuracy, disaster management, and ecosystem services mapping. These approaches exemplify the potential of combining high-resolution images, field surveys, and deep learning models to enhance accuracy and efficiency in environmental analysis.

For the productivity-pillar, the chosen studies have used remote sensing for crop yield forecasting (Yli-Heikkilä et al., 2021), administrative data for studying the impact of innovation on firm performance (Bedford et al., 2021), and web data for analysing economic patterns (Zhang et al., 2022). These publications emphasise the potential of integrating various data sources, including Earth Observation data and financial transaction data, to develop more nuanced and comprehensive analytical models.

Relating to the pillar of equity, the chosen publications employ financial transaction data to explore wealth inequality (Waltl, 2020) and crowdsourced narratives to understand the experiences of older adults during the pandemic (Kyröläinen et al., 2022). These innovative approaches demonstrate the use of detailed transaction records and linguistic analysis to gain deeper insights into social issues.

The pillar of empowerment is addressed through the analysis of social media data (Al Tamime et al., 2022), urban big data for traffic patterns (Rahman et al., 2021), and governmental data for political survey inferences (Ghitza & Gelman, 2020). These studies illustrate the potentials of utilising large-scale datasets and advanced algorithms for enhancing understanding in fields such as gender disparities and democratic engagement.

In research related to the pillar of security, we noted the use of IoT data for real-time traffic safety (Li, 2021) and social media data for mental health community analysis (Tang et al., 2021). These approaches represent opportunities within a potential shift towards more dynamic and comprehensive data utilisation, contributing to fields like traffic management and mental health support.

Finally, cross-pillar studies, such as those analysing start-up firm dynamics (Couture & Houle, 2020) and energy models using IoT data (Happle, 2020), highlight the interdisciplinary nature of contemporary research. These publications demonstrate the effectiveness of combining different data sources, like administrative records and environmental sensors, to create more robust and insightful analyses across various domains.

[1] It is important to disclose that these illustrative examples originate from an extraction assisted by chatGPT, which was reviewed for accuracy and modified accordingly.

Table 2: Map overview of illustrative examples

Section	Pillar	Dimension	Data Category	Data Source
3.2.1	Environment	Environmental sustainability	Remote Sensing and Satellite Data	FROM-GLC30 2017 dataset
		Climate change	Urban Big Data	Images from smart urban infrastructures
		Biodiversity	Social Media Data	Flickr images
		Pollution	Administrative and Governmental Data	GHG Platform India
3.2.2	Productivity	Productivity	Remote Sensing and Satellite Data	Earth Observation data
		Technological Innovation	Administrative and Governmental Data	Intellectual Property Government Open Data
		Economic Growth	Web Data	Online news data
		Labour Market	Social Media Data	LinkedIn Internal Audit Function data
3.2.3	Equity	Equity	Financial Transaction Data	Limited entry permit transactions database
		Poverty	Financial Transaction Data	National accounts and Household Finance Consumption Survey
		Welfare	Crowdsourced Data	Wellness diaries
		Inclusivity	Web Data	Google Trends data
3.2.4	Empowerment	Empowerment	Social Media Data	Online advertising data
		Collective Action	Urban Big Data	Microwave Vehicle Detection System
		Democratic Engagement	Administrative and Governmental Data	Voter Registration Databases
		Social Cohesion	Administrative and Governmental Data	Administrative dataset of civil complaints
3.2.5	Security	Human Security	Crowdsourced Data	Barometer Initiative
		Safety	Internet of Things Data	Connected Vehicles and smartphone sensors
		Solidarity	Social Media Data	Online Depression-Focused Communities data
		Judicial System	Administrative and Governmental Data	Russian criminal court decisions
3.2.6	Cross-pillars	Economy Cross	Administrative and Governmental Data	National Accounts Longitudinal Microdata File
		Energy	Internet of Things Data	Wearable sensors
		Research & Innovation	Internet of Things Data	Wi-Fi signal detector

3.2.1. Environment

Environmental sustainability: *Remote Sensing and Satellite Data* (Guo et al., 2020)
The publication's innovative use of the FROM-GLC30 2017 dataset in studying land cover accuracy at a continental scale lies in several key aspects. Firstly, it utilises a novel method based on watershed sampling units, which is rare in continental-scale studies. This method, more representative than traditional pixel-based methods, offers a new approach to understanding land cover accuracy. Additionally, the study covers a vast area, including the Pan-Third Pole Area, which spans parts of Asia, Europe, and Africa. The use of high-resolution remote sensing images from Google Earth, coupled with field surveys, allows for a comprehensive and accurate assessment of land cover. The study's results, which include detailed analyses of accuracy variations across different regions and land cover types, contribute significantly to the field by providing insights into the spatial variation of land cover accuracy and laying the groundwork for improved applications of global land cover datasets.

