Sustainable Quality in Data Preparation

BARBARA PERNICI, Politecnico di Milano, Milan, Italy
CINZIA CAPPIELLO, Politecnico di Milano, Milan, Italy
CARLO ALBERTO BONO, Politecnico di Milano, Milan, Italy
CAMILLA SANCRICCA, Politecnico di Milano, Milan, Italy
TIZIANA CATARCI, Università degli Studi di Roma La Sapienza, Rome, Italy
MARCO ANGELINI, Università degli Studi di Roma La Sapienza, Rome, Italy
MATTEO FILOSA, Università degli Studi di Roma La Sapienza, Rome, Italy
MATTEO PALMONARI, Università degli Studi di Milano-Bicocca, Milan, Italy
FLAVIO DE PAOLI, Università degli Studi di Milano-Bicocca, Milan, Italy
SONIA BERGAMASCHI, Università degli Studi di Modena e Reggio Emilia, Modena, Italy
GIOVANNI SIMONINI, Università degli Studi di Modena e Reggio Emilia, Modena, Italy
ANGELO MOZZILLO, Università degli Studi di Modena e Reggio Emilia, Modena, Italy
LUCA ZECCHINI, Università degli Studi di Modena e Reggio Emilia, Modena, Italy

Data preparation is crucial for achieving good data management following the four foundational FAIR principles—Findability, Accessibility, Interoperability, and Reusability. Processing datasets to achieve high data (and metadata) quality is mandatory in modern applications. However, the data preparation activities that are needed to reach such levels may easily become unsustainable due to, for example, resource intensity or scalability challenges. Moreover, some preparation efforts may become unnecessary if they result in negligible improvements or duplicate actions. This article examines the sustainability aspects of data preparation through the lens of a circular economy. Within the data landscape, this perspective encourages practices that minimize waste, extend the data life cycle, and maximize reuse in alignment with the FAIR principles. We explore these practices and their impact on selecting and configuring effective data preparation strategies to design sustainable, high-quality pipelines. To this end, we propose an evaluation model that integrates data quality metrics with sustainability parameters for human and computational tasks. Finally, we apply the model

This work has been supported by the MUR PRIN 2022 Project "Discount quality for responsible data science: Human-in-the-Loop for quality data". It also includes research funded by the Horizon Europe project enRichMyData (HE 101070284).
Authors' Contact Information: Barbara Pernici (corresponding author), Politecnico di Milano, Milan, Italy; e-mail: barbara.pernici@polimi.it; Cinzia Cappiello, Politecnico di Milano, Milan, Italy; e-mail: cinzia.cappiello@polimi.it; Carlo Alberto Bono, Politecnico di Milano, Milan, Italy; e-mail: carlo.bono@polimi.it; Camilla Sancricca, Politecnico di Milano, Milan, Italy; e-mail: camilla.sancricca@ polimi.it; Tiziana Catarci, Università degli Studi di Roma La Sapienza, Rome, Italy; e-mail: catarci@diag.uniroma1.it; Marco Angelini, Università degli Studi di Roma La Sapienza, Rome, Italy; e-mail: angelini@diag.uniroma1.it; Matteo Filosa, Università degli Studi di Roma La Sapienza, Rome, Italy; e-mail: m.filosa@diag.uniroma1.it; Matteo Palmonari, Università degli Studi di Milano-Bicocca, Milan, Italy; e-mail: matteo.palmonari@unimib.it; Flavio De Paoli, Università degli Studi di Milano-Bicocca, Milan, Italy; e-mail: flavio.depaoli@unimib.it; Sonia Bergamaschi, Università degli Studi di Modena e Reggio Emilia, Modena, Italy; e-mail: sonia.bergamaschi@unimore.it; Giovanni Simonini, Università degli Studi di Modena e Reggio Emilia, Modena, Italy; e-mail: giovanni.simonini@unimore.it; Angelo Mozzillo, Università degli Studi di Modena e Reggio Emilia, Modena, Italy; e-mail: angelo.mozzillo@unimore.it; Luca Zecchini, Università degli Studi di Modena e Reggio Emilia, Modena, Italy; e-mail: luca.zecchini@tu-berlin.de.

ACM 1936-1955/2025/12-ART23
https://doi.org/10.1145/3769120

in a comparative analysis of key data preparation methods, demonstrating its effectiveness in assessing sustainability and quality tradeoffs.

CCS Concepts: • **Information systems → Information integration;**

Additional Key Words and Phrases: Data preparation, data quality, sustainability

ACM Reference Format:
Barbara Pernici, Cinzia Cappiello, Carlo Alberto Bono, Camilla Sancricca, Tiziana Catarci, Marco Angelini, Matteo Filosa, Matteo Palmonari, Flavio De Paoli, Sonia Bergamaschi, Giovanni Simonini, Angelo Mozzillo, and Luca Zecchini. 2025. Sustainable Quality in Data Preparation. *ACM J. Data Inform. Quality* 17, 4, Article 23 (December 2025), 33 pages. https://doi.org/10.1145/3769120

1 Introduction

Data preparation is crucial to FAIRification, i.e., the process of making data FAIR [53], supporting the transformation of datasets into structured, high-quality resources that adhere to the FAIR [120] principles, and in particular of reusability, which is a key goal associated with the FAIR principles. Key preparation tasks, such as data cleaning, format harmonization, schema mapping, and metadata annotation, directly contribute to improving a dataset's findability, accessibility, interoperability, and reusability, thereby enabling responsible and efficient data sharing and reuse.

Despite its importance, data preparation is not always sustainable, primarily due to its high resource demands. It often involves computationally intensive processes, leading to significant energy consumption, especially when repeated across multiple projects without reuse. Moreover, some data curation activities must be performed manually, raising concerns about economic efficiency and scalability. The lack of optimized, standardized, and reusable pipelines also contributes to redundant effort and waste. Consequently, it is possible to state that FAIRification aims to promote better data management and reuse, however it can introduce sustainability challenges, as also noted in early discussions of the proposal [52, 120].

This article proposes an approach to make data preparation and, thus, FAIRification more sustainable. Sustainability, originally focused on mitigating the environmental impact of human activities in different sectors and the necessity of developing better living conditions [58], has evolved into a concept based on three factors or pillars: environmental, economic, and social. These dimensions form the backbone of discussions on this topic, and their interplay determines the trajectory of global sustainability initiatives [90]. In 2015, the United Nations, developing a new sustainable development agenda, aimed to overcome the previous attempts of integrating economic and social development with environmental sustainability, adopting a goal-setting strategy [21], with the establishment of 17 **Sustainable Development Goals** (**SDGs**). The important aspect of this initiative is the definition, for each goal, of targets and measurable indicators for being able to assess the evolution towards the achievement of the goals.

In recent years, the accelerating pace of the so-called "digital revolution" has underscored the critical role of digital technologies in achieving sustainability objectives. On the one hand, digital tools and technologies, including **Artificial Intelligence** (**AI**), offer significant potential to expedite attaining these goals. For instance, AI has been identified as a key driver in addressing challenges related to poverty, health, and education, among others (see, e.g., [113]). On the other hand, the rapid spread of digital technologies can worsen sustainability challenges if not responsibly designed and implemented with a holistic view of sustainability [106]. The concept of "digital sustainability" [23] promotes the responsible use of digital technologies by integrating environmental, social, and economic considerations across their life cycle, while balancing diverse stakeholder needs.

On the social dimension, digital technologies hold promises for driving transformative change. However, the social implications of digital technologies are multifaceted. Issues related to privacy, security, and human rights require careful consideration. The massive use of personal data by AI systems risks reinforcing inequalities and discriminatory practices.

Economic issues are equally critical in the discourse on digital sustainability. Digital technologies can drive productivity, foster innovation, and create opportunities for sustainable economic growth. However, the economic benefits must be weighed against the long-term costs associated with resource extraction, energy consumption, and waste management. Policymakers and organizations must adopt a holistic approach to ensure that economic growth does not come at the expense of environmental degradation or social inequities.

Digital sustainability is particularly critical when considering data analysis, **Machine Learning** (**ML**), and AI. Electricity consumption trends show a fast-growing percentage of use of electricity for IT and data centers[1] and projections show a considerable increase in the coming years.[2,3] Green IT, Green software [25] and, in particular, Green AI [97] and Sustainable AI [110] approaches are emerging, mainly focusing on energy efficiency and environmental impacts. Data preparation plays a central role in this context, as the growing demand for high-quality data to train models drives the emergence of **data-centric green AI (DCGAI)**. DCGAI emphasizes improving model performance and sustainability by optimizing data quality rather than focusing on model complexity.

Data preparation is fundamental for FAIRification, and its sustainability is increasingly recognized. Emerging guidelines emphasize the use of automated tools to curate tabular data, both in data and metadata [46]. Social implications are also increasingly arising when collecting and analyzing data, in particular concerning bias, misinterpretation, and incomplete coverage in data intended for multiple purposes. The sustainability of FAIR principles needs to focus on the entire life cycle, with human control integrated into the design and development processes [99].

The goal of this article is to introduce a framework and a modeling approach to evaluate the sustainability of data preparation methods. Considering that (i) it is largely recognized that the preparation of information products shares many characteristics with production processes [10] and (ii) circular economy is recognized as a way to achieve sustainable development in companies [116], this work proposes applying circular economy strategies as a compelling paradigm that we extend to foster sustainability in data preparation and FAIRification, adopting a broader concept of sustainable production and consumption in data preparation, which encompasses sharing, reducing, reusing, repairing, and refurbishing existing products to maximize the utility of their life cycle [61, 89].

