SCIENTIFIC Publishing Limited

Research on European Social Issues

http://ojs.ukscip.com/index.php/resi

Article

Integrating Microdata, Paradata, Metadata and Administrative Data in Quantitative Social Science Research

Apostolos Linardis [1,*] **and Anna Moscha**[2]

[1] *Institute of Social Research, National Centre for Social Research, Athens 10552, Greece*

[2] *Computer Center, National Technical University of Athens, Athens 17773, Greece*

[*] Correspondence: Correspondence: alinardis@ekke.gr

Received: 17 March 2025; **Revised:** 5 April 2025; **Accepted:** 21 April 2025; **Published:** 1 May 2025

Abstract: Integrating survey microdata with auxiliary sources—such as administrative records, metadata, and paradata—significantly enhances the analytical potential of quantitative social science research. This article examines the methodological frameworks of three major international surveys: the EU Gender-Based Violence (EU-GBV) survey, the European Social Survey (ESS), and the Programme for the International Assessment of Adult Competencies (PIAAC). By analyzing their approaches to data collection, management, dissemination and integration, we explore how microdata, paradata, metadata, and administrative records are structured, utilized, and aligned with the principles of FAIR (Findable, Accessible, Interoperable, and Reusable) and Open Data. Our comparative analysis highlights key differences in data accessibility, interoperability, and technological infrastructure, reflecting variations in methodological design and institutional priorities. Metadata emerges as a crucial element for ensuring data transparency, documentation, and reusability, while paradata plays a vital role in monitoring microdata quality and optimizing fieldwork procedures. Administrative data, in turn, provide valuable macro-level insights that support multi-level analyses and a more comprehensive understanding of complex social phenomena. Furthermore, we discuss the role of research infrastructures and international organizations in fostering standardized frameworks for data integration. By synthesizing these insights, this study contributes to ongoing discussions on best practices for managing and integrating complex datasets in social science research. Ultimately, we argue that strengthening these integration efforts can enhance the comparability, transparency, and efficiency of international survey methodologies.

Keywords: Gender Based Violence (GBV); European Social Survey (ESS); Programme for the International Assessment of Adult Competencies (PIAAC); Data Management; FAIR Data; Open Data.

1. Introduction

Quantitative research is a systematic method focused on the collection and analysis of numerical data, aiming to quantify phenomena, behaviors, attitudes, and trends, as well as to identify correlations, differences, or -ideally- causal relationships between variables. The collection of primary data is carried out in various ways, such as through structured questionnaires and recording forms, experiments, or structured observation [1, 2], allowing for the investigation of a wide range of information, including demographic characteristics, behaviors, opinions, emotions, knowledge, and intentions. Social research is the systematic study of human behavior, social structures, and cultural patterns. It seeks to understand how individuals, groups, and societies function, interact, and change over time.

In the context of quantitative social research, data is categorized into several types. Microdata includes information derived from participants' responses or recordings and pertains to the main content of the study. Paradata is auxiliary information that is collected during the microdata collection process but is not part of the survey data itself. Metadata describes the structure, content and context of research data, aiding in its interpretation, while administrative data refers to information produced and managed by public or private organizations during their routine operations. Administrative data are mainly produced by National Statistical Institutes (NSIs) using data originally collected and maintained by other governmental bodies, such as ministries, public agencies, tax authorities, social security institutions, educational bodies, and healthcare systems. While NSIs are often responsible for compiling, processing, and disseminating the statistical outputs, the primary data sources remain within the administrative systems of these public institutions. For example, ministries of health provide hospital records, tax authorities supply income and employment data, and civil registries provide vital statistics such as births and deaths. The collaboration between NSIs and data-owning institutions is essential to transform raw administrative data into standardized, high-quality statistics that support evidence-based policymaking and research. Although administrative data originally collected for operational – non-research - purposes, such data is frequently repurposed for secondary analysis, providing valuable insights to inform policy-making and social research.

Simultaneously, strategic guidelines at both the European and national levels, as well as international best practices for data, demand compliance with the principles of FAIR (Findable, Accessible, Interoperable, and Reusable) and Open Data. Open and FAIR data are not distinct categories of data but rather approaches to managing and sharing the data themselves. Open data refers to data that is freely available for anyone to access, use, modify, and share, subject only to, at most, requirements that preserve provenance and openness. This means the data should be accessible in its entirety at no more than a reasonable reproduction cost, preferably downloadable via the internet, and provided in a convenient and modifiable form. Additionally, the terms of use should permit re-use and redistribution, including the intermixing with other datasets, without discrimination against any person or group [3]. FAIR data refers to data that is Findable, Accessible, Interoperable, and Reusable, principles introduced in 2016 to enhance scientific data management and sharing. To be Findable, data must have a globally unique and persistent identifier, core metadata elements (creator, title, data identifier, publisher, publication date, summary and keywords) and metadata offered in such a way that it can be retrieved programmatically and be registered or indexed in a searchable resource. Accessible data should be retrievable through standardized protocols that support authentication and authorization when necessary, the metadata should contain access level and conditions of the data with metadata remaining available even if the data itself is not. Interoperable data uses standardized formats of metadata by using formal knowledge representation language and controlled vocabularies for specific metadata fields, enabling integration and analysis across different platforms and contexts. Finally, Reusable data must be well-documented, include clear usage licenses under which data can be reused, provenance information about data creation, and adhere to community standards for data and metadata, ensuring others can replicate, validate, and build upon it. These principles promote transparency, collaboration, and efficiency in research, as outlined by the GO FAIR initiative [4, 5].

In modern quantitative social research, all the aforementioned data categories and management processes are seamlessly incorporated into the research workflow.

2. Materials and Methods

Quantitative research is a fundamental tool for major international organizations, which rely on systematic and numerical data to monitor, analyze, and support decision-making on a global scale. Particularly in the social domain, quantitative research is widely used to understand social issues, evaluate policies, and develop programs that promote well-being and social justice. The findings of these international studies significantly influence national and international policies, as the data generated guide the formulation of strategies and the adoption of policies.

International organizations such as the United Nations (UN), the World Bank, the Organisation for Economic Co-operation and Development (OECD), the European Union (through Eurostat), the World Health Organization (WHO), etc. conduct large-scale transnational quantitative social research. These surveys are often comparative across countries, as the comparative approach is vital for understanding differences, similarities, and patterns that impact social, economic, and environmental development on a global scale. However, these organizations usually do not carry out the research-field themselves in every country. Instead, they delegate its implementation to reli-

able local entities, such as national statistical agencies, research centers, private agencies, university departments or other statistical organizations (e.g. ministries). To achieve comparability across countries, harmonization is a key aspect of this process, with three main types being identified: ex-ante input harmonization, ex-ante output harmonization, and mixed harmonization. Ex ante input harmonization means that the institutions that participate in the study have agreed on common concepts, common measurement patterns of the concepts and also on common questions based on a common source questionnaire – usually written in English and then translated by each country- while in ex ante output harmonization the institutions that participate in the study have agreed on common concepts and common measurement patterns but the choice of suitable questions is left to participating research groups who adapt the questions to the cultural particularities of the universe that they study [6]. Some studies, while they mainly follow the strategy of ex ante input harmonization, in certain selected data elements they apply the ex ante output harmonization strategy (mixed harmonization). Most of the surveys conducted by these international organizations follow ex-ante input harmonization strategy, while ex-ante output harmonization is applied selectively for certain variables.

In addition to studies conducted by major international organizations, methodologically rigorous research is carried out by established research infrastructures of the social sciences, such as the ESS (European Social Survey)-ERIC [7] and the SHARE (Survey of Health, Ageing and Retirement in Europe) -ERIC [8]. Both infrastructures are included in the Roadmap of the ESFRI (European Strategy Forum on Research Infrastructures) [9] and adhere to strict methodological approaches for the collection and management of their data. The research infrastructures are also focused on evaluating the comparability of results across countries. To ensure consistency, they adopt standardized practices for data collection internationally, delegating the implementation of these processes to local organizations.

To ensure effective coordination between countries and, in particular, between organizations collecting data at national level, and to ensure data comparability, large multinational organisations and research infrastructures are implementing a set of measures. In addition to training seminars and numerous meetings preceding data collection, they develop methodological manuals and comprehensive websites detailing the rigorous methodology to be followed by the implementing organizations. These manuals include specific guidelines and are often based on general principles and regulations defined by the organizations themselves. At the same time, the use of common software and tools for data collection and quality control is recommended.

In most cases, the organizations responsible for data collection submit a research proposal to the commissioning entities. Data collection cannot commence unless the proposal is approved by the commissioning entities. Additionally, international organisations and research infrastructures provide the necessary expertise to monitor procedures and ensure data quality, intervening where deviations from agreed methodological procedures or contract terms are detected.