Climate change: *Urban Big Data* (Chaudhuri & Bose, 2020)
The publication's use of images from smart urban infrastructures to study disaster-hit environments in Central Mexico (2017 earthquake) is considered innovative for several reasons. Firstly, it identifies a novel source of data, namely images from smart urban infrastructures, which are crucial for effective disaster management decision-making. This approach is innovative as it involves collecting and analysing images from the affected areas, providing a unique and direct perspective of the disaster's impact. Secondly, the study utilises advanced technology in the form of a deep learning convolutional neural network to analyse this unique dataset. This method allows for a more detailed and sophisticated analysis of the images, enhancing the ability to understand and respond to the crisis effectively. The use of TensorFlow for image classification further exemplifies the cutting-edge nature of this research. The findings from this study are significant as they offer a new way of utilising technology for effective disaster response, demonstrating both the effectiveness and technical viability of this approach in crisis scenarios.

Biodiversity: *Social Media Data* (Mouttaki et al., 2022)
The use of Flickr images to study Cultural Ecosystem Services (CES) mapping is innovative because it applies a deep learning model to automate the classification of natural and human elements relevant to CES. Traditional methods of quantifying and mapping CES are complex due to their intangible nature, and manual content analysis of large numbers of photographs is time-consuming. This study utilised a convolutional neural network architecture to analyse over 29,000 photographs from the Lithuanian coast, employing hierarchical clustering to group these photographs. This approach not only accurately categorised the photographs but also saved a significant amount of manual work (approximately 100 km worth). By enabling the analysis of large numbers of digital photographs, this method expands the analytical tools available to researchers, facilitating the quantification and mapping of CES on a large geographic scale. This automated assessment and

mapping technique can potentially inform urban planning and enhance nature reserve management, making it a significant advancement in the field of CES research.

Pollution: *Administrative and Governmental Data* (Dhingra et al., 2019)
The publication's use of GHG Platform India for studying Greenhouse Gas Emission Estimates from the AFOLU (Agriculture, Forestry and Other Land Use) sector in India at the subnational level is innovative for several reasons. Firstly, the estimates are primarily based on aggregated secondary data collected by the Vasudha Foundation from nationally recognised documents and reports, involving interactions with external experts for inputs on data availability and emission estimation approaches. This approach facilitates a comprehensive and collaborative assessment. Secondly, the study provides time-series emission estimates for the period 2005 to 2015 at the state (sub-national) level, adhering to the 2006 IPCC Guidelines for National GHG Inventories and other internationally accepted guidance. This level of granularity in data allows for more precise and localised understanding of emission sources and trends. Lastly, the study's findings are publicly available, aiming to enhance discussion, address data gaps, and contribute to policy dialogue on India's GHG emissions profile and related policies. This transparency and commitment to engaging a broader audience in climate change discussions represent a significant step forward in environmental research and policy development.

3.2.2. Productivity

Productivity: *Remote Sensing and Satellite Data* (Yli-Heikkilä et al., 2021)
The publication's innovative use of Earth Observation (EO) data in crop yield forecasting lies in its comprehensive and automated approach, which integrates EO data, administrative data, agro-meteorological data, and historical crop statistics. This method streamlines what was previously a labour-intensive process, enabling more efficient and cost-effective forecasting. The system utilises advanced artificial intelligence techniques to provide pre-harvest yield predictions for key cereal crops. Moreover, the implementation of this EO-based forecasting method represents a significant shift towards automated, data-driven agricultural statistics, potentially offering more accurate and timely insights for sustainable agriculture and informed decision-making.

Technological innovation: *Administrative and Governmental Data* (Bedford et al., 2021)
The publication's innovative approach lies in its application of the Intellectual Property Government Open Data 2019 to extend and replicate prior findings on the relationship between innovation, future profitability, and stock returns in the context of Australian firms. This study diverges from previous U.S.-focused research, addressing the unique Australian setting where innovation and patenting activities are less prevalent. By integrating patent data with firm-level financial and market data, it provides new insights into the impact of innovation on firm performance in a different legal and economic environment. This contributes to a more nuanced understanding of how innovation influences profitability and stock market performance in various contexts, offering valuable implications for policymakers and investors in countries with different levels of innovation intensity.

Economic growth: *Web Data* (Zhang et al., 2022)
The publication's innovative use of online news data from Baidu, Jinritoutiao, Fengguangwang, Sina News, Souhu News, and Wangyi News to study spatial-temporal patterns of economic output values in Yinzhou, Ningbo, China, lies in its pioneering approach in a region where this methodology is relatively unexplored. While previous studies have used online news and search engine data to forecast economic patterns and stock market trends, particularly in the U.S., the application of this technique in China, especially for understanding spatial-temporal dynamics, is limited. This study is innovative in its exploration of these patterns and the development of models like Generalized Linear Model (GLM) and Geographically Weighted Regression (GWR) to predict economic dynamics using online news data. This approach not only demonstrates the potential of online news in forecasting economic trends in China but also suggests a new dimension to augment traditional surveillance systems with a spatial-temporal perspective, which could be applicable to larger geographical areas.