A classification of sustainability strategies for data preparation is presented, focusing in particular on the impact of AI-based data preparation techniques, which are increasingly being adopted in this domain. According to the goal-setting approaches proposed for sustainability governance, we discuss and propose measurable indicators, focusing on tabular data, to evaluate and compare the sustainability of alternative data preparation methods. We discuss how assessment dimensions for sustainability aspects vary depending on the characteristics and context of each strategy.

The article is structured as follows. In Section 2, we introduce a running scenario to motivate the approach and as a basis for illustrating the proposed framework on concrete examples. The main contributions in the state-of-the-art related to the research discussed in this work are presented in Section 3. In Section 4, we present the proposed framework for introducing sustainability in the different data preparation phases, we discuss how data quality dimensions can be extended

[1]International Energy Agency, Report Electricity, 2024 https://www.iea.org, visited on 10 June 2025

[2]EPRI Electric Power Research Institute White paper on Powering intelligence, 2024 https://www.epri.com/research/products/000000003002028905, visited on 10 June 2025

[3]Amodei, D., Hernandez, D., Sastry, G., Clark, J., Brockman, G., Sutskever, I. (2018). AI and Compute. https://openai.com/index/ai-and-compute/, visited on 10 June 2025

and classified for considering sustainability aspects, and we propose an assessment model to compare data preparation methods. In Section 5, we exemplify the application of the approach to some selected data preparation methods, discussing their sustainability aspects. Limitations of the proposed approach are discussed in Section 6 and conclusions and future work are presented in Section 7.

2 Motivating Scenario

As a motivating scenario, we consider data analysis in the digital marketing domain, providing an extended version of a real-world use case implemented by a Spanish digital marketing company within the European enRichMyData project, an example of which is discussed in [33]. The company operates digital marketing campaigns on online advertising platforms, e.g., Google Ads, which place ads in the search results returned for specific keyword-based searches made by users as a result of a bidding process. The overall objective is to maximize the performance of the campaigns for a given budget, which can be measured with different KPIs; for simplicity, in our scenario, we consider *impressions*, which represent the number of times the ad associated with the campaign has been displayed in search results. Several data-driven methods are used to maximize impressions, which depend on campaign configuration aspects, e.g., choice of keywords and allocated budget, and external factors, e.g., trending topics, events, and so on. Data-driven campaign optimization can use visual analysis, leveraging visual analytics discipline and techniques [49, 59, 108], statistical models, and ML—all of which require intensive data preparation steps.

To ground these considerations into a concrete data preparation example, we consider a slightly revised example of data preparation required in the digital marketing domain in Germany [33]. In this context, data preparation is a massive operation involving the management of 5 billion active keywords and the generation of more than 1,000,000 clicks per day in 234 countries and 15 languages.

The input data, a snapshot of which is shown in Step 0 of Figure 1, consist of a table containing the number of impressions collected daily for the specific keywords (shown with a code) that have been selected for the campaign, for each locality (city and state are given). As shown in the figure, accuracy and completeness issues may occur in the input data, such as misspellings, missing values, and so on.[4] To prepare data for the analysis, two data-enrichment tasks have to be performed on the table: (i) getting information about the location to normalize data based on the population (state population in this case) and (ii) getting weather information at the location on that day (temperature in this example), as the campaign optimization is based on the weather conditions of the considered geographical areas, since it has been demonstrated that weather influences user purchasing behavior [33].

Data on the population at the state level can be collected from an open data source such as Wikidata [31]; weather data can be collected using weather services, e.g., Open Weather.[5] The data about the population may be affected by timeliness issues. Temperature and population data can also be affected by the presence of outliers. Further data issues arising in the weather service are related to the availability of data for the specified location and the requested ISO data format for locations, different from the format used in the input data. In addition, as locality data can be incomplete or inaccurate, some disambiguation actions might be needed. For example, a geocoding service can retrieve coordinates of cities, e.g., HERE,[6] while a Wikidata reconciliation service, e.g., OpenRefine,[7] can support the disambiguation of regions. Observe that these matching operations

[4]Data collected from Google Ads seldom present these issues, which may occur at a larger scale in other input data sources
[5]https://openweathermap.org/, visited on 10 June 2025
[6]https://www.here.com/platform/geocoding, visited on 10 June 2025
[7]https://wikidata.reconci.link/, visited on 10 June 2025

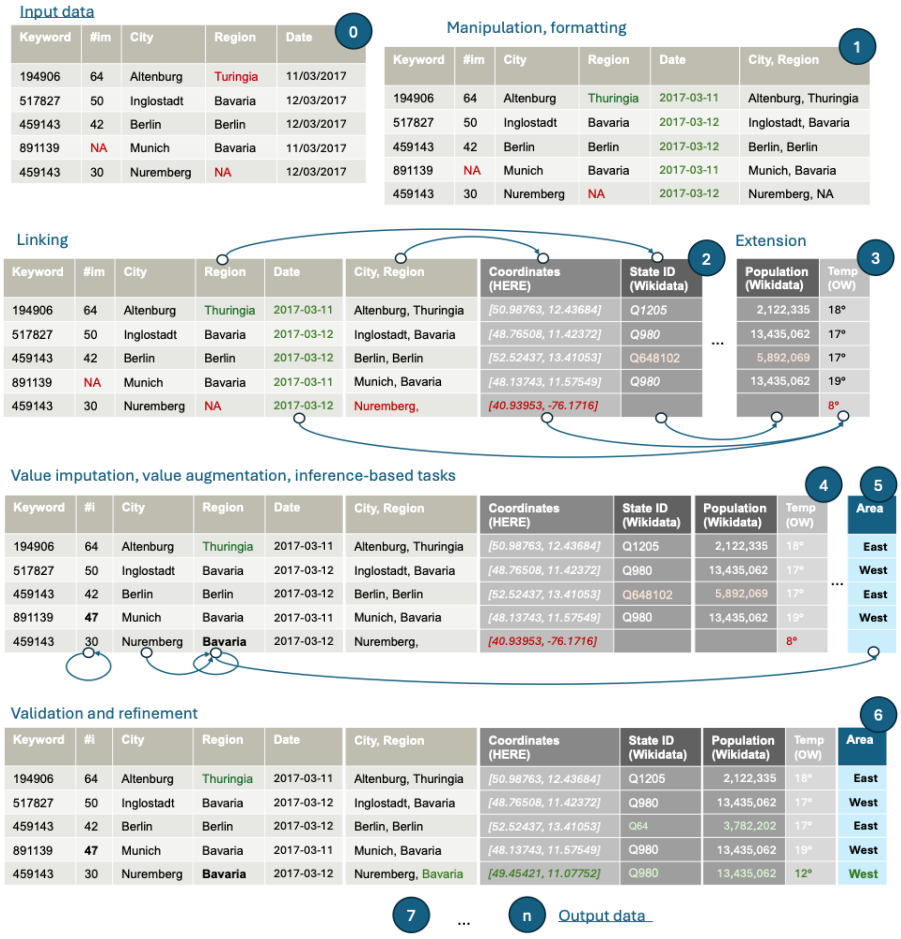

Fig. 1. Example of data preparation for optimizing digital marketing campaigns. Numbers identify data preparation steps. Column colors represent data from different sources. Arrows trace how columns contribute to derived content. Text colors indicate data quality: red for missing or incorrect values; orange for imprecise yet acceptable data; green for improved, reliable values.

are intrinsically subject to uncertainty and mistakes. A further enrichment action is based on the assumption that user behavior may differ in eastern and western German states; to this purpose, parametric knowledge contained in **Large Language Models (LLMs)** is exploited.

The following pipeline has been developed to perform the relevant preparation steps (see Figure 1):

Step 1: Manipulation and formatting. Input data is manipulated to correct spelling errors and convert date formats in accordance with the expected input format of the weather data service.

Step 2: Linking. Associating location-based values with the **identifiers** (**IDs**) necessary for retrieving additional data from third-party services. The downstream impact of matching errors varies in severity. For example, the geocoding service mislinks 'Nuremberg' to a U.S. location, producing entirely incorrect coordinates. In contrast, Wikidata reconciliation maps "Berlin" to the Berlin-Brandenburg Metropolitan Region[8] instead of the city itself. While the city might be a better match, the difference may have limited effect on downstream analysis.

[8]https://www.wikidata.org/wiki/Q648102, visited on 10 June 2025

Step 3: Extension—data enrichment. With collected identifiers and reformatted dates, it is possible to extend the table with data collected from third-party sources necessary to support the downstream analysis. Observe that the errors introduced in Step 2 are propagated here.

Step 4: Value imputation and other inference-based tasks. Missing data can be addressed using different techniques for value imputation, e.g., by computing the average or adding the region with manual intervention. Outliers must be identified with detection techniques and subsequently handled, for example, by removing or correcting them with imputation techniques.

Step 5: Augmentation. In this step, data are augmented by exploiting distinctive features of LLMs to classify regions as belonging to Eastern or Western Germany.

Step 6: Validation and refinement. Validation and revision steps are applied to improve the data. For example, if we run reconciliation and extension for Nuremberg once the correct region is specified, inaccuracies in coordinates, weather data, and east/west classification can be solved.