International social surveys are not legally mandatory for participating countries, as involvement depends on political will and the availability of national funding. Eurostat follows a mixed approach: certain surveys like EU-SILC (European Union Statistics on Income and Living Conditions) and the Labour Force Survey (LFS) are mandatory under EU regulations, as they are essential for producing harmonized European statistics that support EU-wide social monitoring and policymaking. Their mandatory nature ensures cross-country comparability, continuity, and data availability. In contrast, other Eurostat surveys such as EU - Gender-Based Violence (EU- GBV) are optional and rely on political commitment and co-funding arrangements. Similarly, ESS and SHARE operate on a voluntary basis within their ERIC framework, while other international organization surveys are entirely optional and require full national funding. In all cases, lack of financial support at the national level can lead to withdrawal or non-participation, undermining data continuity and comparability across time and countries.

To this end, a literature review will be conducted on the methodological guidelines, regulations and websites of three major transnational comparative surveys: the EU- GBV survey organized by Eurostat, the European Social Survey (ESS) implemented by ESS-ERIC ESFRI infrastructure, and the Programme for the International Assessment of Adult Competencies (PIAAC) conducted by the OECD.

The selection of the three surveys —EU-GBV, ESS, and PIAAC— was based on a combination of strategic and scientific criteria. These surveys are internationally recognized for their methodological rigor and are conducted by leading institutions in the field of social statistics and research: the EU-GBV is implemented by Eurostat, the official statistical office of the European Union; the ESS is coordinated under the institutional framework of ESS-ERIC; and

the PIAAC is a flagship initiative of the OECD. Notably, the ESS is a European Research Infrastructure Consortium (ERIC) and is also listed on the ESFRI Roadmap, which affirms its strategic importance for European science policy and ensures its long-term institutional sustainability and transnational recognition.

Beyond the credibility of the supporting institutions, each of these surveys addresses distinct yet critical dimensions of contemporary social life: EU-GBV explores gender-based and interpersonal violence, ESS systematically measures social attitudes and values, while PIAAC focuses on adult skills and their connection to employment and education. This thematic diversity enables a multidimensional approach to the empirical investigation of complex social phenomena.

Equally important is the methodological complementarity of the three surveys, which provides a valuable opportunity for comparing different strategies of data collection and management. The EU-GBV, while centrally coordinated, is a particularly sensitive survey both in political and methodological terms, as it captures personal experiences of violence and involves vulnerable populations; it thus requires flexible and context-sensitive implementation with a strong emphasis on confidentiality and ethical safeguards. The ESS, in contrast, represents a model of standardized academic research, prioritizing international comparability through harmonized instruments and protocols. PIAAC adopts a more technocratic orientation, employing psychometric tools to assess core adult competencies through scientifically grounded measurement approaches. Taken together, these diverse methodological perspectives offer a fertile ground for the comparative study of international and European survey practices.

This article focuses on these three surveys to: a) compare similarities and differences in the production and management of microdata, paradata, metadata, and administrative data, b) evaluate compliance with OPEN and FAIR principles and c) examine how the aforementioned data can be integrated adopting the best practices of all three surveys.

3. Results

The upcoming literature review will focus on addressing several critical questions concerning microdata, paradata, metadata and administrative data. Specifically, it will examine which data collection methods are implemented in research to prevent errors and incosistencies, the techniques developed for detecting and correcting structural and logical errors prior or during the data submission as well as the data dissemination procedures. Furthermore, it will examine how interviewers' performance is detected and controlled during fieldwork and examine the practices that organizations use to ensure the quality and transparency of the research process. Finally, it will analyze the use of administrative data in research and the ways it is utilized.

Initially, the findings will be presented in juxtaposition by survey and data category rather than synthesis. Next, the findings will be brought together to highlight key similarities and differences across the surveys. This analysis will also assess how well they follow OPEN and FAIR data principles and explore how their best practices can be combined to create a more effective and streamlined research system. When the methodological framework is more flexible and allows countries to take the initiative, we will give the case of Greece as an example. In Greece, data collection for all waves of the aforementioned surveys has been carried out by the National Centre for Social Research.

3.1. EU-Gender Based Violence (EU-GBV)

The EU-GBV survey on gender-based violence against women and other forms of inter-personal violence coordinated by Eurostat. The purpose of the EU-GBV survey was to gather comprehensive, comparable, and reliable data to measure the prevalence of different forms of violence against women and its impact on their lives. The population of the survey includes women aged 18–74. The first and only wave so far involved 18 EU countries and took place in 2020–2023. The survey was implemented by statistical agencies and authorities in Europe, following the methodological framework and guidelines provided by Eurostat [10]. Statistical bodies are also required to comply with the 16 principles of Eurostat's European Statistics Code of Practice [11] when conducting their surveys. These 16 principles relate to the institutional environment, statistical procedures and statistical outputs. National statistical authorities usually specify these principles through specific guidelines, which are addressed to other national statistical bodies. For example, in Greece, the Hellenic Statistical Authority has developed a relevant analytical guide [12]. Consequently, the GBV survey is built upon a structured and comprehensive methodological

framework, aligned with the principles of the European Statistics Code of Practice. It's important to note that the EU-GBV Eurostat consortium takes a more flexible approach to how countries submit their deliverables, avoiding strict deadlines and frequent reporting requirements to ease the process. This is because the survey is conducted by national statistical authorities or statistical agencies, which already operate within established data collection frameworks.

3.1.1. Microdata

In the methodological framework provided by Eurostat, it was allowed for data collectors to select the data collection methods that best suited their needs, resulting in variation in methods across countries. Eurostat recommended prioritizing face-to-face and computer-assisted methods, though self-completion methods were also endorsed. Some countries, including Greece, opted for a combination of methods to minimize non-response and dropouts, while face to face recruitment was used to minimize the non-coverage error. In mixed methods approaches, the sequence of methods varied by country; for instance, Greece used CAPI (Computer-Assisted Personal Interviewing), CATI (Computer-Assisted Telephone Interviewing), and CAWI (Computer-Assisted Web Interviewing) in that order, while Slovenia used CAWI followed by CAPI. Eurostat provided detailed methodological questionnaire guidelines, including routing (which population is suitable to answer a question), tailoring/piping (how a question is verbally transformed according to the population that answers), and validation rules (range, format, consistency and logical checks) [13, 14] in the methodological manual and related files distributed to data collectors. These rules were to be implemented electronically in the questionnaires during data collection preventing from errors and discrepancies. Eurostat did not recommend the use of common software for the survey but provided the necessary rules that needed to be implemented.

In addition, Eurostat provided countries with SAS/SPSS code to identify any logical and structural errors. Errors had to be corrected before submitting microdata to the EDAMIS (Electronic Data files Administration and Management Information System) platform. The Eurostat EDAMIS is the information system developed by Eurostat to manage the flow of data between national statistical authorities and Eurostat. It is a digital platform that facilitates the collection, submission, and monitoring of microdata related to the official statistics of the European Union. Any microdata with logical or structural errors would fail to upload to EDAMIS.

Indicator 15.2 of the European Statistics Code of Practice refers to access, openness and dissemination services as follows:

> *15.2. Dissemination services use modern information and communication technology, methods, platforms and open data standards (Standards that enable the data to be freely accessed, used, modified, and shared for any purpose (subject, at most, to requirements that preserve provenance and openness).*

In alignment with Indicator 15.2, Eurostat has made the microdata from the EU-GBV survey accessible for scientific research purposes. Researchers affiliated with recognized research entities can request access to the anonymized datasets, by submitting a research proposal to Eurostat [15]. In typical dissemination processes, users are generally able to search the survey metadata before submitting a data access proposal. However, this is not the case for EU-GBV.

3.1.2. Paradata

The paradata help in monitoring the quality of data collection, understanding the conditions under which the microdata were collected, and identifying any issues that may affect the consistency and accuracy of the microdata. Eurostat required some paradata to be submitted with the main data set. These were the variables:

- Mode of Data Collection: information was collected on whether the interview was conducted via CAPI (Computer-Assisted Personal Interviewing), CATI (Computer-Assisted Telephone Interviewing), CAWI (Computer-Assisted Web Interviewing), or other methods.
- Timing of Data Collection: this included details such as the month and year of the interview, which helps in understanding the temporal distribution of data collection activities.
- Interview Duration: the time taken to complete the interview was also recorded, providing insights into the

length and potential respondent burden of the survey.

In addition to the core paradata variables mandated by Eurostat, countries may utilize Case Management Systems (CMS) or Contact Forms to facilitate more specific management and monitoring of interviewers, supervisors, and sample participants. Thus, Greece additionally used the following info/variables collected through CMS:

- Sampling information: number of eligible persons in each household.
- Disposition codes/Response behavior: precoded response behavior such as item non-response, break-offs, and other indicators of respondent engagement are captured, which are critical for evaluating mcrodata quality and response rates.
- Sample dwellings assigned to supervisors/interviewers.
- Contact attempts: the survey records the number of contact attempts made with respondents, which is crucial for assessing the effort required to achieve the final response rate.
- GPS dwelling coordinates.