Labour market: *Social Media Data* (Renschler et al., 2023)
The publication's innovative use of LinkedIn Internal Audit Function (IAF) data from Revelio Labs in studying IAF competency and financial reporting failures is notable for several reasons. Firstly, it introduces a new data source to the IAF literature, transitioning from traditional survey-based data to a more robust and objective dataset. The LinkedIn IAF data, aggregated by Revelio Labs, provides several advantages. It allows for a more objective sample composition, as it does not rely on voluntary participation, and offers more detailed and accurate data collection on individual personnel on a longitudinal basis. This shift enables a more comprehensive understanding of IAF competency and its impact on financial reporting quality, leading to insights that were previously unattainable with survey-based approaches.

3.2.3. Equity

Equity: *Financial Transaction Data* (Ayers & Chan, 2020)
The publication's use of the limited entry permit transactions database to study the ownership of Hawai'i longline fleet limited entry permits is innovative primarily due to its comprehensive and detailed approach. The dataset includes records of 6,171 transactions from 1991 to 2017, encompassing permit transfers, renewals, and changes in permit status. Notably, the study differentiates between 'permit holders' (entities whose names are on the permits) and 'permit applicants' (entities that applied for the permits). This distinction is crucial because it acknowledges the complexity of ownership in this context, where an entity may be an individual vessel owner, a business entity (like an LLC or a corporation), or an agent representing others. The separate analysis of permit holders and applicants allows for a more nuanced understanding of ownership changes and trends in the fleet, which is essential for effective fisheries management and policy-making.

Poverty: *Financial Transaction Data* (Waltl, 2020)
The innovative aspect of using National accounts data and top-tail-adjusted HFCS (Household Finance and Consumption Survey) data to study wealth inequality lies in their ability to create a more accurate and comprehensive view of wealth distribution. This method effectively addresses the limitations of traditional surveys, which often underrepresent the wealthiest individuals. By integrating national accounts data with survey data, and making adjustments for the top tail of the wealth distribution using models like Pareto and Generalized Pareto, the approach offers a more accurate representation of the wealth distribution across different sections of society. This enhanced accuracy is crucial for understanding the true extent and nature of wealth inequality, allowing for more effective policy-making and economic analysis.

Welfare: *Crowdsourced Data* (Kyröläinen et al., 2022)
The use of the CoSoWELL project to study written narratives by North American older adults before and during the pandemic is considered innovative for several reasons. Firstly, it provides a unique corpus of English narratives written by this demographic, capturing their experiences both before and during the pandemic. This corpus is enriched with detailed lexical and syntactic annotations at the word level, along with demographic and psychological participant metadata from the CoSoWELL survey. Secondly, the project's design and initial data collection predate the COVID-19 pandemic, allowing for a broader scope of research beyond just the psychological outcomes of the pandemic. The comprehensive goal of the project is to study changes in language use as a function of age, social isolation, and perceived loneliness, using personal life stories to gain insights into the memory functioning and overall mental and emotional well-being of older adults. This approach is innovative as it combines linguistic analysis with demographic and psychological data to provide a multifaceted understanding of the effects of ageing and social factors on older adults.

Inclusivity: *Web Data* (Boss et al., 2023)
The publication's innovative use of Google Trends to study refugee flows from 150 origin countries to the EU27 lies in its development of monthly refugee flow forecasting models that leverage machine learning and high-dimensional data, including digital trace data from Google Trends, and integrates a wide range of classical predictors capturing socio-economic, labour, and monetary statistics. This approach is particularly innovative because it allows for near real-time availability of data, which is a significant advantage over traditional methods that often suffer from data availability and publishing lags. For instance, the ensemble model, which combines Random Forest and Extreme Gradient Boosting algorithms and is based exclusively on Google Trends variables, consistently outperforms other models for forecasting horizons between 3 to 12 months. Especially in the top-20 corridors with significant refugee flows, this Google Trends Index (GTI)-only specification surpasses all other models, highlighting its potential as a tool for refugee forecasting in large-flow corridors.

3.2.4. Empowerment

Empowerment: *Social Media Data* (Al Tamime et al., 2022)
The publication's innovative approach lies in its utilisation of aggregate anonymous online advertising audience estimates from platforms like Facebook, Google, and LinkedIn. This method provides real-time data on demographic distributions, particularly gender disparities, in the professional domain. Traditional data sources such as surveys or censuses often lack the geographical and temporal resolution, especially in lower-income countries. By leveraging these digital platforms' data, the study offers more immediate and detailed insights into gender inequalities, filling existing data gaps with a novel, real-time, and large-scale data source. This approach is particularly valuable in analysing gender-related aspects of Sustainable Development Goal 5, which focuses on gender equality and women's empowerment.