As shown above, data preparation challenges depend on the quality of the input data and of the external data sources or services being used, and may depend on the order of execution of the actions in the pipeline. In the following, we examine the aspects related to assessing the sustainability of a data preparation pipeline and its components and evaluating alternatives, focusing on the FAIRification process and reuse.

3 Related Work

With the recent spread of data analysis applications, we are witnessing a growing number of tools for helping users explore, clean, and analyze data. Several approaches have been developed to support the early stages of the data science pipeline, such as data exploration, profiling, and **Data Quality (DQ)** assessment. Data exploration and profiling are the focus of [34], which proposes a comprehensive survey of tools for profiling and DQ measurement and monitoring; [51] supports the identification and exploration of data inconsistencies. The article [123] proposes a reinforcement learning system to support data transformation (split, concatenate, index, etc.) and table manipulation operators (join, union, pivot/unpivot, etc.). However, most of the contributions focus on the design of data preparation pipelines [77] oriented to ML applications.

The role of data in AI development has gained increased importance with the advent of **data-centric AI** (**DCAI**) [55], which highlights the importance of having high-quality input data to obtain reliable results rather than improving models. In this direction, [86, 101] proposes an approach to support the exploration and the DQ assessment for data used in AI systems. [85] groups several tools for DQ analysis in the context of ML (e.g., Pandas Profilers, Amazon Deepqu, and IBM's Data Quality for AI), discussing their strengths and similarities. They also discuss potential factors affecting classification performance, such as missing data, outliers, duplication, correlation, feature relevance, label correctness, balance, or overlap. In addition to the previous contributions, the work described in [86] proposes a framework for data exploration and DQ for DCAI.

As regards data cleaning operations, [94] automatically generates data repairs through a probabilistic approach; in [19, 20], the design of a data preparation pipeline is driven by the maximization of the quality of the results of an ML model. The authors propose adopting a reinforcement learning approach (i.e., Q-learning) to explore the space of all the possible data-cleaning tasks to perform and select the best ones. In this framework, as stated above, the selection of the data preparation tasks is also driven by the quality of the results, but the recommendations are based on empirical evidence. Moreover, [54] focuses on data imputation and develops an automated ML-based approach for error detection and cleaning based on the estimation and balance of the introduced ML model uncertainty; [75] focuses on supporting non-expert users by automating several preprocessing activities; [86] extracts suggestions on relevant data preparation actions to apply through the assessment of potential data issues.

Widespread AutoML approaches also offer the possibility to perform data preparation automatically [40, 79, 98], but they offer only standard preparation strategies such as percentile-based outlier detection and mean/most frequent-based data imputation; other approaches leverage external or previously collected knowledge related to data preparation pipelines performed in the past to define and suggest promising pipelines for new datasets [29, 70, 71, 122, 124]; this knowledge can include, for example, the set of cleaning activities adopted in the past by expert data scientists, or previously cleaned datasets with their own set of metadata or profiling characteristics.

It is worth emphasizing that all the above-mentioned approaches aim to facilitate the preparation pipeline design by focusing on performance optimization without considering the amount of resources needed and the sustainability of the approach. Moreover, it is worth noting that in general data cleaning is not a fully automated process since many tasks require human intervention; thus, a set of interactive approaches has been designed to involve users while preparing data: [68] develops a data visualization process exploiting user feedback to find new data errors and suggest possible repairs; [72] builds an iterative process in which the data analyst progressively identifies DQ issues and addresses them; [62] proposes an iterative data cleaning approach to support statistical modeling. On the one hand, all these approaches oriented to data preparation emphasize the complexity of such a task, but on the other hand, they do not consider the sustainability aspect.

Sustainability has been addressed in Green IT, Green software [25, 112], Green AI [97], and Sustainable AI [110]. In particular, techniques for measuring the energy impact of IT applications and making them more sustainable have been proposed. To consider the environmental impact, the first step is in the direction of accountability [65], for which **metric tons of carbon dioxide-equivalent** (**MTCO2e**) are being adopted, as in many other fields. On the other hand, several other metrics have been proposed [115], and, in particular, in addition to direct consumption of energy, also idle time and wasted resources, e.g., in data centers, should be considered [73]. Reporting emissions is considered a first step in the sustainability direction. However, evaluating the environmental impact appears to be challenging [27], even if some efforts are being made, such as for instance training and usage estimations for BLOOM, a 176-billion parameter language model [67]. As reported in [111], the choice of the models can also significantly impact reducing energy consumption. Research directions towards efficient deep learning are discussed in [76].

The literature has scarcely considered energy consumption in data preparation. Assessing the carbon footprint of data preparation activities requires carefully analyzing key sustainability metrics in data preparation pipelines. Machado et al. [69] focus on data cleaning in ETL (Extract, Transform, Load) pipelines and emphasize that optimizing such processes for energy efficiency can significantly reduce their carbon footprint. Thus, monitoring and minimizing energy usage are essential for organizations implementing sustainable data practices. Furthermore, carbon disclosure metrics enable organizations to transparently report their emissions and sustainability initiatives, which are becoming increasingly important for regulatory compliance and stakeholder engagement. Janssen et al. discuss the significance of harmonized carbon disclosure metrics in the financial sector, an approach that can also be extended to industries reliant on ETL pipelines [56].

Data-centric approaches are starting to be used to improve AI energy efficiency, for instance, [111] presents experimental results showing that, with modifications on datasets, reductions up to 92.16% of energy consumption have been achieved, often with negligible or even absent accuracy decline. Examples of techniques that can be applied include carefully selecting and preprocessing data, choosing the appropriate samples and their size, feature selection, reducing the number of data points, reducing data redundancy, and avoiding overfitting. In general, data reduction emerges as an important direction. An in-depth analysis of statistical data reduction methods for sustainable deep learning is presented in [88], focusing on data reduction for image classification based on numerical tabular data for the images, discussing also metrics for evaluating different approaches.

Although limited to these specific data types, the article shows the potential of data reduction in training deep learning models.

In the present article, we attempt to provide a general framework for evaluating sustainability aspects in data preparation, focusing on strategies to sustainably extend the data life cycle. Since we have to guarantee the sustainability of their impact on human activities and socio-economic aspects, we extend quality dimensions in data preparation tasks beyond computational efficiency.

4 Framework for Data Preparation Sustainability

Preparing data involves multiple processing stages, often starting with the discovery of datasets, continuing through integration and transformation processes, and ending with the storage of the curated data. Although there is no universally defined sequence of steps for data preparation, several commonly supported functionalities are broadly recognized as standard practices. These include the tasks illustrated in the reference scenario, as well as common techniques such as data profiling, schema matching and mapping, deduplication, entity resolution, format transformation, and data repair [39]. Note that data preparation plays a critical role in supporting the implementation of the FAIR principles by ensuring that data are accurately cleaned, consistently structured, and enriched with appropriate metadata.

We can identify two main phases in data preparation: (i) *data inspection,* in which users explore the dataset to understand the characteristics of the values to be analyzed; (ii) *data improvement,* in which actions for enhancing the quality of data and metadata are performed, to ensure specific requirements or standards are met. This process often includes detecting and repairing errors, filling in missing values, standardizing formats, deduplication, enriching data with additional information, and optimizing its structure for better efficiency and reliability in analysis or operational use.

In this section, an approach for assessing and improving sustainability in data preparation is presented. In Section 4.1, we discuss quality dimensions relevant to sustainability assessment. Then, in Section 4.2, we introduce an evaluation model to compare methods concerning their sustainability and, in Section 4.3, we propose a circular economy perspective for framing data preparation activities aimed at FAIRification and we discuss relevant strategies and the application of the evaluation model in their context. In particular, strategies for smarter product use and manufacture are illustrated in Section 4.4, and strategies for extending the lifespan of products are discussed in Section 4.5.

4.1 Quality Dimensions

Quality dimensions relevant for assessing the sustainability of data preparation actions span traditional data quality, DCAI, and visual analytics dimensions.

4.1.1 Traditional DQ Dimensions. Traditional DQ models include multiple dimensions to identify and resolve issues from different perspectives, commonly related to values, schema, and usage [118]. Various DQ dimensions can be used depending on the context [11, 118], with the most common being *Accuracy*, *Completeness*, *Consistency*, *Uniqueness*, and *Timeliness* [12]. *Accuracy* is defined as the closeness between a data value v and a v' value, correctly representing a real-life phenomenon of interest [12] and is measured as the number of correct values over the total number of non-null values. *Completeness* reflects how comprehensively a dataset represents the real-world [12]; it can be assessed by the ratio of non-null values to total cells in a table. *Consistency* relates to the capability of the information to comply without contradictions to all properties of the reality of interest, e.g., integrity constraints, data edits, business rules, and other formalisms [12]. A way to assess it is to compute the number of violations of semantic rules defined over data, which can be either functional dependencies or business rules. *Uniqueness* reflects the amount of duplication and can be measured by counting duplicated records in a dataset [11]. *Timeliness* expresses how

current the data are for the task to be performed. Timeliness implies that data are current and in time for events that correspond to their usage [12].