In conclusion, the variables required by Eurostat, such as mode of data collection, timing, and interview duration, offer valuable insights into the procedural aspects of the survey. Additionally, the use of CMS software in countries like Greece enriches the dataset further by incorporating critical indicators like sampling details and interviewer behavior. This allows for the assessment of microdata quality and the early identification of any issues during the fieldwork. Paradata, such as disposition codes are crucial for assessing quality criteria such as the response rates and their recording and monitoring are common practices in small and large-scale surveys [13].

3.1.3. Metadata

The European Statistics Code of Practice highlights the critical role of metadata in ensuring the quality, transparency, and usability of statistical outputs. Principles 8 & 15 of the European Statistics Code of Practice refer to metadata and are specialized to the following indicators:

> *8.4 Metadata related to statistical processes are managed throughout the statistical processes and disseminated, as appropriate.*
>
> *15.1. Statistics and the corresponding metadata are presented, and archived, in a form that facilitates proper interpretation and meaningful comparisons.*
>
> *15.5. Metadata related to outputs are managed and disseminated by the statistical authority according to the European standards.*

According to the aforementioned indicators, Eurostat provides to statistical agencies the ESS-MH (Metadata Handler) [16] web application that is a platform designed to manage and disseminate metadata associated with the European Statistical System (ESS). Documentation based on metadata is referred to by Eurostat as Quality Reporting and is mandatory alongside the submission of microdata to the EDAMIS system. The ESS-MH is structured according to a standard called SIMS (Single Integrated Metadata Structure) that contains the following information: contact, metadata update, statistical presentation, unit of measure, reference period, institutional mandate, confidentiality, release policy, frequency of dissemination, accessibility and clarity, quality management, relevance, accuracy, timeliness and punctuality, coherence and comparability, cost and burden, data revision and statistical processing. Subsequently, the metadata based on SIMS is made available from Eurostat to the public [17]. It should be clarified that each country creates its own SIMS file but the one available from Eurostat is an integrated and comparative file for all countries.

3.1.4. Administrative Data

In the context of the EU-GBV framework, member states are encouraged to integrate administrative data (e.g., police and court records, helpline calls, and shelter accommodations for women) to provide comprehensive insights into the prevalence and handling of gender-based violence. These data are transmitted to Eurostat to develop a limited set of standardized indicators [10] (p. 580).

As depicted in Figure 1, the layered model of violence data highlights the progression from reported violence - cases formally documented in administrative records - to disclosed violence, captured through self-reported survey

data, and ultimately to actual prevalence, encompassing the true scale of violence, including unreported and hidden cases. This hierarchy underscores the limitations of relying solely on administrative data, as it represents only a fraction of the broader phenomenon. By integrating multiple data sources, such as surveys, researchers can bridge the gap between recorded and real-world experiences, offering a more comprehensive understanding of violence and informing more effective interventions.

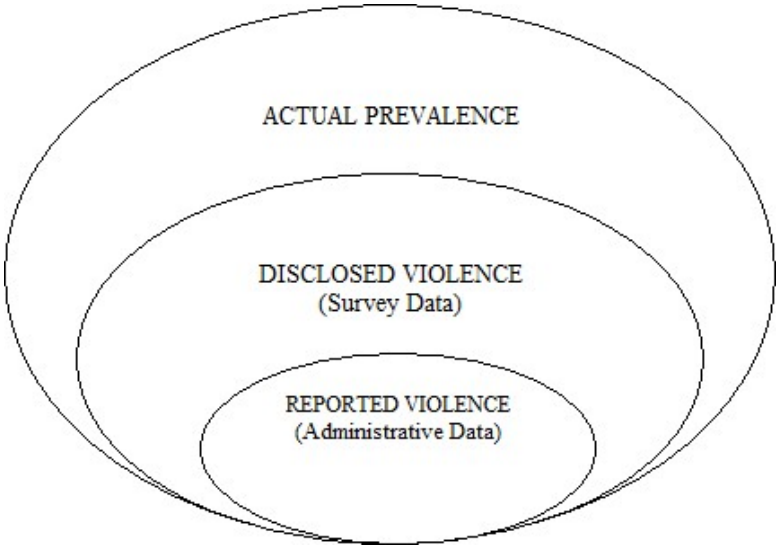

Figure 1. Differences in violence due to data sources. Source EIGE [18] (p. 121).

Survey microdata and administrative data serve distinct yet complementary purposes, each addressing different aspects of gender-based violence. Survey microdata provide valuable insights into the severity and frequency of violence, while also highlighting the socio-economic and cultural factors that influence its occurrence. In contrast, administrative data offer practical information on how cases are managed within the system, focusing on reporting, registration, and processing by police and judicial institutions. Moreover, administrative data help assess the capacity of government agencies and evaluate the effectiveness of victim support services. By integrating these two types of data, a more comprehensive and nuanced understanding of gender-based violence can be achieved, enabling policymakers to address both its root causes and systemic responses effectively.

Other administrative data—such as census data or even more up-to-date figures from the census, including births, deaths, and migration flows—are utilized for the creation of weighting variables and for ensuring the representativeness of the sample.

3.2. ESS

The ESS is a scholarly, cross-national survey that has been conducted across Europe since its inception in 2001. The survey population of the ESS consists of individuals aged 15 years and older, who are residents of private households within the participating countries, regardless of nationality, citizenship, or legal status. Carried out biennially, the survey involves face-to-face interviews with newly selected, cross-sectional samples. It provides valuable insights into the attitudes, beliefs, and behavioral patterns of diverse populations across more than 30 countries. The methodology followed is set out in detail via the infrastructure's website [19]. Participants are selected through rigorous random probability sampling methods at every stage, utilizing sampling frames based on individuals, households, or addresses. At the time this article was written, data from the 11th round of the ESS had been released, and preparations were underway for the 12th round. The ESS Consortium undertakes comprehensive quality assessment activities [19] to ensure reliable and comparable data across countries. These include evaluating question measurement quality using MTMM (Multitrait-Multimethod) experiments and the SQP (Survey Quality Predictor) tool [20], ensuring measurement equivalence for valid cross-national comparisons, and assessing sample representativeness through comparisons with benchmarks like the EU Labour Force Survey. Countries participating in the ESS are continuously monitored by the ESS Consortium throughout the entire process. From de-

sign to final data submission, they are required to deliver multiple reports, which the Consortium centrally oversees to ensure compliance and quality.

3.2.1. Microdata

The ESS traditionally relied on face-to-face interviews for data collection. However, the COVID-19 pandemic highlighted challenges with this approach, leading to the temporary adoption of self-completion methods, such as online and paper questionnaires, in some countries during Round 10 [21]. Following a strategic review, the ESS decided to gradually transition from face-to-face interviews to self-completion methods. According to the plan, in round 12 (2025/26), participating countries will implement a mixed data collection mode: half of the sample will participate through face-to-face interviews, while the other half will complete online or paper questionnaires. From round 13 (2027/28) onwards, data collection will be conducted exclusively through self-completion methods [21]. This transition aims to adapt to changing circumstances while recognizing the challenge of maintaining high response rates.

The source questionnaire (see the round 11 source questionnaire [22]) contains the common questions in English as well as rules for tailoring/piping, routing, and validation that must be incorporated into the CAPI/CAWI questionnaires. Countries are responsible for programming the CAPI questionnaire and digital contact forms, as well as preparing all other fieldwork documents (e.g., advance letters, showcards). However, many countries utilize Centerdata's DataCTRL Survey Tool Suite [23], so the CAPI and CAWI questionnaires, as well as the contact forms and the monitoring mechanisms can be centrally programmed by Centerdata.

Upon completing data collection, participating countries in the ESS are required to deliver a comprehensive set of data and documentation to the ESS Archive (Sikt). This includes datasets from the main questionnaire, the interviewer questionnaire, the contact form, Sample Design Data File (SDDF), raw data and verbatim recorded answers and documents such as the National Technical Summary (NTS), including appendices (education, income, political parties, marital and relationship status), population statistics, main and interviewer questionnaires, contact form, showcards, interviewer and fieldwork instructions, interviewer training material, advance letters and brochures, CAPI programs and interim dataset analysis report [24]. All data files must adhere to ESS-defined variable names, labels, and formats, include consistent respondent identifiers and country codes, and undergo disclosure risk assessments to ensure no identifiable information is present in publicly distributed data.

After the submission, a total of 16 data programs are applied during the data processing stages, starting from the moment the National Coordinators deposit the files into the ESS Archive Intranet until the final draft files are prepared for control and approval. Some of these programs perform automatic checks on the data files, while others generate outputs that require manual verification. All programs, files, and outputs are accessible to the National Coordinators, ensuring complete transparency throughout the processing workflow [19].