Collective action: *Urban Big Data* (Rahman et al., 2021)
The publication's use of the Microwave Vehicle Detection System (MVDS) to study hurricane evacuation traffic patterns is considered innovative due to its development of a comprehensive data pipeline that incorporates automated data quality checking, data imputation, and spatiotemporal visualisation of large-scale traffic data. This approach allowed for a detailed understanding of the changes in traffic patterns during Hurricane Irma's evacuation. By collecting MVDS data from four major highways in Florida, which handled the majority of the evacuation traffic, the study provided valuable insights into the network-wide spatiotemporal evacuation traffic patterns. This methodology represents a significant advancement as it leverages large-scale, real-time data, which was previously underutilised, to enhance the understanding of the extent and distribution of evacuation traffic, thereby aiding transportation agencies in better planning and managing such scenarios.

Democratic engagement: *Administrative and Governmental Data* (Ghitza & Gelman, 2020)
The publication's innovative use of Voter Registration Databases (VRDs) for improving inferences from political surveys lies in its ability to leverage large-scale political variables such as party registration and past voting behaviour, which are free of overreporting bias or endogeneity between survey responses. The developed process utilises multilevel regression and poststratification to produce vote choice estimates for the presidential election, projecting these estimates to millions of registered voters in a post-election context. This approach results in stable and reasonable inferences across demographic subgroups and small geographic areas, down to the county or congressional district level. This methodology addresses the limitations of traditional exit polls and enhances the precision and reliability of political survey inferences.

Social cohesion: *Administrative and Governmental Data* (Wang & Liu, 2022)
The publication's innovative aspect lies in its comprehensive analysis of parking issues in Brisbane, Australia, using a large-scale administrative dataset of civil complaints. This approach provides a unique, real-world perspective on parking problems, differing from traditional studies that often rely

on smaller-scale surveys or theoretical models, with the use of spatial regression modelling. By examining spatial and temporal patterns of parking complaints in both inner and outer city areas, the study offers a nuanced understanding of how parking issues vary across different urban settings and times, emphasising the need for sustainable mobility and liveability in suburban neighbourhoods. This data-driven approach enables a more targeted and effective policy response, considering the distinct needs of various urban areas and contributing to sustainable urban planning.

3.2.5.Security

Human security: *Crowdsourced Data* (Paladini & Molloy, 2019)
The publication's use of the Barometer Initiative in Colombia to study inclusive monitoring mechanisms in the implementation of the peace process is innovative due to its approach to more inclusive monitoring. It emphasises the opening up of peace process implementation to a broader range of actors, including those who were previously marginalised. This inclusivity is crucial as it deals with the transformation of commitments into norms, institutions, policies, and actions, addressing the challenges of implementation in a complex, evolving political and social landscape. The Barometer Initiative thereby contributes to a more sustainable and participatory peace process by ensuring diverse voices and concerns are incorporated and addressed throughout the implementation phase.

Safety: *Internet of Things Data* (Li, 2021)
The publication's innovative approach in the realm of real-time traffic safety studies lies in its integration of Connected Vehicles (CV) and mobile sensing techniques. This approach represents a significant shift from traditional methods that relied heavily on infrastructure-based data, limited by fixed-location devices and prone to environmental influences. By leveraging vehicle-based data through CV and mobile sensors, the study offers a more dynamic and comprehensive analysis of traffic conditions. Key innovations include using smartphone sensors for detecting vehicle manoeuvres, semi-supervised learning algorithms to utilise massive unlabelled data efficiently, and real-time crash likelihood prediction models based on CV emulated data. Additionally, the research introduces novel surrogate safety measures for pedestrian and bicycle safety analysis and employs data fusion techniques to enhance model accuracy, illustrating the potential of these technologies in proactive traffic management and safety applications.

Solidarity: *Social Media Data* (Tang et al., 2021)
The publication's innovative approach lies in its comparative analysis of user behaviours in managed and unmanaged Online Depression-Focused Communities (ODCs) in China. Utilising deep learning for text classification and econometrics for behavioural analysis, it offers novel insights into how community management styles impact user participation and support provision. This research enhances the understanding of ODC user dynamics and aids in designing and managing these communities more effectively, potentially improving support for individuals with depression. This

approach is particularly innovative as it bridges the gap in understanding the differences in user behaviour between different types of ODCs, which was previously underexplored.

Judicial system: *Administrative and Governmental Data* (Kudryavtsev, 2021)
The publication's innovative approach lies in its use of an extensive dataset from Russian criminal court decisions between 2009 and 2012 to analyse the link between homicides and holidays. This study diverges from previous research, primarily US-focused, by examining this phenomenon in Russia, thus offering external validity. It challenges the 'broken promises theory,' which suggests marginalised individuals show greater violence during holidays. Surprisingly, the research finds that well-integrated individuals (employed, first-time offenders) display a higher propensity for homicide during holidays than marginalised groups, contradicting existing theories and sparking new discussions on the nature of holiday-related violence.