4.1.2 DQ Dimensions in DCAI. Incorrect, incomplete, inconsistent, and duplicated values are known to affect ML results [45, 63, 92, 93, 128]. Beyond traditional DQ issues, it has been shown that data profiling characteristics directly influence ML and AI outcomes. A high *dimensionality*, i.e., the number of features and instances of a dataset, can potentially result in overfitting [2, 35, 66]. The presence of redundant or highly *correlated* features can also increase computational complexity, slowing the training time and causing overfitting [8, 81]. Classification tasks are particularly sensitive to issues such as noise, errors, or inconsistencies in the target labels [48, 114]; moreover, the target label must also be not *overlapping*, with similar entities associated with different labels [114], and not *biased* or *unbalanced*, with unevenly distributed values. If data contain bias, there is the risk that the model will give biased and unfair predictions, which can be discriminatory for under-represented groups [1, 105, 114]. Ethical aspects and fairness must also be considered when utilizing AI systems, especially in sensitive contexts such as healthcare, finance, justice, and education. Detecting and measuring bias in training data is then essential; for example, *coverage* metrics have been proposed to assess how faithfully a dataset represents the real world, measuring "the proportion of entities represented in the dataset relative to the number of real-world entities" [78]. Another way to detect the presence of bias is by measuring how densely concentrated specific values are within a feature; *density* is defined as "a measure of appropriate numerosity and intensity between different real-world entities available in the data" [78]. A measure of the value heterogeneity within a feature is *diversity*, defined as "the degree to which different kinds of objects are represented in a dataset" [43]. Other metrics have been proposed to assess whether predictions across different demographic groups are fair [74].

4.1.3 Visual Analytics Dimensions. These dimensions address key aspects of data quality that influence the effectiveness of visual analytics. Exploratory tasks in visual analytics integrate interactive exploration and human cognition to extract insights from complex datasets, and are linked to issues such as excessive latency [14] and usability concerns. Visual analytics dimensions specifically related to exploratory tasks are discussed in the following. *Granularity* strongly influences how data is explored and visualized. A high granularity in a well-designed visualization system provides a high level of detail, allowing users to explore disaggregated data and detect patterns or anomalies. Conversely, low granularity offers a comprehensive overview by reducing the level of detail while preserving critical information. Granularity should be controllable by users, providing "overview first, zoom and filter, then details-on-demand" [100]. *Latency* affects the fluidity of exploration, as delays in system responses can disrupt analytical workflows. It can be measured through system performance benchmarks (e.g., time-to-first-render, response times to interactions) and subjective assessments via user surveys, where delays above 500 ms are often reported as disruptive in the exploration [64, 117]. *User Control* directly relates to how much a user can steer the exploration in a visualization system. Another factor influencing cognitive load is the *mental effort*, which can be measured using subjective rating scales such as the Paas scale or physiological indicators like gaze shift rate [82]. A lower number of gaze shifts per second suggests better assimilation of visual information, indicating reduced cognitive load. Lastly, *performance efficiency* is a key indicator of cognitive load, which can be evaluated by analyzing task accuracy relative to completion time. Higher efficiency suggests a lower cognitive load, as users can process information more effectively [107]. System *Usability* quantifies how effectively, efficiently, and satisfactorily users achieve goals. According to ISO 9241-11:2018, *effectiveness* measures the accuracy and completeness of achieving goals; *efficiency* the resources expended relative to task success; and *satisfaction* the user's comfort and acceptability of use [50]. To measure usability, common metrics include the *success rate* of tasks,

time on task measuring task completion time, and the *error rate* [80]. *User satisfaction*, another important aspect of system usability, is often measured via post-task questionnaires or standardized tools like the **System Usability Scale** (**SUS**) [24].

4.2 Evaluation Model

This section illustrates the proposed model to assess the costs and benefits of data preparation activities, inspired by the model proposed in [9, 10], which considers the information products as the result of an information manufacturing process. In particular, we consider aspects related to sustainability to compare alternative data preparation approaches, focusing on tabular data—the most common data source—with individual table cells representing the data elements.

In the following, the index i denotes a *dataset*, j a *quality dimension*, and k a *method*; J denotes the number of quality dimensions considered in a given context, K indicates the number of methods being used, considering the use of a specific method in a given context. Each method is applied to a data element (e.g., to perform the imputation of a specific column, in some cases, multiple methods can be applied, depending on the element). The model considers the following elements, partially inspired by [10]:

- n: number of datasets.
- N_i: cardinality of dataset i, i.e., number of tuples. We use N when considering a single data source (in this case, the i index is omitted in the next sections for readability).
- $e_{i,j}$ rate of tuples to be improved (error rate) for dataset i and quality dimension j.
- $c_{i,j,k}$—improvement cost—unit cost of improvement actions for each element of dataset i and quality dimension j with method k.
- $ac_{i,j,k}$—assessment cost—unit cost of examining elements of the dataset i and quality dimension j with method k.
- $p_{i,j,k}$ effectiveness of method k in improving data for dataset i and quality dimension j with (percentage of actually improved data). Effectiveness should also be evaluated considering other dimensions $j' \neq j$ that may be impacted. For instance, data imputation improves completeness but lowers accuracy.
- $perc_{i,j,k}$: percentage of data being considered in the assessment and/or improvement process.

The cost of executing the actions is a combination of the costs of the following items: *Performance-related costs*: Execution time, Processing (e.g., FLOPS), RAM, Storage (volume); *Environmental impact*: Energy consumption, Carbon emissions; *Human resources*: human resources involved in data assessment and preparation **human-in-the-loop (HITL)**; *Cost of external services*. In [10], the cost to the organization of each undetected error in dataset i is also considered. An example of parameters that can be used to compute these costs and their derivation is illustrated in Appendix A.

The overall data quality improvement is defined as

$$DQImprovement_i = N_i \sum_{j=1}^{J} \sum_{k=1}^{K} perc_{i,j,k} \cdot e_{i,j} \cdot p_{i,j,k}, \tag{1}$$

where J data qualities are considered and K actions are applied.

To assess sustainability, we define three measures, $DQAssessmentCost_i$, for evaluating the cost of assessment actions, $DQImprovementCost_i$, for a general assessment of data quality improvement costs in preparation phases for a dataset i, and $DQWaste_i$, for assessing actions wasting resources.

$$DQAssessmentCost_i = N_i \cdot \textstyle\sum_{j=1}^{J} \sum_{k=1}^{K} perc_{i,j,k} \cdot ac_{i,j,k}, \tag{2}$$

$$DQImprovementCost_i = N_i \cdot \textstyle\sum_{j=1}^{J} \sum_{k=1}^{K} perc_{i,j,k} \cdot e_{i,j} \cdot c_{i,j,k}, \tag{3}$$

$$DQWaste_i = N_i \cdot \sum_{j=1}^{J} \sum_{k=1}^{K} perc_{i,j,k} \cdot e_{i,j} \cdot c_{i,j,k} \cdot \left(1 - p_{i,j,k}\right), \tag{4}$$

In the following sections, the model is adopted for evaluating and comparing the sustainability of alternative data preparation actions and pipelines.

4.3 Circular Economy Strategies

We propose adopting the main concepts of a circular economy for data management, focusing on addressing issues related to making data FAIRification sustainable. The strategies towards sustainability presented in the following are based on the 9Rs approach proposed by [89] for measuring the transition towards a circular economy in a product chain. Circular economy is defined in the article as "an economic system based on the reusability of products and product components, recycling of materials, and conservation of natural resources while pursuing the creation of added value in every link of the system." Three categories of strategies are discussed: "Smarter product use and manufacture", "Extend lifespan of product and its parts", and "Useful application of materials", detailing ten strategies to achieve them.

A *circular economy for data management* aims to maximize the value of the data life cycle while minimizing waste and inefficiencies. Such a concept is strongly linked with the FAIR principles: they share the same goals. Both perspectives promote efficient, sustainable, and responsible use of data. It is worth highlighting that data quality—and, therefore, data preparation—is a key enabler in this regard, as high-quality data favors reuse and sharing. Circular economy has the following goals: (i) *Maximizing versatility*: well-documented, accurate, and standardized data are more likely to support different uses and objectives; (ii) *Ensuring efficiency*: reducing redundant or unnecessary data collection, storage, and processing; (iii) *Extending the data life cycle*: reusing, repairing, and maintaining data for a longer time interval through proper curation, documentation, and updating. We propose using these goals as the primary drivers for designing data preparation pipelines, aiming to minimize waste and guarantee sustainability and cost savings.

We focus in particular on five strategies included in the categories *Smarter product use and manufacture* and *Extend lifespan of product and its parts* discussed in [89], while we do not consider other strategies in the category Useful application of materials (R8-9 Recycle and Recover), as they focus more on physical aspects of the products and would need a much broader discussion on components of the IT infrastructure, out of the scope of the present work (for possible directions, see, for instance, [67]). We are also not considering strategy Refuse (R0), defined within the Smarter product use and manufacture category, as it is out of the scope for this article, and strategies R6-7 Remanufacture and Repurpose within the category Extend lifespan of product and its parts, as they focus again on physical components in the production.

Considering data preparation actions within circular economy concepts, the category "Smarter product use and manufacture" corresponds to strategies in the *design phase* of data preparation, when the data products are being prepared to facilitate reuse, while the considered strategies in the category "Extend the lifespan of product and its parts" focus on actions performed during the *consumption* of a specific data product that, due to some quality issues, is no longer fit for immediate reuse. In the following, we discuss the characteristics of these strategies in the context of data preparation and outline the key aspects and criteria for assessing their sustainability in comparative terms, based on the evaluation model presented in Section 4.2.