Finally, the consortium releases data and documents on a predefined date for each round, rather than by country. These are made available through the ESS data portal [25]. Through the Data Portal, researchers have access to all data categories: research data, paradata, administrative data and documents (e.g. questionnaires, contact forms etc.) via a structured and user-friendly way. Access to the data is available through a simple registration process, complemented by user-friendly tools like the "Data Wizard," which streamline data selection and download.

3.2.2. Paradata

The ESS collects various types of paradata to monitor and enhance data quality. These include:

- Contact form data [26]: detailed records of each contact attempt with potential respondents, documenting interview outcomes and interactions with the interviewers. In the ESS, great importance is placed on this data, as it is a distinct file submitted to ESS Archive and accessible to third parties through the Data Portal [27].
- Time stamps & data collection mode: Accurate records of the start and end times for each interview and questionnaire module are essential for identifying issues such as rushing through questions. They serve as a critical tool for monitoring interview quality and detecting undesirable behaviors like speeding. Time stamps at the seconds level must be included at the beginning and end of each module. The mode of data collection and time stamps can be accessed through the ESS Data Portal under the 'Administrative Variables' group of the integrated data file.

- Interviewer observations [28]: Notes and assessments made by interviewers (included in the module J of the main questionnaire) regarding the interview process and respondent behavior. The ESS places significant importance on this data, making it available as a separate file accessible to third parties through the Data Portal.

These paradata are instrumental in ensuring adherence to fieldwork standards and identifying potential issues during data collection. They are all accessible through the ESS Data Portal.

3.2.3. Metadata

In the ESS, countries do not submit metadata separately using a structured template. Instead, metadata is embedded within the data and documents provided by each country. The consortium centrally manages the metadata documentation and uses a documentation process that is compliant with the Data Documentation Initiative Lifecycle (DDI-L) metadata standard [29]. The ESS Data Portal provides access to rich metadata for each variable, as well as extensive documentation for each round of data collection. The data presentation in portal is centered around variables, offering detailed views for each one. The metadata in integrated datasets (covering multiple waves) make it easy to track how variables evolve over time, whether they stay the same or change. Each collection round is comprehensively documented with links to related materials, covering aspects like universe, time, participating countries and methodology. Country-specific details, such as sampling and collection notes are also documented. For each variable, extensive metadata is provided, including variable groups, response categories, question text, data types, and interviewer prompts.

3.2.4. Administrative Data

The European Social Survey Multilevel Data (ESS MD) [30] is a supplementary dataset designed to complement the individual-level data collected in the ESS. Its primary purpose is to enable researchers to analyze respondents' survey data within the broader contextual environments they live in, such as countries or regions. The ESS MD incorporates contextual and macro-level variables derived from various external data sources, such as national statistics and international databases. These variables typically include socio-economic, political, demographic, and cultural indicators, which provide a richer understanding of the conditions that may influence individual responses. For example, it includes information on GDP, unemployment rates, governance indicators, or cultural diversity indices. The selection of variables in the ESS MD is not exhaustive but is carefully curated based on expert recommendations. This makes the ESS MD a valuable resource for conducting multilevel analyses, where researchers can examine the interplay between individual-level data (e.g., attitudes, behaviors) and their broader societal or contextual frameworks. The ESS MD is also available through the Data Portal.

Additionally, variables like ancestry, citizenship, country of birth, father's country of birth and mother's country of birth, and language most often spoken at home must all be assessed with respect to disclosure risk comparing frequencies and cross tabulations with population statistics at a national and regional level. It is strongly recommended that citizenship, country of birth and country of birth of parents are assessed together since they might represent combination of minority groups [31].

Furthermore, ESS uses post-stratification weights [32] that use auxiliary information to reduce the sampling error and potential non-response bias. They have been constructed using information on age group, gender, education, and region. The population distributions for the adjusting variables were obtained from the European Union Labour Force Survey.

Finally, efforts are being made within the Data Portal to integrate datasets from other sources that can be linked to the social variables of the ESS, such as data from the Climate Neutral and Smart Cities Science Project [33] and cross-national data based on the CRONOS probability-based harmonized web panel.

3.3. PIAAC

The PIAAC is a multi-cycle international computer-based household survey of adults aged 16-65 years sponsored by the OECD. It focuses on assessing adult skills and competencies crucial for participation in 21st-century economies and societies. The basic survey questionnaire collects data on participants' educational backgrounds, professional achievements, and their use of information and communication technologies and is accompanied by a psychometric assessment design that focuses on measuring four key domains: literacy skills, numeracy skills,

reading and numeracy components, and adaptive problem-solving [34, 35]. PIAAC is designed as 10-yearly cycles. The first cycle of the PIAAC survey took place in three rounds, with a total of 39 countries participating. The second cycle began with 31 countries in Round 1, culminating in the release of data on 10 December 2024.

PIAAC data collection is governed by uniform standards to ensure data comparability across countries. This comparability is achieved through the use of not only common data collection instruments but also standardized methods of data administration. Participating countries are committed to adhering to these standards and must continuously demonstrate their compliance throughout the implementation process. Each country is responsible for conducting PIAAC in compliance with the PIAAC Technical Standards and Guidelines [34, 35] provided by the Consortium to ensure that the survey design and implementation yield high-quality and internationally comparable data.

In PIAAC, each participating country plays a crucial role in ensuring the success and accuracy of the survey by providing a series of key deliverables at different stages of the process. During the planning phase, countries develop detailed sampling plans to carefully design their approach to selecting participants, ensuring fair representation and reliable response rates. At the same time, they adapt the survey instruments to fit their national context, making necessary linguistic and cultural adjustments to maintain consistency across countries. Before data collection begins, they conduct thorough interviewer training and submit reports documenting how interviewers are prepared to administer the survey effectively.

Once fieldwork is underway, countries continuously track progress through Sample Monitoring Forms (SMFs), ensuring that data collection remains on track and meets the required quality standards. After the fieldwork is completed, they conduct a Nonresponse Bias Analysis (NRBA) to examine whether certain groups were underrepresented and take steps to correct any imbalances. A weighting plan is also submitted to ensure that survey results accurately reflect the broader population.

Throughout the process, each country compiles National Survey Design and Planning Reports (NSDPR), which document their approach, adjustments made along the way, and lessons learned. Finally, after the end of the collection process, one of the most critical deliverables is the database submission, where countries provide their clean and verified data in a standardized format, allowing for seamless integration into the international dataset.

3.3.1. Microdata

In-person interviews are required to complete the background questionnaire and administer the direct assessment. A computer-assisted data collection method must be utilized at all stages of the data collection process. The direct assessment is delivered in a tablet-based format, using the same tablet system employed for collecting background questionnaire data [35]. The background questionnaire is completed by the interviewer, while the assessment is completed directly by the respondent. If the respondent is unable to use the CAPI system, the assessment is administered in a paper-and-pencil format. The international master questionnaire [36] includes inference rules, consistency checks, and routing codes that must be implemented within the CAPI software. The PIAAC consortium provides specific CAPI software to countries. Besides the CAPI software, the consortium provides Data Integration Software, which integrates data from different sources to one single database. The database integrates diverse datasets - data and paradata - to support both computer-based and paper-pencil assessments. It includes: a)sample design data from countries, processed into the Sampling Data Information File (SDIF) dataset, which is a file that includes all sampled persons, nonrespondent sampling units (e.g., dwelling units) and unsampled dwelling units, that is created through the CMS system and information from the sampling frame and is essential for ensuring consistency and facilitating weighting adjustments, b)background questionnaire coded responses, ex-post coded responses of verbatim responses and background questionnaire log files, c)Computer-Based Assessments (CBA) responses and detailed log files for tracking respondent interactions, d)scoring for the paper and pencil assessments and e) post-interview questions, answered by the interviewers. The Data Integration Software provided by the consortium generates the necessary reports to facilitate within-country verification during the survey's progression and prior to data submission.

Following data submission, the Consortium conducts identical checks to identify any unresolved issues not previously addressed by national organizations. Further verification checks and cleaning logic, including those for multivariate inconsistencies, will be implemented at the international level and reported back to the countries for review, comments, and corrections based on both within-country and cross-country analyses.

Finally, users can download Public Use Files (PUFs), codebooks, background questionnaires, and all necessary materials to conduct their own analyses from the "PIAAC 1st Cycle Database" [37] and "PIAAC 2nd Cycle Database" [38] pages. To access the PUF files, users should answer a questionnaire providing their contact details and optionally research project details [39]. The PIAAC survey utilizes complex sampling and psychometric designs, requiring advanced statistical methods to estimate sampling and measurement variances and run the appropriate statistical analysis. To assist users of SAS, SPSS, Stata and R, specialized tools have been developed to facilitate data exploration and statistical computations [40]. In addition to the above tools, there is a specialized web application, the PIAAC Data Explorer [41], designed to enable users to explore and analyze the collected data online as well as the PIAAC LogDataAnalyzer [42] which is a tool that facilitates analysis of the log data. Due to the complexity of the PIAAC design and data and the complex statistical analysis required for proper statistical inference, it appears that PIAAC places a strong emphasis on providing statistical tools rather than structured metadata and documentation.