3.2.6.Cross-pillars

Economy cross: *Administrative and Governmental Data* (Couture & Houle, 2020)
The use of the National Accounts Longitudinal Microdata File (NALMF) in this study to analyse Canadian start-up firms from 2005 to 2013 is considered innovative due to its comprehensive and longitudinal approach. This dataset, created by linking various administrative data sources, allowed for a detailed analysis of start-up firms over time, focusing on aspects like survival rates, gender of ownership, and labour productivity. The study's ability to track cohorts of new firms over an extended period and across different industries provided unique insights into the dynamics and performance of start-up firms in Canada, particularly highlighting differences based on the gender of ownership. This approach is noteworthy for its depth and breadth of analysis, providing a richer understanding of the start-up landscape in Canada.

Energy: *Internet of Things Data* (Happle, 2020)
The publication's innovation in using wearable Internet of Things (IoT) sensors in Singapore for studying occupant presence and behaviour in urban building energy models lies in its novel approach to collecting and analysing large-scale, real-time environmental data. This approach significantly advances the field of Urban Building Energy Models (UBEM) by providing detailed insights into occupant behaviour and environmental exposure. The use of wearable IoT sensors allows for the gathering of granular data on air-conditioning exposure and other environmental factors, which are crucial for creating accurate and diverse occupant behaviour models. This data-driven method enables more precise simulations of energy demand and supply on an urban scale, leading to better-informed decisions for urban planning and energy system design. This approach contrasts with traditional methods that often rely on standardised, less specific assumptions, highlighting its innovative contribution to the field.

Research & innovation: *Internet of Things Data* (Ding et al., 2019)
The publication presents an innovative approach to traffic state monitoring using Wi-Fi signal data, a novel data source in this context. The system is distinguished by its use of an Internet of Things (IoT)-based Wi-Fi signal detector, which is solar-powered and collects Wi-Fi signal data. A filtration and mining algorithm extracts traffic state information, such as travel time, traffic volume, and speed, from the data. This method was tested in a practical field test on a major corridor in China. The results, compared with traditional loop data, showed a high correlation and accuracy in traffic speed and volume estimation. The system's distinctiveness lies in its sustainable, cost-effective deployment, and its ability to provide real-time traffic state monitoring with a novel data source, positioning it as a supplementary tool to traditional traffic monitoring systems.

4. Conclusive remarks

This part of the report presented a mapping of innovative data sources used in various disciplines for the study of sustainability. The research process relied on Large Language Models to enable the distant reading of a large corpus of academic texts. These academic texts have been collected based on an expansive operationalisation of the concept of sustainability. The data sources mentioned in the texts have been extracted and classified according to a coding scheme and served as basis to present illustrative examples.

Our results show that different types of data sources have been claimed as innovative in relation to the different pillars of sustainability, as outlined in the SPES framework (Biggeri et al., 2023). In many cases, methodological innovation derives from identifying sources of data previously unavailable, generated as a by-product of other activities or with other goals in mind than sustainability research. However, novelty is also associated to the innovative use or re-combination of existing data sources, as well as from the integration of different types of data, as testified by the relative prevalence of Administrative and Governmental Data and by the several studies integrating different types of data. These considerations raise two important methodological issues. First, data originally developed for engineering or marketing goals, and only secondarily repurposed for research, come with important issues of representativity, validity and reliability. Second, combining data of different types from heterogeneous sources poses a challenge of integration and harmonisation. For this reason, the second part of the report will focus on methods for data integration and harmonisation.

This report has been based on exploratory research that, despite its limitations, yields great potential for mapping methodological innovation in a field such as sustainability studies. Traditional literature review techniques fall short in capturing the multi-dimensional and inter-disciplinary character of sustainability. Inspired by the procedure laid out in this report, future research can cope with some of its limitations and expand its scope. For example, using more specialised Large Language Models would allow to take advantage of broader context windows. Moreover, it would have great value not to limit the analysis to data, but to include methods of analysis in the scope of the mapping.

References

Al Tamime, R., Fatehkia, M., Weber, I., & Kashyap, R. (2022). Leveraging online advertising data for measuring the sustainable development goals: Applications for gender gaps and SDG 5.

Ayers, A. L., & Chan, H. L. (2020). Rights-Based Management, Competition, and Distributional Equity in Hawai'i's Largest Commercial Fishery. International Journal of the Commons, 14(1), 262–277. https://doi.org/10.5334/ijc.996

Bedford, A., Ma, L., Ma, N., & Vojvoda, K. (2021). Future profitability and stock returns of innovative firms in Australia. Pacific-Basin Finance Journal, 66, 101508. https://doi.org/10.1016/j.pacfin.2021.101508

Biggeri, M., Ferrannini, A., Lodi, L., Cammeo, J., & Francescutto, A. (2023). The "winds of change": The SPES framework on Sustainable Human Development (SPES Project – Sustainability Performances, Evidence and Scenarios. SPES Working paper no. 2.1,). University of Florence. https://www.sustainabilityperformances.eu/publications-deliverables/