4.4 Strategies for Smarter Product Use and Manufacture

R1 Rethink. *Rethink aims to design systems that can be used more intensively and durably.*

In the context of data preparation, the main aim of this strategy is to facilitate the sharing of data products at a large scale. Two main trends are currently gaining importance. First, *data sharing*

initiatives, which include open data portals, e.g., government data, data natively created to be reused across applications, e.g., Wikidata [31], and data sharing environments, such as data spaces [7]. This process of re-thinking the data production and consumption life cycle is supported by cross-national coordination activities, e.g., Europe has funded a Data Spaces Support Center,[9] and the cooperation of different bodies such as international associations, e.g., GAIA-X[10] and IDSA,[11] standardization organizations and consortia, e.g., ISO [44] and W3C [30].

Second, in recent years, evidence suggests that AI supports the development of solutions that can *perform multiple tasks* with one model. General models can be adopted as the basis for data improvement actions. Pieces of this evidence are emerging across tasks and include approaches based on encoders that perform multiple matching tasks [36], LLMs fine-tuned on tabular data or even generalistic LLMs to perform a variety of tasks (e.g., from entity linking to question answering) [16, 127].

Concerning sustainability evaluation, the data owners preparing data for sharing need to balance costs by monetizing reuse (e.g., in data spaces) or achieving a broader societal impact (e.g., in open data portals and other open resources), offering to their potential data consumers a better tradeoff between *DQImprovement* and *DQImprovementCost* for their pipelines.

On the other hand, considering ML-based approaches to curate data, sustainability can be improved by adopting general models as the basis for data improvement actions. While the development of such models is known to consume a large amount of resources, several techniques are proposed to improve the energy efficiency of model training and execution [121]. The costs incurred for training more general models can be amortized through their more intensive use and a reduced investment of resources in training models tailored to specific tasks. The $DQImprovementCost_i$ can be reduced by applying the data cleaning techniques only to data of interest. In addition to traditional DQ dimensions, among the DQ dimensions considered in DCAI, data imbalance dimensions need to be evaluated to avoid bias and ensure good coverage.

R2 Reduce. *In the taxonomy of circular economy, Reduce means to "increase efficiency in product manufacture or use by consuming fewer natural resources and materials" [89].*

Within this strategy, the main goal is to create methods and techniques to make data preparation more efficient: for example, efficiency can be reached by eliminating redundant or low-value data, focusing on collecting and managing data that aligns with specific quality criteria, working on samples, preparing only a subset of data of interest instead of the entire dataset or reducing the number of preparation actions to be performed, minimizing performance losses. In the context of dynamic (e.g., streaming or frequently updated) data, on-demand data exploration (through progressive visualization), data quality assessment, and data preparation can increase efficiency by reusing the same actions/functionalities for new incoming data; the amount of processed data in table integration can be reduced by selecting and aggregating only the relevant or high-value/quality data needed for specific purposes [125]; in this regard, the effort required for data preparation can also be adaptive, focusing on cleaning the most influential data issues not only for a specific data analysis application, but also based on the context.

A Reduce action in sustainable data preparation then emphasizes the efficiency in resource consumption while enhancing—or, at least, without compromising—data quality. Aligning with the broader principles of sustainability, efficiency is not solely about cost savings but also about lessening the environmental impact, which translates to efforts for minimizing the *DQImprovementCost* and *DQWaste* measures, as defined in the model, while controlling the *DQImprovement* measure.

[9] https://dssc.eu/, visited on 10 June 2025
[10] https://gaia-x.eu/, visited on 10 June 2025
[11] https://internationaldataspaces.org/, visited on 10 June 2025

Several types of Reduce actions can be leveraged. The resulting improvement could impact different aspects of the preparation tasks. Considering the prominent computational aspects of time and space, a Reduce action typically improves one of the two, directly affecting *DQImprovementCost*. The observed quality dimensions, or a subset, can then be enhanced, preserved, or controlled, possibly allowing a tradeoff between sustainability and the quality of the results. Moreover, the introduction of HITL models can be leveraged to benefit quality, computational resources efficiency, or both.

Regarding exploration tasks, efficient interactive visualization tools, possibly progressive, can trade precision (e.g., granularity and visual integrity) for benefits in computational time, and hence *DQImprovementCost*, during discovery and assessment phases. LLMs can be used to enhance data cleaning and enrichment, generally exchanging DQ dimensions (e.g., higher accuracy and completeness) for an increase in *DQImprovementCost* and *DQWaste*, which usually involves a tradeoff. When only a part of the data is relevant, on-demand data cleaning can greatly reduce the time to availability of processed data, increasing timeliness and accuracy with the advantage of limiting *DQImprovementCost* and *DQWaste* to the required operations. Moreover, minimizing the information overlap from datasets extracted from different sources can enhance data quality and optimize space requirements (e.g., consistency, uniqueness, dependency, and dimensionality) through reconciliation, matching, or join operations, resulting in an initial investment of *DQImprovementCost* which, however, results in a *DQImprovementCost* optimization along the use. Finally, the optimization of the processing steps, in their selection, order, and parameters, can improve the global execution time while controlling the final quality of the output, possibly enhancing it.

4.5 Strategies for Extending the Lifespan of Products

R3 Reuse. *Reuse, in terms of circular economy, is defined as "reuse by another consumer of a discarded product that is still in good condition and fulfills its original function" [89].*

In the present work, such a product can be either a data source or a sequence of already implemented actions (a pipeline or a subset of it). Reuse involves the efforts needed to create accurate descriptions of (meta)data to facilitate reuse or to make data accessible using paradigms that improve semantic interoperability (e.g., W3C recommended formats that support shared vocabularies such as RDF or JSON-LD, use of global (and local) identifiers, and APIs that simplify data access from third-party sources).

The main challenge in reusability is that reusability depends not only on data but also on its context of use *Reuse*(*Data*, *Context*). Some data sources are developed by design to be reused across applications (e.g., an encyclopedic knowledge graph like Wikidata or weather data served through an API). However, in most scenarios, data have been gathered for a given primary use, and reuse occurs in a different context (secondary use). Data preparation for reusability needs to be robust to context changes and allow for the evaluation of the differences between the context in which data is produced for its primary use and future contexts in which reuse can occur. For data preparation for reuse, we need to associate rich metadata to (i) facilitate the location of datasets for different users and applications and (ii) evaluate the data reliability and suitability on the basis of the data history and quality level. Metadata can be added in different stages. When the dataset is created for its primary use, profiling and quality metadata need to be stored. When the dataset is used, provenance information indicating previous processing tasks must also be provided (for instance, imputation with an average value could have been performed for primary use, but may not be suitable for a specific secondary use). To be able to reassess these metrics in secondary use, data enrichment should provide further information, including coverage and data granularity.

Different techniques and approaches can be adopted to provide these metadata, including the use of external services and semantic reconciliation. The challenge in adding the metadata needed for

Table 1. Methods for Data Preparation and Their Applicability to Sustainability Strategies

Methods	Strategies	Section
Progressive visualization	All strategies	5.1
Reconciliation-based data enrichment	Rethink, Reduce, Reuse, Repair	5.2
Cleaning on demand	Reduce, Reuse	5.3
Dynamic data cleaning	Rethink, Refurbish	5.4
Pipeline design	Reduce, Reuse, Repair	5.5

improving reusability is the cost of performing *a priori* enrichment actions on complete datasets, which involves not only performance costs, but also estimating other dimensions such as bias and coverage, which involve manual activities, so that visual analytics dimensions should also be taken into consideration.

R4 Repair. *Repair focuses on maintaining defective products so that they can be used with their original function.*

In data preparation, this can be achieved by cleaning data with quality issues, supporting a longer use and a higher value. In general, when cleaning is carried out based on a repair strategy, all quality dimensions may be relevant, depending on the context.

Sustainability for repair strategies requires a sustainable inspection phase, considering the sustainability of the exploration for the user. Progressive visualization can improve the exploration of the data in terms of assessment costs and usability dimensions, reducing the $DQAssessmentCost_i$. When considering improvement, the selection of additional sources and techniques is critical for reducing both $DQImprovementCost_i$ and $DQWaste_i$. Concerning costs, performance dimensions are crucial, while considering waste, the effectiveness of the applied methods has to be considered. A further aspect to consider for Repair is the construction of sustainable data preparation pipelines, particularly in selecting the data cleaning methods and their order.

R5 Refurbish. *Refurbish refers to restoring and bringing an old product up to date.*

In this strategy, the emphasis is on extending the data source with new data that have become available over time. The product can be a pipeline or an outdated data source, which can be enriched by adding new up-to-date information to extend its life cycle.

The incremental enrichment of previous datasets with new up-to-date data can increase the effectiveness of using the same dataset for the same or different purposes; this can be achieved by performing regular audits to maintain data updated. Differently from a Repair strategy, Refurbish mainly focuses on improving the timeliness dimension. For outdated data sources, sustainability aspects are similar to those discussed for data cleaning methods in repair. Additional challenges are posed for dynamic data, for which new data should be considered to prevent the reduction of the quality of data. In these cases, the goal is to consider performant methods but also to reduce $perc_{i,j,k}$ to decrease the number of data items to be considered and, therefore, $DQImprovementCost_i$.