3.3.2. Paradata

As already mentioned, at the end of data collection, each participating country must submit a complete database that is exported through the Data Integration Software. This database includes data and paradata and sampling information. Besides, the data - (a) background questionnaire responses from the PIAAC CAPI, (b) the coding of education, occupation, industry, language, country and region and (c) the cognitive assessment responses and scores for automatically scored items - it also includes many paradata such as [35]:

- The CMS paradata, such as disposition codes describing the contact results and contact details, incorporated in the SDIF file,
- Background questionnaire log file, which is a log/audit dataset that holds information about the interviewer's actions during the CAPI (e.g. time stamps for interview start, interview paused, interview end, open help and other actions).
- Computer based assessment log files, that provide a detailed understanding of respondents' interactions during computer-based assessments, allowing researchers to derive process indicators that complement traditional assessment scores. The PIAAC log file data is of potential relevance to researchers and others interested in better understanding a range of issues relating to test-taking behavior and the strategies and processes followed by respondents in responding to test items.

It should be noted that the consortium provides participating countries with an international CMS, a software platform designed to organize, track, and manage information, processes, and workflows related to the enumeration and selection of cases. Alternatively, countries may choose to use their own CMS, provided it adheres to the standards outlined in the methodological guidelines for software.

The PIAAC 2012 study was the first large-scale educational assessment to be conducted entirely on a computer. As participants completed the assessment, their interactions within the system were carefully tracked and recorded with time stamps in special log files. These files, which provide valuable insights into response patterns and behaviors, are available through the GESIS repository [43].

3.3.3. Metadata

In PIAAC, countries do not submit metadata separately using a structured template. Instead, some unstructured metadata is embedded within the data and documents provided by each country. The consortium centrally manages metadata documentation, but it doesn't seem to place any particular emphasis on it. In PIAAC, there is no specific metadata standard used for data dissemination; however, metadata and documentation are available – through a rather unstructured way - via the corresponding website [44]. The metadata for variables is accessible through the PUF files, while additional metadata for each country is provided in the technical reports [45].

3.3.4. Administrative Data

In PIAAC, administrative data play a dual role, both in ensuring data accuracy and expanding research possibilities. On one hand, they are crucial for Nonresponse Bias Analysis (NRBA) and weighting adjustments, helping to correct biases related to eligibility, nonresponse, and benchmarking, ensuring that survey results truly reflect

the broader population. The NRBA and weighting adjustments rely on auxiliary variables, primarily derived from registries, censuses, or other administrative sources.

At the same time, in Nordic and Baltic countries like Denmark, Estonia, Finland, Norway, and Sweden, decades of individual-level register data allow researchers to link PIAAC data across cycles, offering deep insights into how cognitive skills, education, and employment shape people's lives and influence broader societal trends. This linkage enables longitudinal studies, helping to analyze skill demands in the labor market, the effectiveness of education systems, gaps between education and employment, and young people's transition into the workforce, while also exploring connections between skills, income, health, and lifelong learning [46].

4. Discussion

Building on the literature review and the findings from the three surveys, which were presented in juxtaposition by survey and data category, this section addresses the three previously stated research questions.

4.1. Synthesizing Similarities and Differences

The synthesis below highlights both commonalities and distinctions across the surveys while maintaining a clear comparative structure.

4.1.1. Microdata

All three surveys emphasize methodological rigor, but their approaches to data collection, quality control tools and data dissemination diverge. ESS and PIAAC traditionally reliant on computer-assisted in-person interviews. Nevertheless, ESS is transitioning to self-completion methods to adapt to changing contexts, such as those highlighted during the COVID-19 pandemic. In contrast, the EU-GBV survey allows participating countries more flexibility in choosing data collection methods, based on local needs. This flexibility was allowed because the survey dealt with a highly sensitive topic, was carried out by trusted statistical agencies, and marked Eurostat's first time conducting it.

Quality control and data integration processes also reveal differences. The ESS relies on centralized tools like the DataCTRL Survey Tool Suite to ensure consistent programming and compliance across countries. Similarly, PIAAC employs consortium-provided software for database construction and validation, ensuring that all participating countries adhere to uniform standards. Moreover, for both aforementioned surveys, specific files were submitted to validate the probabilistic sampling methodology. The EU-GBV survey, however, combines country-specific methods with Eurostat's overarching guidelines, enabling tailored solutions while maintaining a baseline of comparability. This flexibility provided by the EU-GBV stems from the fact that Eurostat cooperates with organisations such as national statistical offices and authorities, which guarantee methodologically sound data collection, the application of quality control and ensure appropriate probabilistic sampling.

Comparing research data accessibility, the ESS stands out by offering open access to data following a simple registration process. PIAAC also offers immediate access to PUFs after completing a short questionnaire, requiring mandatory fields such as name, affiliation, and email address. In contrast, the EU-GBV survey imposes stricter data governance policies, granting access only upon request.

Despite these differences, the shared purpose across these surveys is evident: providing high-quality, comparable data to inform policy and academic research. The ESS seeks to illuminate cross-national differences in social attitudes and behaviors, the EU-GBV survey focuses on understanding the prevalence and impact of gender-based violence, and PIAAC evaluates adult competencies critical for participation in modern economies.

4.1.2. ParaData

The three surveys share several similarities in their use of paradata, highlighting their common goal of ensuring high-quality data collection. All three surveys incorporate paradata to monitor the quality and efficiency of the data collection process. This includes tracking variables such as the timing of interviews, contact attempts, and disposition codes. Moreover, the surveys integrate paradata into centralized systems or tools, enabling real-time validation and monitoring. These systems not only ensure data quality but also enhance transparency by providing detailed insights into the conditions under which the data was collected, the performance of interviewers, and the

behavior of respondents.

Despite these shared practices, the surveys differ significantly in the depth, tools and accessibility of their paradata collection and processing. The ESS employs an extensive array of paradata, including detailed contact form data, timestamps for all survey modules, and interviewer observations. The PIAAC incorporates a vast array of advanced paradata, including detailed audit logs from computer-based assessments and the background questionnaire, seamlessly into its workflow. In contrast, the EU-GBV Survey collects a narrower range of paradata, focusing on essential variables such as mode of data collection, interview duration, and timing.

In terms of technological integration, the ESS and PIAAC demonstrate a higher reliance on advanced tools for paradata management. The ESS employs the DataCTRL Survey Tool Suite, while PIAAC takes a more technology-intensive approach. The PIAAC system includes unique "technical" features like audit logs and adaptive workflow data, which are integrated with responses for comprehensive analysis. The EU-GBV survey, while technologically supported, does not have a unified framework for paradata processing, instead relying on a combination of Eurostat guidelines and country-specific systems.

Additionally, all ESS paradata is made accessible through the ESS Data Portal, reflecting its emphasis on open data practices while a part of PIAAC paradata is made accessible through the PUFs containing individual unit record data files and other through the PUFs that contain the log data from the PIAAC cognitive assessments that can be downloaded from the repository of GESIS [43]. EU-GBV paradata variables may be provided by Eurostat in a micro-data request.

Overall, while the three surveys share a common emphasis on the importance of paradata for maintaining quality and transparency, their approaches differ in scope, depth, accessibility, and technological sophistication.

4.1.3. Metadata

All three surveys recognize metadata as essential for ensuring transparency and usability of their datasets. Each survey incorporates metadata into their workflows, documenting survey processes and methodologies to facilitate proper data interpretation, support research reproducibility, and enhance comparability across participating countries.

However, there are clear differences in metadata submission, structure and completeness. Eurostat is the only organization that requires countries to submit their metadata through the ESS-MH. In contrast, for the ESS, the consortium itself handles the metadata centrally. A similar approach is taken for PIAAC, where only limited metadata is provided by the consortium. Concerning the structure, the ESS employs the DDI-L standard, offering a highly structured and user-friendly metadata system accessible through its Data Portal. Researchers can view detailed metadata for each variable, including its definition, data types, and associated collection rounds. The ESS also provides metadata linked to integrated datasets for longitudinal analysis and maintains rich documentation for cross-national comparisons. The EU-GBV survey, by contrast, uses the SIMS standard provided by Eurostat. While this ensures adherence to European standards, the metadata system is less interactive compared to the ESS Data Portal and doesn't integrate variable metadata. PIAAC differs further by lacking a structured metadata standard like those used by the ESS or EU-GBV. Limited and unstructured metadata is available through PIAAC's website, where users can access documents such as codebooks and technical reports, but the information is less organized and requires more effort to navigate.