Boss, K., Groeger, A., Heidland, T., Krueger, F., & Zheng, C. (2023). Forecasting Bilateral Refugee Flows with High-dimensional Data and Machine Learning Techniques (Working Papers 1387). Barcelona School of Economics. https://EconPapers.repec.org/RePEc:bge:wpaper:1387

boyd, danah, & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. Information, Communication & Society, 15(5), 662–679. https://doi.org/10.1080/1369118X.2012.678878

Chaudhuri, N., & Bose, I. (2020). Application of Image Data Analytics for Immediate Disaster Response. Proceedings of the 21st International Conference on Distributed Computing and Networking, 1–5. https://doi.org/10.1145/3369740.3372729

Couture, L., & Houle, S. (2020). Survival and performance of start-ups by gender of ownership: A Canadian cohort analysis. Statistics Canada= Statistique Canada. https://epe.lac-bac.gc.ca/100/201/301/weekly_acquisitions_list-ef/2020/20-37/publications.gc.ca/collections/collection_2020/statcan/11f0019m/11f0019m2020013-eng.pdf

Del Río Castro, G., González Fernández, M. C., & Uruburu Colsa, Á. (2021). Unleashing the convergence amid digitalization and sustainability towards pursuing the Sustainable Development Goals (SDGs): A holistic review. Journal of Cleaner Production, 280, 122204. https://doi.org/10.1016/j.jclepro.2020.122204

Del Vecchio, P., Mele, G., Ndou, V., & Secundo, G. (2018). Open Innovation and Social Big Data for Sustainability: Evidence from the Tourism Industry. Sustainability, 10(9), 3215. https://doi.org/10.3390/su10093215

Dhingra, S., Singh, D., & Mehta, R. (2019). Greenhouse Gas Emission Estimates from AFOLU (Agriculture, Forestry and Other Land Use) Sector in India at the Subnational Level. GHG Platform India Report – Vasudha Foundation. https://www.ghgplatform-india.org/wp-content/uploads/2023/03/AFOLU-Methodology-Note-GHG-Phase-III.pdf

Ding, F., Chen, X., He, S., Shou, G., Zhang, Z., & Zhou, Y. (2019). Evaluation of a Wi-Fi Signal Based System for Freeway Traffic States Monitoring: An Exploratory Field Test. Sensors, 19(2), 409. https://doi.org/10.3390/s19020409

ElMassah, S., & Mohieldin, M. (2020). Digital transformation and localizing the Sustainable Development Goals (SDGs). Ecological Economics, 169, 106490. https://doi.org/10.1016/j.ecolecon.2019.106490

Fritz, S., See, L., Carlson, T., Haklay, M., Oliver, J. L., Fraisl, D., Mondardini, R., Brocklehurst, M., Shanley, L. A., Schade, S., Wehn, U., Abrate, T., Anstee, J., Arnold, S., Billot, M., Campbell, J., Espey, J., Gold, M., Hager, G., … West, S. (2019). Citizen science and the United Nations Sustainable Development Goals. Nature Sustainability, 2(10), 922–930. https://doi.org/10.1038/s41893-019-0390-3

Ghitza, Y., & Gelman, A. (2020). Voter Registration Databases and MRP: Toward the Use of Large-Scale Databases in Public Opinion Research. Political Analysis, 28(4), 507–531. https://doi.org/10.1017/pan.2020.3

Guo, Z., Wang, C., Liu, X., Pang, G., Zhu, M., & Yang, L. (2020). Accuracy Assessment of the FROM-GLC30 Land Cover Dataset Based on Watershed Sampling Units: A Continental-Scale Study. Sustainability, 12(20), 8435. https://doi.org/10.3390/su12208435

Happle, G. (2020). Data-driven occupant presence models for urban building energy simulation [PhD Thesis, ETH Zurich]. https://www.research-collection.ethz.ch/bitstream/handle/20.500.11850/438825/3/Thesis_Happle_26871_w_cover_wo_CV.pdf

Ilieva, R. T., & McPhearson, T. (2018). Social-media data for urban sustainability. Nature Sustainability, 1(10), 553–565. https://doi.org/10.1038/s41893-018-0153-6

Kharrazi, A., Qin, H., & Zhang, Y. (2016). Urban Big Data and Sustainable Development Goals: Challenges and Opportunities. Sustainability, 8(12), 1293. https://doi.org/10.3390/su8121293

Kitchin, R. (2022). The data revolution: A critical analysis of big data, open data & data infrastructures (Second edition). SAGE.