5 Sustainability of Data Preparation Methods

We illustrate now how the proposed framework can be applied to assess the sustainability and related tradeoffs of a selection of data preparation methods, chosen to cover all five strategies (as synthesized in Table 1), explaining also the different sustainability concerns in the application of a method within different strategies. For each method, first, we illustrate the goals of the method and we discuss how tradeoffs can be evaluated with the proposed framework, then we illustrate with an example the assessments of possible alternatives.

5.1 Progressive Visualization

Data visualization is an activity that is performed during data preparation in a transversal way across strategies. In the data inspection phase, it reveals value distributions and enables steering, letting users interactively focus on quality issues and guide the exploration [83]. In data improvement, it displays maturity metrics (the proportion of data rendered so far) to determine when and where to apply fixes [91]. Visualization is also important in evaluating the results of analytic tasks performed on the prepared data. Reducing the costs involved in a visual analytics pipeline requires strategic tradeoffs between time, storage, data quality, and human involvement. *Progressive visualization* provides an effective means to mitigate these costs by delivering incremental updates in the displayed data, allowing users to engage with data earlier while balancing computational constraints. For instance, during the exploratory task in data visualization, data could be progressively fetched and rendered based on the level of detail the user is interested in (i.e., a user zooming by a great extent in a scatter plot), instead of directly loading all the data chunks at once, reducing the computational overhead in the visualization system and avoiding latency [5]. With progressive updates, avoiding long waiting times associated with fetching large volumes of data all at once, the users feel more in control of the system, reducing possible user frustration in data exploration, which often leads to errors or, worse, to the abandonment of the task. By providing users with immediate feedback and a gradual introduction to complex data, progressive visualization can facilitate learnability and user familiarity with the visualization system, allowing users to build an understanding of the system's functionalities and data structure over time. Managing cognitive load is another crucial challenge, as complex analytical tasks can become mentally demanding if users must simultaneously interpret evolving visualizations and maintain context across updates. Progressive visualization addresses this by summarizing information in early iterations and refining details progressively, although this approach introduces a tradeoff between visual stability and result quality, as initial updates may be incomplete or misleading, requiring careful uncertainty management [109]. Lastly, usability can be enhanced by carefully analyzing users' interaction patterns in the visualization system, analyzing *user traces*, which encompass low-level interactions performed by users, identifying, for instance, not yet explored portions of the visualization system or suffering from excessive latency [15, 17, 41]. Considering the proposed evaluation model, the goal is to select data granularity and data items to visualize a smaller number of items, reducing $DQAssessmentCost_i$ with a reduction of $perc_{i,j,k}$ in the exploration phase. Finally, although it is beyond the scope of the subsequent example, the cost of assessment $ca_{i,j,k}$ can also be reduced by devising efficient techniques considering the evolving visualization context of the users while exploring the data, progressively fetching the data based on progressive data analysis and visualization techniques [38, 109] or prefetching them based on predictive techniques of the user actions (see [13, 18]).

5.1.1 *Example.*

Design alternatives. To begin, the marketing analyst opens a dashboard, displaying a choropleth map, where each state is shaded according to its total impressions (lighter tones for fewer impressions, darker tones for more), and an interactive bar chart showing, for each state, how many city-level records still require lookup or imputation. As lookups or imputations complete, the map's state colors update and the bars shrink, giving the user immediate visual feedback on data-quality improvements and highlighting any outliers needing manual review.

Two alternatives are evaluated:

(i) **No Progressivity (bulk):** All city records and missing impressions are processed in a single batch. The map and bar chart remain static until the very end, at which point they are both fully refreshed, incurring the full assessment cost $DQAssessmentCost_{i,\text{bulk}}$;

Table 2. Tradeoff Comparison between Bulk (No Progressivity) and Progressive Approaches for Dataset i

Metric	Bulk	Progressive
API + Compute Cost	€16.75	€1.88
$DQAssessmentCost_i$	€513.05	€50.00
Latency to First Update	31 min	200 ms
Steering Corrections	2	10
Steering Time	1 min	5 min
Iterations	1	10
Maturity Progress	0 → 100% (single step)	0 → 100% (10 steps)
CO_2 Emissions	0.0033 kg	0.0009 kg
perc (processed at city granularity)	100 %	≈ 10 %

(ii) **High Progressivity (incremental):** The system first fetches only the records identified as "dirty" at state granularity, populating those city markers immediately. Then it runs five iterative imputation passes in mini-batches until all dirty records are processed. After each mini-batch completes, the map and bar chart update, letting the analyst intervene only on the remaining ambiguous cases. The goal is to minimize $DQAssessmentCost_i$ by reducing the fraction of rows processed in a single bulk pass ($perc_{\text{bulk},i}$) and instead handling more rows through small, iterative batches ($perc_{\text{prog},i}$).

Parameters. With reference to the scenario in Section 2, we consider examining 51,305 vs. 5,000 rows in Step 2 and 10,000 incomplete tuples in Step 4, in addition to the parameters in Appendix A.

Evaluation. The comparison of the two approaches is shown in Table 2.

We compare the two approaches across six key dimensions. First, *feedback latency*: the bulk workflow must process all 51,305 rows before updating the map or chart, taking approximately 0.514 h (≈ 31 min), while the progressive strategy delivers the first partial update on ≈ 5,000 "dirty" rows in under 200 ms, enabling immediate feedback and avoiding error propagation in the pipeline. Second, *compute and API cost*: bulk geocoding of 51,305 records at €0.0003 each plus 0.514 h of AWS at €2.73/h costs 51,305×0.0003 + 0.514×2.73 ≈ €16.75, while progressive geocoding of 5,000 records and 0.139 h of compute costs 5,000×0.0003 + 0.139×2.73 ≈ €1.88 (89 % saving). Third, $DQAssessmentCost_i$: bulk flags all 51,305 records, costing 51,305×€0.01 = €513.05; progressive flags only ≈ 5,000, costing 5,000×€0.01 = €50.00 (a 90 % reduction). Fourth, *steering effort*: bulk yields 2 corrections (1 min of labor), available only after 31 min of processing; progressive yields 10 targeted corrections (5 min of labor), each applied immediately as mini-batch updates arrive. Fifth, *maturity progression and iterations*: bulk jumps from 0% to 100% in a single iteration, while progressive climbs from 0% to 100% over 10 mini-batches, giving the analyst visibility into intermediate completeness. Sixth, *environmental impact*: assuming 0.0064 kg CO_2/h of compute, bulk emits 0.514×0.0064 ≈ 0.0033 kg, while progressive emits 0.139×0.0064 ≈ 0.0009 kg (a 73 % reduction).

Discussion. Underlying all gains is the *granularity management* of the progressive method. By first aggregating at the state level to isolate approximately 5,000 problematic rows and then processing only those at the city level, the method avoids unnecessary work on the remaining 46,305 rows, thereby enabling all of the above efficiencies. Indeed, in the bulk case every row is handled at city granularity ($perc_{\text{bulk},i} = 1$), whereas in the progressive case only ∼ 5 000 of the 51 305 rows are processed in detail, giving $perc_{\text{prog},i} \approx \frac{5000}{51305} \approx 0.10$.

5.2 Reconciliation-based Data Enrichment

A number of data preparation tasks for the enrichment of tabular data are based on matching and data fusion techniques. When an input tabular data source must be enriched with content from one or more third-party data sources, a possible choice is to obtain a local copy of those data

and apply data integration techniques [12]. An alternative approach, aligned with sustainability strategies, is based on ***entity reconciliation (ER)*** and fetches only the data relevant to the input table. With ER, we refer to a family of techniques that map values of an input table to shared systems of identifiers and can be invoked as services (e.g., W3C Reconciliation Service APIs[12]). Reconciliation is functional to data enrichment because it supports a "*link & extend*" approach, where links returned by reconciliation services are used to query third-party data sources.

In relation to **rethink** strategies based on data sharing initiatives, making reconciliation services (and associated data extension services) first-class citizens in these initiatives increases the chances that data sources can be used by consumers from different domains, thus also supporting data **reuse**. A flagship example of an interoperable data source used in our scenario is Wikidata, which provides access to a vast amount of encyclopedic knowledge using Semantic Web standards [30]. While Wikidata tradesoff some data quality aspects (e.g., consistency with conceptual schemas) for others (e.g., data size and coverage), it offers unified access to a vast range of real-world entity features, reducing *DQImprovementCost* for data enrichment (e.g., quick access to German states' population in the reference scenario). This tradeoff can be captured by the *DQImprovement* formula, where the percentage of data that are improved by enrichment can increase ($perc_{i,j,k}$ in Equation (1)), while the effectiveness of the enrichment may decrease because of data quality issues in the source ($p_{i,j,k}$ in Equation (1)). Moreover, API-based solutions supporting multiple input formats (e.g., the HERE geolocation service applied to get coordinates of cities in the reference scenario) minimize the need for expertise in handling complex data formats, and projections, significantly lowering *DQImprovementCost*, particularly in terms of human effort. While geocoding services like HERE may not traditionally be viewed as reconcilers, they may effectively serve that function by resolving input ambiguities and providing identifiers that enable integration with other sources, such as OpenWeather, in this scenario.