Besides the fact that the ESS offers a highly structured and user-friendly metadata system, according to Gregory, Wackerow & Orten (2023) [47], the ESS metadata lacks granularity in two key areas. First, formal definitions of concepts are not directly provided; researchers must refer to survey instruments and questionnaires to understand the context of questions and possible responses. These questionnaires, available at the country level, offer valuable insights for cross-national comparisons. Second, data processing details are documented but require researchers to download and review related documents to determine specifics for each variable. The ESS is addressing these limitations by developing a new dissemination system, which will include metadata in XML format following the DDI-L standard, along with improved tools for comparing variables across data collection rounds.

Overall, although all three surveys emphasize metadata as a cornerstone of their research frameworks, their approaches to structuring and detailing this information differ. All survey metadata are freely available and contribute to a deeper understanding of the data themselves. However, they vary in terms of structure and completeness.

4.1.4. Administrative Data

The ESS, EU-GBV, and PIAAC integrate administrative data in two key ways: enhancing data accuracy and broadening research opportunities. Data accuracy concerns survey design and implementation as well as quality checks. Administrative data are used to create accurate and up-to-date sampling frames, define the target population, and allocate the sample based on actual population distributions. They also support the construction of post-stratification weights, which adjust survey data according to real demographic distributions. By comparing key demographic or socio-economic variables—such as age, gender, education, or employment status—between the survey sample and corresponding administrative population data, researchers can identify coverage errors, non-response biases, or deviations from expected distributions. Such comparisons allow for the detection of underrepresented or overrepresented groups in the sample and guide the application of post-stratification or calibration weighting adjustments. Additionally, administrative data constitute a powerful asset for the social sciences, as they greatly expand the range and depth of research possibilities. Their comprehensive coverage, low bias, and longitudinal nature enable the exploration of complex social phenomena with enhanced empirical robustness and broader generalizability—often making feasible research designs that would otherwise be unattainable. Within this broader context, the integration of administrative data into machine learning models can further improve the precision, scalability, and policy relevance of analytical tools used in evidence-based decision-making.

A distinctive feature of administrative data is that, unlike survey data, paradata, or metadata, they are not generated by the data-producing organization itself, but rather originate from NSIs or other public and private bodies, and are collected for administrative and operational purposes, not with research in mind. Access to such data typically requires lengthy and complex procedures, and their content is often not fully aligned with research needs, as it reflects institutional mandates rather than scientific standards. Moreover, their use is subject to strict legal constraints, frequently necessitating special authorizations or formal data access agreements. In many cases, individual-level register data are not available, limiting the scope for linking with other individual-level datasets. Administrative statistics are also commonly disseminated in aggregated or non-standardized formats, which further hampers their integration with survey-based microdata. Consequently, and despite their considerable analytical potential, administrative data remain, in practice, only partially compatible with the FAIR and OPEN data principles, highlighting the need for institutional commitment, legal clarity, and technical interoperability to enable their responsible and effective use in research.

Despite these similarities, administrative data vary significantly in scope, with each survey incorporating different types of data to serve its unique objectives. The ESS primarily uses administrative data as a contextual supplement through its ESS MD, which includes macro-level indicators like GDP and unemployment rates. The EU-GBV survey integrates administrative sources such as police reports, court records, and helpline calls to generate standardized indicators of gender-based violence. In the case of PIAAC, some countries link survey responses with register data, allowing for a more comprehensive approach by connecting longitudinal administrative datasets. This enables advanced analyses of education, employment, and broader socio-economic trends.

The accessibility of administrative data also varies across these surveys. ESS offers open access to its multilevel data through its Data Portal, making it widely available for research. In contrast, PIAAC provides structured access and allows data linkage in certain countries, but integration depends on national frameworks. These varying approaches highlight the distinct goals and operational frameworks of each survey. The ESS focuses on providing contextual insights for all countries, while PIAAC builds on national initiatives to enable in-depth longitudinal analysis.

4.2. Compliance with FAIR and OPEN principles

A key finding that emerged from the synthesis of the surveys is that the FAIR and OPEN principles do not apply to the surveys as a whole but concern each data category separately - microdata, paradata and administrative data. On the other hand, metadata is a crucial tool for ensuring fairness and openness, as the majority of the criteria for aligning data with the FAIR and OPEN principles rely on metadata [48]. Consequently, metadata itself will not be evaluated below, as it serves as the foundation for assessing other data. Additionally, interoperability goes beyond standardized metadata formats, encompassing seamless integration between microdata, paradata, metadata, and administrative data. Each survey adopts different practices and levels of compliance depending on the data category,

highlighting the need for more targeted approaches to their management and dissemination.

Concerning research microdata, the ESS excels in openness and FAIR compliance, providing freely accessible data with comprehensive metadata and minimal access barriers, ensuring findability and reusability. Its data is also highly reusable with clear usage licenses under which data can be reused and disseminates data in standardized community data formats such as csv, SPSS, and STATA and metadata formats such as DDI-L. Administrative data and paradata are disseminated in separate files while maintaining the same level of openness and fairness as research microdata. All data categories can be seamlessly integrated using key variables that ensure smooth interoperability. Furthermore, demonstrating its interoperability with other surveys, the ESS portal also disseminates data from the Climate Neutral and Smart Cities Science Project. These datasets are integrated with ESS data, combining environmental data with insights on people's attitudes and behaviors to support social, political, and scientific analysis.

The EU-GBV survey includes a small amount of paradata variables in its main data - microdata - and there are no specific files available for administrative data or paradata. Researchers can request access to microdata through Eurostat's microdata access portal by submitting a project proposal. However, there is no detailed metadata available for individual variables, which can make it harder to understand the context and structure of the data before applying for access. The indicators generated from the microdata are searchable, available and accessible through Eurostat's database [49], specifically under the path: "Detailed datasets → Population and social conditions → Living conditions and welfare → Gender-based violence against women". Each indicator is accompanied by an "M" symbol, which links to the EU-GBV SIMS metadata. The same metadata can also be accessed when viewing each EU-GBV indicator in table format. So, concerning microdata, EU-GBV demonstrates moderate FAIR compliance, adhering to Eurostat's guidelines to ensure quality and comparability, but restricts access, requiring formal approval for data use, which limits openness. Aggregated EU-GBV data, such as key indicators, are freely accessible.

PIAAC microdata is considered open, as it is accessible through PUFs following the completion of a short questionnaire. Access to country specific PUFs differs by country. PIAAC integrates advanced paradata, including audit logs and timestamps, which provide valuable insights into respondent behavior and interviewer performance, but access is partially restricted. PUFs containing log data from the PIAAC cognitive assessments can be downloaded from the GESIS Data Catalogue; however, they are only available for the first cycle and first round of data collection. PIAAC exhibits moderate compliance with FAIR principles, primarily due to its relatively limited and unstructured metadata. Regarding administrative data, they are generally unavailable. However, in countries like Denmark, Estonia, Finland, Norway, and Sweden, statistical offices provide access to annual individual-level register data spanning several decades that enable researchers to link PIAAC data from both Cycle One and Cycle Two with register data for research purposes.

4.3. Integrating the Best Data Practices

This section describes how microdata, paradata, metadata, and administrative data can be integrated adopting the best practices of all three surveys. The main approaches for integrating these elements are summarized below and illustrated in Figure 2.

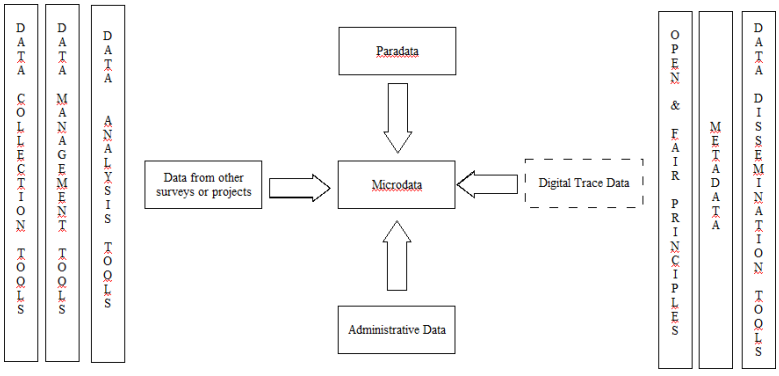

Figure 2. The integration of survey microdata (dashed lines indicate potential integrations not analytically examined in this article).

4.3.1. Connecting microdata with paradata, metadata and administrative data: Business and Physical Linkage

Surveys emphasize the need for a seamless and well-integrated workflow, where different data elements come together to improve both the quality and clarity of microdata. A key part of this process is paradata, which helps ensure methodological rigor by tracking data collection in real time and spotting potential biases or errors. At the same time, metadata plays a crucial role in keeping things transparent and user-friendly. It provides detailed documentation on the structure of the data, key variables, and the overall process—from data collection to final analysis—making secondary research more reliable and insightful. Beyond this, administrative data is incorporated to add depth, allowing researchers to connect individual survey responses to broader societal trends. By combining all these elements—survey microdata, paradata, metadata, and administrative records—we create a rich, multidimensional dataset that supports strong analysis and informed, evidence-based decision-making.