Kudryavtsev, V. (2021). Do the Marginalized Kill More during Holidays? SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3855539

Kyröläinen, A.-J., Gillett, J., Karabin, M., Sonnadara, R., & Kuperman, V. (2022). Cognitive and social well-being in older adulthood: The CoSoWELL corpus of written life stories. Behavior Research Methods, 55(6), 2885–2909. https://doi.org/10.3758/s13428-022-01926-0

Li, P. (2021). Real-Time Traffic Safety Evaluation in the Context of Connected Vehicles and Mobile Sensing [PhD Thesis, University of Central Florida]. Electronic Theses and Dissertations, 2020-. https://stars.library.ucf.edu/etd2020/1144

Mouttaki, I., Bagdanavičiūtė, I., Maanan, M., Erraiss, M., Rhinane, H., & Maanan, M. (2022). Classifying and Mapping Cultural Ecosystem Services Using Artificial Intelligence and Social Media Data. Wetlands, 42(7), 86.
https://doi.org/10.1007/s13157-022-01616-9

Paladini, B., & Molloy, S. (2019). More inclusive monitoring of peace agreement implementation: Barometer Initiative in Colombia. Navigating Inclusion in Peace Processes," Edited by Andy Carl. Special Issue, Accord, 28, 33–35.

Rahman, R., Roy, K. C., & Hasan, S. (2021). Understanding Network Wide Hurricane Evacuation Traffic Pattern from Large-scale Traffic Detector Data. 2021 IEEE International Intelligent Transportation Systems Conference (ITSC), 1827–1832.
https://doi.org/10.1109/ITSC48978.2021.9564480

Renschler, M., Ahn, J., Hoitash, R., & Hoitash, U. (2023). Internal Audit Competency and Financial Reporting Quality: Evidence from LinkedIn Human Capital Data. SSRN Electronic Journal.
https://doi.org/10.2139/ssrn.4335236

Tang, J., Yao, X., & Yu, G. (2021). Exploring the Online Behavior of Users of Online Depression-Focused Communities: Comparing Communities with Different Management Types. Psychology Research and Behavior Management, Volume 14, 1707–1724.
https://doi.org/10.2147/PRBM.S323027

UN. (2015). Transforming our world: The 2030 Agenda for Sustainable Development (A/RES/70/1).
https://sustainabledevelopment.un.org/content/documents/21252030%20Agenda%20for%20Sustainable%20Development%20web.pdf

van den Brakel, J., Buelens, B., Curier, R., Daas, P., Gootzen, Y., de Jong, T., Puts, M., & Tennekes, M. (2019). Aspects of existing databases, traditional and non-traditional data sources and collection of good practices (The MAKSWELL Project). Statistics Netherlands (CBS).
https://www.makswell.eu/attached_documents/output_deliverables/deliverable_2.1.pdf

van Dijck, J. (2014). Datafication, dataism and dataveillance: Big Data between scientific paradigm and ideology. Surveillance & Society, 12(2), 197–208.
https://doi.org/10.24908/ss.v12i2.4776

Waltl, S. R. (2020). Multidimensional wealth inequality: A hybrid approach toward distributional national accounts in Europe. INEQ Working Paper Series, 16, 1–62.

Wang, S., & Liu, Y. (2022). Parking in inner versus outer city spaces: Spatiotemporal patterns of parking problems and their associations with built environment features in Brisbane, Australia. Journal of Transport Geography, 98, 103261.
https://doi.org/10.1016/j.jtrangeo.2021.103261

Weber, I., Imran, M., Ofli, F., Mrad, F., Colville, J., Fathallah, M., Chaker, A., & Ahmed, W. S. (2021). Non-traditional data sources: Providing insights into sustainable development. Communications of the ACM, 64(4), 88–95.
https://doi.org/10.1145/3447739

Yli-Heikkilä, M., Wittke, S., Kokkinen, M., & Partala, A. (2021). CropYield—Towards pre-harvest crop yield forecasts with satellite remote sensing: Final report (p. 28). Natural Resources Institute Finland (Luke).
https://jukuri.luke.fi/handle/10024/547998

Zhang, Y., Sun, H., Gao, G., Shou, L., & Wu, D. (2022). Developing Spatio-Temporal Approach to Predict Economic Dynamics Based on Online News. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.4276222

Appendix

This section provides an illustration of the workflow, presenting an overview of the input given to the model and the output returned.

Information extraction

Prompt passed with "system" role:
"0) You are a research assistant.
1) You will get as input the text of an academic article which reference, among other things, data sources.
2) You will identify a list of data sources that the text claims to be innovative.
3) For each innovative data source, you will be asked to extract three types of information: the data source itself; the phenomenon that the data source is used to study; the method used to analyse the data; an explanation of how data are used innovatively. And some additional metadata.
4) The types of information are related to each other: the data sources are used to analyse some phenomena based on a certain method of analysis.
5) For each type of information, you will be asked to provide a literal description, extracted from the text.
6) If you do not find any of the information, please store an N/A value in the same data structure. Always keep the data structure consistent.
7) Collect also the metadata.
8) You will not make up any information."