A second aspect of **rethink** concerns the impact of LLMs on reconciliation tasks. Although Wikidata includes a built-in lookup service that can be used for ER, the process can be challenging when dealing with massive knowledge bases containing hundreds of millions of items, such as Wikidata itself. Scientists have proposed several methods to improve ER for tabular data [32], shifting from heuristic or ML-based methods to multi-task frameworks like UNICORN [36] and general-purpose table understanding models such as TableLlama [32, 127]. LLMs fine-tuned for various semantic tasks on tabular data (e.g., ER and question answering) have outperformed specialized ER models across multiple datasets and settings [16]. Even more strikingly, top-tier generalist LLMs (e.g., GPT-4o) have surpassed fine-tuned models on this task [16]. This suggests that a single model, primarily designed for language-intensive tasks yet trained on code and tabular data, can now effectively support a wide range of use cases (increased $p_{i,j,k}$ in Equation (1)). However, this comes at a cost, particularly in terms of resource consumption, introducing another tradeoff between *DQImprovementCost* and *DQImprovement*.

The discussion on LLM-based ER highlights the relevance of the **reduce** principle. While GPT-4o offers unmatched performance [16], its high resource consumption, both in terms of execution time and cost, makes it unsuitable for large-scale data processing. Lighter variants, such as GPT-4o-mini, offer a significantly better tradeoff between *DQImprovementCost* and *DQImprovement*. However, smaller ML-based models can still outperform LLMs on certain datasets. To further reduce resource consumption, techniques like quantization have been applied to on-premise LLM-based reconciliation models such as TableLlama, addressing memory usage and execution time constraints. Similar considerations apply when assessing other methods to **repair** a defective dataset with data cleaning operations to improve accuracy and completeness.

[12] https://www.w3.org/community/reports/reconciliation/CG-FINAL-specs-0.2-20230410/, visited on 10 June 2025

Table 3. Comparison of Enrichment Approaches, Costs, and Waste

Method type	Matching method	LLM	Meas.	$p_{j,k}$	$c_{j,k}$ x1000	Impr. costs	Ass. costs	Total costs	Waste
Pairwise	GPT4o	✓	F1	0.8774	12.125	3,891	120	4,011	477
Pairwise	GPT4o-mini	✓	F1	0.8084	0.728	233	120	353	44
Reconc.	HERE	✗	Acc	0.998	0.3	96	0	96	0
Reconc.	Alligator + Geo	✗	Acc	0.8	0.266	85	0	85	17
Reconc.	TableLlama + Geo	✓	Acc	0.823	1.367	438	0	438	77
Reconc.	GPT4o + Geo	✓	Acc	0.862	5.2	1,668	0	1,668	230
Reconc.	GPT4o-mini + Geo	✓	Acc	0.804	0.312	100	0	100	19

5.2.1 Example.

Design alternatives. Let us consider linking operations on our digital marketing dataset, with reconciliation-based enrichment in Step 2, i.e., *cities-to-coordinates*, where city names are linked to their coordinates. We consider the following design alternatives:

(i) **Pairwise-based:** following a traditional data integration pipeline, the user will (1) search, understand and download data of cities and their coordinates online and (2) use an approach for pairwise record matching. We consider the two LLM-based methods evaluated in [87] that are more resource-aware, using GPT4o and GPT4o-mini as LLMs. It is reasonable to assume that pairwise comparison should be applied after some sort of blocking; for comparability with reconciliation-based methods, we hypothesize that top-50 most similar records are collected for comparison;

(ii) **Reconciliation-based:** we consider a *service-based reconciliation* method based on the HERE APIs and several methods based on *ER*. With HERE APIs, coordinates are retrieved using "city, state" patterns for queries. In ER, a reconciler for Wikidata is expected to disambiguate 50 candidates retrieved by an external service like in [16]; for this category, we consider Alligator (based on a neural network and not on an LLM), GPT4o, and GPT4o-mini, which return IDs used to retrieve the coordinates.

Parameters. We consider the improvement of a single dataset, the digital marketing dataset. Therefore, in all formulas, we drop the index i. We consider $perc_{j,k} = 1$ range) and $e_{j,k} = 1$ because all records must be enriched. The number of distinct values for cities in the data is 6,419.

Evaluation. Results for the evaluation are listed in Table 3.

$DQImprovement = 6{,}419 \times p_{j,k}$, j refers to the quality measure, and k refers to the matching method selected in the alternative. The list of considered matching methods is listed in the second column of Table 3. For estimating the effectiveness of all methods, we use average quality measures reported in related work, which differ: $j = F1$ of the positive class is used for pairwise matching [87] and $j = Acc$ (short for "Accuracy") for reconcilers [16]. In this specific setting, formulas for the two measures are similar but not fully comparable. For pairwise matching we consider scores reported for WDC [87], considering best one-shot prompts. For reconciliation-based matching, we provide estimates, fine-tuned models where relevant were considered. Also, for $DQImprovementCost$, calculations change according to the methods. For pairwise matching $DQImprovementCost = 6{,}419 \times 50 \times c_{F1,method}$ because we need to evaluate 50 pairs for each city; coordinates are collected during assessment. For reconciliation, $DQImprovementCost = 6{,}419 \times c_{Acc,method}$. For LLM-based solutions, costs are estimated similarly as in pairwise matching; for on-premise solutions, estimates are based on average inference time per entity and prices of required machines, i.e., with GPUs for TableLlama and w/out GPUs for Alligator [16].

Concerning $DQAssessmentCost$, reconciliation-based data enrichment approaches do not require a proper assessment step, as enrichment is only mediated by the reconciler; in contrast,

for approaches based on pairwise matching, the user must search online and collect the data beforehand (e.g., from Simplemap.com or Worldcitiesdatabase), which requires manual labour for data exploration and query fine-tuning. Therefore, *DQAssessmentCost* is only considered for pairwise matching. In this case, the cost $ac_{j,k}$ is calculated per country (Germany, Spain, and USA), that is, $N = 3$, $j = fitness$ (the data is suitable for matching and has coordinates), $k = labour$ at €40.00 per hour. For each country, we estimated about 40 minutes of manual labour to search online, select Wikidata, customize the queries, and assess the results.

Finally, *DQWaste* is computed by using the above-discussed parameters for each method and quality measure in the $c_{j,k}$ and $p_{j,k}$ terms.

Discussion. The table presents several tradeoffs. Pairwise matching adds significant assessment costs due to the need to search the data and understand its structure, due to country-level specificity. In this case, reconciliation based on HERE APIs offers exceptional quality at a limited cost. However, in other data enrichment scenarios (e.g., enrichment of company data), such vertical, high-quality reconciliation APIs may not be available. An ER method like Alligator can offer a reasonable quality-cost tradeoff; in addition, it could be deployed on a commodity server, further reducing data improvement costs. However, another interesting tradeoff is offered by using cheap instances of top-tier LLMs like GPT4o-mini. These methods are particularly amenable if we consider that they are used in zero-shot settings and do not require fine-tuning and have been shown to generalize much better than previous methods on unseen data [16, 87].

5.3 Data Cleaning on Demand

In recent years, the paradigm for data integration and cleaning has more and more shifted from ETL (Extract, Transform, Load), where raw data are collected, cleaned, and stored in a data warehouse, towards ELT (Extract, Load, Transform). In this scenario, a huge amount of data—on which the ETL approach would be prohibitively expensive or technically unfeasible—is collected and directly stored in its raw form, e.g., in a data lake. Then, practitioners can integrate and clean portions of this large data corpus according to the needs of the task at hand.

Operating with huge amounts of data poses several challenges to the traditional data cleaning paradigms. In many situations, it is not possible (or at least not efficient) to apply expensive and time-consuming offline approaches to the entire data: *DQImprovementCost* is too high. In particular, this happens when practitioners need to deal with tasks characterized by *(i)* an information need, i.e., only a portion of the data is relevant to the task, hence cleaning the entire data is a waste of resources, and/or *(ii)* time constraints, i.e., time is limited and decisions based on the clean data have to be made quickly before data gets outdated. For instance, when dealing with dynamic sources such as Web data, we would like to be able to transform only the relevant portion of the data and return results promptly. Under this scenario, the user is not interested in the entire dataset, thus for any subset of interest there will be a $DQWaste_i$ corresponding to the cleaning of the unused data, represented by $(1 - perc_{i,j,k})$. To apply a **reduce** strategy to data cleaning, novel *on-demand* solutions need to be designed to clean only the data actually needed for the task at hand while returning results promptly, often in a pay-as-you-go fashion. This allows saving time, resources, and money (e.g., for pay-as-you-go contracts), producing results in a limited amount of time so that they can be useful to the practitioner. This comes with a minimal *DQAssessmentCost*, required to assess the priority of the data to be prepared (e.g., sorting records by similarity in a matching step).

An example of this paradigm shift is represented by entity resolution (a.k.a. deduplication or record linkage), one of the core challenges in data integration and cleaning [28]. Entity resolution is typically employed as an expensive offline cleaning step applied to the entire dataset before being able to use it. Building on previous solutions designed to overcome the limitations of this offline approach, such as progressive [42, 84, 102, 119] and query-driven [3, 4] methods, BrewER

[103, 126] performs entity resolution on-demand driven by an SQL SP (selection and projection) query expressing the interest of the practitioner. BrewER evaluates such queries on dirty data and returns results as if they were issued on the cleaned version of this data, allowing users to run clean queries on dirty data. In particular, this is done progressively, incrementally returning the most relevant entities to the practitioner while avoiding as much as possible matching and resolving entities that are not part of the query result.