When managing or analyzing data, it is often consolidated into a single dataset. This dataset typically consists of microdata, though in some cases, it also includes paradata or/and administrative data from one or multiple survey waves. From a technical standpoint, paradata is linked to microdata using key variables, while administrative data can be connected either at the individual level through these same key variables or using geographic identifiers. Among the surveys examined, the ESS stands out for its well-structured and transparent data integration process. In the ESS framework, microdata, paradata (such as contact forms and interviewer observations), and administrative data (ESS-MD) are provided as separate files. However, each file contains unique identifiers, allowing researchers to seamlessly merge them while maintaining flexibility and accuracy. Additionally, when merging microdata across different survey rounds, all common variables retain the same names across waves. This consistency simplifies the process of linking data over time, making longitudinal analysis more straightforward.

However, merging diverse databases requires specialized skills. The role of the National Data Manager in the PIAAC survey [34] is a characteristic example, as it demanded the ability to create comprehensive datasets derived from multiple sources, enabling further processing and analysis.

4.3.2. Technological Integration: Data Collection – Management – Analysis & Dissemination Tools.

The tools for data collection, management, analysis, and dissemination go beyond just microdata—they also cover paradata and administrative data. For microdata, surveys utilize CAI systems and employ both simple and mixed data collection methods. Paradata, which captures metadata about the data collection process itself, is managed through CMS platforms and log tracking tools. Administrative data collection, on the other hand, is a more hands-on process. Researchers manually explore databases such as Eurostat's indicators database, the Eurostat Census Hub, and various registers to extract relevant information. They then construct key variables to integrate this data with microdata, ensuring a more comprehensive analytical framework.

As datasets grow more complex - such as PIAAC microdata - the need for advanced analytical tools becomes even more critical. To meet this challenge, PIAAC offers specialized platforms like the PIAAC Data Explorer and LogDataAnalyzer, while also ensuring compatibility with statistical software such as SAS, SPSS, Stata, and R. As data becomes more intricate, so must our analytical approaches, demanding ever more sophisticated tools and methodologies to extract meaningful insights.

Finally, dissemination tools and platforms aren't limited to microdata - they also support other types of data, including paradata and metadata. The European Social Survey (ESS) stands out as the only survey that uses a single dissemination platform for both paradata and administrative data, ensuring consistency and accessibility across different data types.

4.3.3. The essential role of structured metadata

Metadata plays a crucial role in making sense of any type of data—whether it's microdata, paradata, or administrative records. Without it, understanding and interpreting datasets would be much more difficult. It acts as a detailed guide, explaining variable definitions, methodology, and sampling processes, ensuring that data is both transparent and usable. Keeping metadata well-documented becomes even more challenging for repeated – longitudinal surveys, especially when working with integrated datasets at the country or round level. In such cases, certain variables remain unchanged and should be preserved as they are in relation to their metadata for easy searchability. Others evolve over time, requiring careful monitoring and documentation of changes. Additionally, new variables

should be introduced within a comparative framework to ensure efficient management in the future.

All types of data - microdata, paradata, and administrative data - rely on metadata for proper documentation. However, the ESS is unique in systematically documenting all three data types in the same structured way, ensuring consistency and ease of access for researchers. Additionally, ESS is compliant with DDI-L metadata standard, while EU-GBV is compliant with SIMS metadata standard.

4.3.4. Linkage with other surveys and sources of data

The ESS portal goes beyond just collecting survey data - it also connects with the Climate Neutral and Smart Cities Science Project, making it a valuable resource for understanding how people's attitudes and behaviors relate to environmental issues. By integrating these datasets, researchers can explore the social and political dimensions of climate action, helping to paint a more complete picture of how individuals and communities respond to environmental challenges. This seamless combination of data allows for richer, more meaningful analysis that can inform both policy and scientific discussions.

In Figure 2, although no such cases emerged in the case studies of the surveys we examined, we list the potential link between surveys and digital trace data that can be roughly defined as "records of activity (trace data) undertaken through an online information system (thus, digital)" [50]. As Stier et. al. point out in their article [51] the main advantages of the linkage between surveys and digital trace data are the cross-validation and improvement of measurements, the explanation of human behavior at a large scale, and novel opportunities to improve causal inference in experimental settings. Additionally, declining response rates in surveys, particularly among those most engaged with digital technologies, further challenge the effectiveness of survey methods [51].

4.3.5. Adherence to Standards

The integration process is guided by international standards and principles like FAIR and OPEN data. This ensures that microdata, paradata, metadata, and administrative data are not only linked but also accessible for secondary use and replication. Metadata, when aligned with established standards (e.g., DDI, SIMS), strengthen the FAIR principles of data by ensuring interoperability across different systems, platforms, and tools. Standardized metadata enable seamless data exchange, interpretation, and utilization in a consistent and structured manner.

5. Conclusions

Figure 2 showcases how survey data fits into a larger research ecosystem, bringing together multiple sources of information. This framework is designed to integrate various datasets— survey microdata, administrative records, and paradata—while also recognizing the potential for incorporating additional sources like digital trace data (e.g. social media) or datasets from other surveys and projects. By expanding the range of available data, researchers can achieve deeper and more comprehensive insights. At the heart of this system is microdata, which acts as a bridge connecting different datasets and reinforcing the foundation for broader research efforts. The framework follows OPEN and FAIR principles, ensuring that data remains accessible, reusable, and transparent, thus promoting collaboration and innovation. Metadata plays a crucial role by providing essential context and improving data discoverability, while paradata—which captures details about the data collection process—enhances quality control and interpretation. Additionally, specialized tools are employed at every stage of the data lifecycle, from collection and management to dissemination and analysis, ensuring seamless integration and maximum utility. By embracing this comprehensive approach, researchers can uncover patterns and relationships that might otherwise remain hidden when working with isolated datasets. This enhances the depth and impact of research, leading to more robust, well-rounded findings.

Figure 2 highlights a key challenge: managing and analyzing complex datasets requires advanced technical skills. To tackle this, initiatives like PIAAC have invested in developing analytical management and analysis tools that help researchers work with large and intricate data structures.

However, this article does not explore critical issues like GDPR compliance or the ethical responsibilities involved in research. Bringing together data from multiple sources can increase the risk of exposing sensitive or personally identifiable information, making careful handling essential. Beyond just technical solutions, addressing these challenges requires clear ethical guidelines, strong legal frameworks, and proper training for researchers.

It's not just about having the right tools—it's about ensuring that researchers have the knowledge and support to integrate data responsibly, protect privacy, and uphold research integrity.

Author Contributions

Conceptualization, Apostolos Linardis and Anna Moscha; methodology, Apostolos Linardis; formal analysis, Apostolos Linardis and Anna Moscha; investigation, Apostolos Linardis and Anna Moscha; writing—original draft preparation, Apostolos Linardis; writing—review and editing, Anna Moscha; supervision, Apostolos Linardis. All authors have read and agreed to the published version of the manuscript.

Funding

This work received no external funding.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

No new data were created or analyzed in this study. Therefore, data sharing is not applicable to this article.

Acknowledgments

The authors gratefully acknowledge the researchers of the Institute of Social Research at the National Centre for Social Research (EKKE), Greece - Dimitra Kondyli, George Papadoudis, Antoinetta Capella, Andromachi Hadjiyanni, and Alexandra Theofili - for their invaluable contributions to the implementation of the Greek wave of the EU-GBV survey.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Bryman, A. *Social Research Methods*, 5th ed.; Oxford University Press: Oxford, UK, **2016**; pp 197-282.
2. Babbie, E. R. *The Practice of Social Research*, 15th ed.; Cengage AU: Southbank, VIC, Australia, **2020**; pp 228-288.
3. Open Data Handbook—What is Open Data? Available online: https://opendatahandbook.org/guide/en/what-is-open-data/ (accessed on 31 January 2025).
4. GO FAIR—FAIR Principles. Available online: https://www.go-fair.org/fair-principles/ (accessed on 31 January 2025).
5. Wilkinson, M. D.; Dumontier, M.; Aalbersberg, I. J.; et al. The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Sci. Data* **2016**, *3*, 1-9. [CrossRef]
6. Kallas, J.; Linardis, A. A. Documentation Model for Comparative Research Based on Harmonization Strategies. *IASSIST Q.* **2008**, *32*, 12. [CrossRef]
7. European Social Survey. Available online: https://www.europeansocialsurvey.org/ (accessed on 31 January 2025).
8. SHARE. Available online: https://share-eric.eu/ (accessed on 31 January 2025).
9. ESFRI Roadmap 2021. Strategy Report on Research Infrastructures. Available online: https://roadmap2021.esfri.eu/ (accessed on 31 January 2025).
10. Eurostat. Methodological Manual for the EU Survey on Gender-Based Violence Against Women and Other Forms of Inter-Personal Violence (EU-GBV), 2021. Available online: https://ec.europa.eu/eurostat/web/products-manuals-and-guidelines/-/ks-gq-21-009 (accessed on 31 January 2025).