Prompt passed with "user" role:
[chunk of 13,000 characters from a specific text]

Function's fields description:
{"data_source": {"type": "string", "description": " Literal description of the data mentioned in this text"},
"phenomenon": {"type": "string", "description": "Literal description of the phenomenon that is analysed by the related data},
"method": {"type": "string", "description": "Literal description of the method of analysis used on the related data"},
"how_innovative": {"type": "string", "description": "A short paragraph describing how data are innovative or how they are used innovatively"}}

Output:
{"data_source":"online job advertisements",
"phenomenon":"skill needs in knowledge production and transfer occupations",
"method":"analysis of data from online job advertisements and from the European Skills and Jobs Survey",
"how_innovative":"The analysis of data from online job advertisements and the European Skills and Jobs Survey provides a comprehensive picture of skills needed in occupations related to science, technology, and ICT, as well as teaching positions from higher education in Europe."}

Data source classification

Prompt passed with "system" role:

1. ◦ Code: Citizen Science Data
   ◦ Description: Data collected by the general public through participation in scientific research.
   ◦ Examples: Environmental monitoring, biodiversity tracking, community-led surveys.

2. ◦ Code: Urban Big Data
◦ Description: Large volumes of data generated within urban settings, public transportation systems, urban sensors, etc.
◦ Examples: Traffic flow data, public transportation data, urban infrastructure usage.
3. ◦ Code: Social Media Data
◦ Description: User-generated data from social media platforms which can be analysed for trends, public opinion, and behaviour related to sustainability.
◦ Examples: Twitter feeds on environmental issues, Instagram posts on recycling practices.
4. ◦ Code: Remote Sensing and Satellite Data
◦ Description: Data collected via satellites and aerial imaging, useful for monitoring environmental changes, land use, and urban development.
◦ Examples: Deforestation tracking, urban heat islands, coastline changes.
5. ◦ Code: Mobile Phone and Telecommunication Data
◦ Description: Data generated from mobile phone usage and telecommunication networks, useful for understanding human mobility and communication patterns.
◦ Examples: Movement patterns during natural disasters, communication trends in response to public awareness campaigns.
6. ◦ Code: Internet of Things Data
◦ Description: Data from interconnected devices and sensors, often used for monitoring and optimising resource use.
◦ Examples: Smart energy metres, water quality sensors in rivers and lakes.
7. ◦ Code: Administrative and Governmental Data
◦ Description: Data collected by government bodies and agencies, often as part of their administrative functions.
◦ Examples: Census data, municipal waste management records, public transport usage statistics.
8. ◦ Code: Financial Transaction Data
◦ Description: Data related to economic activities, including transactions, investments, and consumption patterns.
◦ Examples: Spending patterns on sustainable goods, investment trends in green technologies.
9. ◦ Code: Crowdsourced Data
◦ Description: Data collected through voluntary contributions from a large number of people, typically via the internet.
◦ Examples: Crowdsourced environmental monitoring, community reporting platforms.
10. ◦ Code: Web data
◦ Description: Data collected from the world wide web, either available for download or obtained through HTML scraping.
◦ Examples: Sustainable companies websites, open data repositories on climate data, blogs on social justice activism.
11. ◦ Code: Other
◦ Description: Data that do not fit the descriptions of codes from 1 to 10.”

Prompt passed with “user” role:
“Classify the following record according to the code book.
{"data_source":"online job advertisements",
"phenomenon":"skill needs in knowledge production and transfer occupations",
"method":"analysis of data from online job advertisements and from the European Skills and Jobs Survey",
"how_innovative":"The analysis of data from online job advertisements and the European Skills and Jobs Survey provides a comprehensive picture of skills needed in occupations related to science, technology, and ICT, as well as teaching positions from higher education in Europe."}”

Output:
{"data_category":"Web data"}

Example selection

Prompt passed with "system" role:
In this prompt you will receive a list of records. Each record describe a data source that a certain text described as innovative in the study of a phenomena based on certain methods.
Compare all the records, and select the one that, according to all the information provided, is the most innovative and the most informative.
Most innovative here means that it uses a non-conventional data source (such as data that are not generated for research but that are creatively repurposed for research); or that it uses it in a original way (such as combining different type of data).
Most informative here means that the record contains enough information to assess such innovativeness.
As output, give me the selected record maintaining its exact structure.
Prompt passed with "user" role:
[list of records associated to a combination of data category / dimension]

Example refinement

Human agent

[Attaching PDF of article by Ghitza & Gelman (2020)]
You are a good research assistant who responds only in a single paragraph.
How is this publication's use of Voter Registration Databases to study Improving inferences from political surveys considered innovative?
You may ONLY respond with a single paragraph.

ChatGPT4

The publication's innovative use of Voter Registration Databases (VRDs) for improving inferences from political surveys lies in its ability to leverage large-scale political variables such as party registration and past voting behavior, which are free of overreporting bias or endogeneity between survey responses. The developed process utilizes multilevel regression and poststratification to produce vote choice estimates for the presidential election, projecting these estimates to millions of registered voters in a post-election context. This approach results in stable and reasonable inferences across demographic subgroups and small geographic areas, down to the county or congressional district level. This methodology addresses the limitations of traditional exit polls and enhances the precision and reliability of political survey inferences.