11. Eurostat. Code of Practice–Revised Edition 2017, 2017. Available online: https://ec.europa.eu/eurostat/web/products-catalogues/-/ks-02-18-142 (accessed on 31 January 2025).
12. Hellenic Statistical Authority. Guide for the Entities of the Hellenic Statistical System on the Implementation of the Code of Practice for European Statistics, 2018. Available online: https://www.statistics.gr/documents/20181/1196143/Odigos_ELSS_1_0.pdf/ (accessed on 31 January 2025).
13. Linardis, A.; Maravelakis, P.; Fragoulis, G. Data Collection Methods Using Digital Questionnaires and Survey Methodology. *Management of Online, In-Person and Telephone Data and Surveys with Limesurvey and SPSS*; Kallipos, Open Academic Publications: Athens, Greece, **2023**; pp 83-88. [CrossRef]
14. Biffignandi, S.; Bethlehem, J. *Handbook of Web Surveys*, 2nd ed.; John Wiley & Sons: Hoboken, NJ, USA, **2021**; pp 237-320.
15. Microdata-EU Survey on Gender-Based Violence. Available online: https://ec.europa.eu/eurostat/web/microdata/gender-based-violence (accessed on 31/01/2025).
16. European Statistical System Metadata Handler (ESS-MH) User Guide. Available online: https://cros.ec.europa.eu/book-page/ess-metadata-handler-ess-mh-user-guide (accessed on 31 January 2025).
17. Gender Based Violence Against Women (GBV), Reference Metadata in Single Integrated Metadata Structure (SIMS), Compiling Agency: Eurostat, the Statistical Office of the European Union. Available online: https://ec.europa.eu/eurostat/cache/metadata/en/gbv_sims.htm (accessed on 31 January 2025).
18. EIGE, Gender Equality Index 2015 - Measuring Gender Equality in the European Union 2005-2012. Available online: https://eige.europa.eu/sites/default/files/documents/mh0415169enn.pdf (accessed on 31 January 2025).
19. ESS Methodology. Available online: https://www.europeansocialsurvey.org/methodology/ess-methodology/survey-specification (accessed on 31 January 2025).
20. GESIS–Survey Quality Predictor. Available online: https://sqp.gesis.org/ (accessed on 31 January 2025).
21. ESS - Modes of Data Collection - The ESS Move to Self-Completion Data Collection. Available online: https://europeansocialsurvey.org/methodology/methodological-research/modes-data-collection (accessed on 31 January 2025).
22. ESS Round 11 (2023) Source Questionnaire. Available online: https://stessrelpubprodwe.blob.core.windows.net/data/round11/fieldwork/source/ESS11%20Source%20Questionnaires.pdf (accessed on 31 January 2025).
23. CENTERDATA - DataCTRL Survey Tool Suite. Available online: https://en.centerdata.nl/werkvelden-2/datactrl-survey-tool-suite (accessed on 31 January 2025).
24. ESS11 2023 Data Protocol. Available online: https://stessrelpubprodwe.blob.core.windows.net/data/round11/fieldwork/source/ESS11_data_protocol_e01_5.pdf (accessed on 31 January 2025).
25. ESS–Data Portal. Available online: https://www.europeansocialsurvey.org/data-portal (accessed on 31/01/2025).
26. ESS11 - Data from Contact Forms, Edition 2.0. Available online: https://ess.sikt.no/en/datafile/065b75fc-67c7-4f24-9c55-5a42fb6a2a21 (accessed on 31 January 2025).
27. ESS - Data Portal. Available online: https://ess.sikt.no/en/ (accessed on 31 January 2025).
28. ESS11 - Data from Interviewer's Questionnaire, Edition 2.0. Available online: https://ess.sikt.no/en/datafile/0c3f5fda-4fe0-42f2-b4c8-4ef093c0508f (accessed on 31 January 2025).
29. Data Documentation Initiative (DDI). Available online: https://ddialliance.org/ (accessed on 31 January 2025).
30. ESS Multilevel Data. Available online: https://ess.sikt.no/en/series/50ebb530-c72e-4002-a7bf-5581e3fdfe21 (accessed on 31 January 2025).
31. ESS Round 12 Survey Specification for ESS ERIC Member, Observer and Guest Countries. Available online: https://www.europeansocialsurvey.org/sites/default/files/2024-04/ESS012_projection_specification_v2.pdf (accessed on 31/01/2025).
32. Survey Weights in the ESS. Available online: https://www.europeansocialsurvey.org/methodology/ess-methodology/data-processing-and-archiving/weighting (accessed on 31 January 2025).
33. EOSC Future–SP9. Available online: https://ess.sikt.no/en/series/a1f81168-b864-4838-9701-39bbc6087e30 (accessed on 31 January 2025).
34. PIAAC Technical Standards and Guidelines June 2014. Available online: https://www.oecd.org/content/dam/oecd/en/about/programmes/edu/piaac/technical-standards-and-guidelines/cycle-1/Cycle_1_PIAAC_Technical_Standards_and_Guidelines_June2014.pdf/_jcr_content/renditions/original./Cycle_1_PIAAC_Technical_Standards_and_Guidelines_June2014.pdf (accessed on 31 January 2025).

35. Cycle 2 PIAAC Technical Standards and Guidelines. Available online: https://www.oecd.org/content/dam/oecd/en/about/programmes/edu/piaac/technical-standards-and-guidelines/cycle-2/PIAAC_CY2_Technical_Standards_and_Guidelines.pdf/_jcr_content/renditions/original./PIAAC_CY2_Technical_Standards_and_Guidelines.pdf (accessed on 31 January 2025).
36. PIAAC 1st Cycle Database–Questionnaires. Available online: https://www.oecd.org/en/data/datasets/piaac-1st-cycle-database.html#questionnaires (accessed on 31 January 2025).
37. PIAAC 1st Cycle Database. Available online: https://www.oecd.org/en/data/datasets/piaac-1st-cycle-database.html (accessed on 31 January 2025).
38. PIAAC 2nd Cycle Database. Available online: https://www.oecd.org/en/data/datasets/piaac-2nd-cycle-database.html (accessed on 31 January 2025).
39. PIAAC PUF Users Data Collection. Available online: https://survey.oecd.org/index.php?r=survey/index&sid=424913&lang=en (accessed on 31 January 2025).
40. PIAAC–Tools for Data Analysis. Available online: https://www.oecd.org/en/about/programmes/piaac/piaac-data.html#tools (accessed on 31 January 2025).
41. PIAAC–Data Explorer. Available online: https://piaacdataexplorer.oecd.org/ide/idepiaac/ (accessed on 31 January 2025).
42. PIAAC–Log Data Analyzer. Available online: https://piaac-logdata.tba-hosting.de/download/ (accessed on 31 January 2025).
43. GESIS Programme for the International Assessment of Adult Competencies (PIAAC), Log Files. Available online: https://search.gesis.org/research_data/ZA6712 (accessed on 31 January 2025).
44. Survey of Adult Skills (PIAAC). Available online: https://www.oecd.org/en/about/programmes/piaac.html (accessed on 31 January 2025).
45. PIAAC Methodology and Manuals- Technical Reports. Available online: https://www.oecd.org/en/about/programmes/piaac/piaac-data.html#manuals (accessed on 31 January 2025).
46. Nordic Network for Lifelong Learning - Access to PIAAC Data. Available online: https://nvl.org/artikler/access-to-piaac-data/ (accessed on 31 January 2025).
47. Gregory, A.; Wackerow, J.; Orten, H. Reuse and Reproducibility: Describing Cross-Domain Research Data in the Science Project Climate Neutral and Smart Cities. *ARPHA Preprints* **2023**, *4*, e115047.
48. F-UJI Automated FAIR Data Assessment Tool. Available online: https://www.f-uji.net/index.php (accessed on 31 January 2025).
49. Eurostat Database. Available online: https://ec.europa.eu/eurostat/web/main/data/database (accessed on 31 January 2025).
50. Howison, J.; Wiggins, A.; Crowston, K. G. Validity Issues in the Use of Social Network Analysis with Digital Trace Data. *J. Assoc. Inf. Syst.* **2011**, *12*, 767-797. [CrossRef]
51. Stier, S.; Breuer, J.; Siegers, P.; et al. Integrating Survey Data and Digital Trace Data: Key Issues in Developing an Emerging Field. *Soc. Sci. Comput. Rev.* **2020**, *38*, 503-516. [CrossRef]