BrewER supports stop-and-resume execution and allows saving the partially cleaned dataset after each query, reusing the results of previous queries and avoiding cleaning the same entity multiple times. This allows saving and **reusing** data at different quality levels, using, for instance, different matchers that operate according to different matching criteria (e.g., two products of the same model but different colors might be considered as the same entity for training a price prediction algorithm, and as different entities for a unified online catalog).

5.3.1 Example.

Design alternatives. We consider a data integration task involving a campaign dataset *(i)* containing ad impressions linked to location, product, and demographic attributes. The goal is to enrich and deduplicate entities (e.g., the same city might appear with slightly different values for the name and state attributes, such as "München" or "Munich," and "BY" "Bayern," or "Bavaria") for campaign analysis across states.

Following the *reduce* strategy, we compare two cleaning alternatives:

(i) **Full Cleaning (Offline/Batch):** Entity resolution and data enrichment are applied to the entire dataset upfront, regardless of downstream query interest.

(ii) **On-Demand Cleaning (e.g., with BrewER):** Cleaning is performed only for the subset of data required to answer a given SQL query (e.g., impressions in the top 5 states with the highest click-through rate), using progressive entity resolution and enrichment as in BrewER.

This design choice directly affects both the *Data Quality Improvement Cost* ($DQImprovementCost_i$) and *Data Quality Waste* ($DQWaste_i$), since full cleaning processes the entire dataset, including data not used in any query.

Parameters. We assume the size of the dataset with impressions over 50 states is $n_i = 200{,}000$. Approximately 20% of records contain missing or ambiguous fields. The query selectivity of on-demand query targets the top 5 states by click-through rate, corresponding to $perc_{i,j,k} = 0.15$.

Cleaning costs are as follows. Entity resolution: €0.01 per pair of record with LLM API call (c_{match}); Enrichment via API: €0.0005 per record (c_{enrich}); Human review (if needed): €0.50 per item. A timeliness constraint requires results within 10 minutes.

We can evaluate as sustainability metric the CO_2 emissions computed from compute time, human involvement, and API usage. For the sake of brevity, we report the full example only for the latter.

Evaluation. For the full cleaning strategy, $DQImprovementCost_{i,\text{full}} = n_i \times (c_{\text{match}} + c_{\text{enrich}}) = 200{,}000 \times (0.01 + 0.0005) = €2{,}100$ and $DQWaste_{i,\text{full}} = (1 - perc_{i,j,k}) \times DQImprovementCost_{i,\text{full}} = 0.85 \times €2{,}100 = €1{,}785$.

For on-demand cleaning with BrewER, only $n_{i,j,k} = 30{,}000$ records are cleaned, with $DQImprovementCost_{i,\text{on-demand}} = 30{,}000 \times (0.01 + 0.0005) = €315$.

Progressive resolution returns first results within 2 seconds and completes within 6 minutes, significantly reducing the costs compared to the batch approach, which must clean the entire dataset upfront to produce the same results.

Discussion. The on-demand approach yields a cost reduction of approximately 85% (from €315 to €2,100) and minimizes waste, as only the relevant data is cleaned. It also enables earlier access to partial results via progressive execution, enhancing responsiveness. While full cleaning ensures completeness, it suffers from high cost and latency. The benefits of on-demand strategies are particularly evident when query focus is narrow, data volumes are large, or timeliness is critical.

Moreover, the choice of an on-demand cleaning strategy should consider the intended use of the cleaned data—whether it supports a one-time analysis by a single user or serves multiple users over time—which may call for adapting the workflow accordingly. For instance, a single user might perform minimal cleaning just sufficient to complete their immediate task, even if the resulting data is not fully standardized or reusable by others. Contextual factors—such as query frequency, reuse potential, and result precision—can influence the relative effectiveness of each approach. Overall, the on-demand strategy with progressive execution represents a sustainable and efficient alternative to traditional cleaning pipelines.

5.4 Cleaning and Enrichment of Dynamic Data

In **Retrieval-Augmented Generation** (**RAG**) pipelines, **refurbishing** plays a crucial role in ensuring that heterogeneous tables remain relevant and usable over time for AI consumption—the *generation* phase is notably performed with an LLMs [37]. By progressively retrieving and enriching data with up-to-date information, these pipelines can adapt to evolving datasets without requiring costly re-training. This continuous refurbishment process helps standardize formatting differences, resolve inconsistencies, and correct errors, ultimately improving the reliability and effectiveness of data retrieval and augmentation. These challenges arise because tables often originate from diverse sources, with varying schemas, formats, and levels of data quality. The end scope of RAG is to provide the right portion of data to a generative model (e.g., LLM); so, similarly to the data cleaning on demand, it is important to assess what part of the data is useful before retrieving it. Thus, *DQWaste* is proportional to the amount of data retrieved and prepared that is not exploited in the generation of the final answer. It should be noted that RAG's generation phase leverages the general skills of LLMs—their ability to handle multiple data formats and perform different tasks from entity linking to question answering. These broad capabilities align well with **rethink** strategies (Section 4.4) that design systems for intensive product usage, making them more frequently applicable, thereby helping to justify the costs of LLM development.

A typical RAG workflow begins by indexing data with embeddings, enabling efficient similarity-based retrieval of relevant information [47]. Once the data is retrieved, it can undergo processing and integration steps, such as entity resolution, missing value imputation, reconciliation, or outlier detection and removal, depending on the specific requirements of the task. By incorporating data cleaning on-demand into this workflow, the cleaning operators can be selectively and iteratively applied only to the data portions that are needed, ensuring that the preparation process remains efficient and converges dynamically as new data is retrieved. This approach not only reduces *DQWaste*, but also adapts seamlessly to the evolving needs of the pipeline.

To further optimize the reuse of heterogeneous tables, dynamic retrieval and preparation techniques are employed to evaluate and enhance the quality of the data as needed. For example, Self-RAG [6] introduces the idea of evaluating the quality of generated outputs using "reflection tokens" to assess when the data retrieved in the RAG pipeline no longer meets quality standards. This concept can be applied to tables by triggering retrieval to enrich the data whenever inconsistencies or missing values are detected, ensuring that the retrieved tables meet the required standards for secondary use. Similarly, Flare [57] monitors the confidence level of the model during the generation phase and triggers retrieval when confidence drops below a certain threshold. This approach can be adapted to tables by initiating additional retrieval when the data quality or coherence of the table decreases, ensuring that the preparation process only focuses on improving the data when necessary. Extending the RAG pipeline, *agentic RAG* [104] has been introduced to leverage autonomous agents to enhance flexibility and decision-making within the pipeline. By leveraging agents, we argue that RAG can be extended with table retrieval and preparation on demand, progressively driven by the need to improve the quality of final results. Thus, for

some applications, this approach may converge with only a few iterations, whereas others with stricter (quality) requirements might necessitate more dynamic retrieval cycles. In such a setting, an agent can take several forms: a general-purpose conversational LLM, an LLM equipped with tool-using capabilities (e.g., a data-cleaning-on-demand tool), or a specialized agent acting as a judge or executor, capable of evaluating and computing data quality metrics independently of an LLM. These agents operate iteratively, retrieving data and applying data preparation progressively and on-demand to address the specific heterogeneity and inconsistencies in the retrieved tables. To address ambiguities and inconsistencies in the retrieved tables, techniques like *ToC* (Tree of Clarifications) [60] can be employed. *ToC* helps guide agents through iterative query refinement, progressively resolving uncertainties and ensuring retrieved data aligns with the intended context. This allows the agents to refine their data retrieval strategy and improve the quality of the tables through multiple cycles, ensuring they are suitable for reuse. The iterative process continues until the agents achieve a satisfactory result or the system reaches a predefined resource budget, such as a limit on API calls. This agent-driven approach ensures that data preparation and retrieval are tightly coupled and responsive, ultimately enabling the pipeline to converge effectively while tackling the unique challenges posed by diverse and imperfect data sources.

5.4.1 Example.

Design alternatives. We consider a data integration task involving a campaign dataset containing ad impressions linked to location, product, and demographic attributes. The goal is to retrieve, enrich, and deduplicate relevant entities (e.g., locations, product codes, and audience segments) to support campaign performance analysis across states. The dataset is heterogeneous and collected from multiple sources (e.g., marketing platforms and demographic providers), often differing in schema and quality. We compare two alternatives for retrieving and preparing tabular data for generation in an RAG pipeline:

(i) **Static Retrieval with Pre-cleaning:** All campaign-related tables are pre-cleaned and standardized offline before indexing. During a query, static retrieval fetches relevant tables, already in their cleaned form.

(ii) **Progressive Retrieval with On-demand Cleaning (Agentic RAG):** Retrieval and cleaning are interleaved progressively. The system dynamically retrieves a minimal set of candidate tables based on embeddings and progressively cleans only those required for the task, applying operations such as imputation, resolution, or schema alignment as needed.

Parameters. The number of candidate tables is 500 (location stats, product impressions, demographics, etc.), with an average of 400 rows per table. Only 10% of tables are relevant per query (i.e., $perc_{i,j,k} = 0.10$).

Cleaning costs include: Entity resolution (location/product): €0.01 per record pair (c_{match}); API enrichment (demographics): €0.0005 per row (c_{enrich}); Missing value imputation: €0.002 per row.

Agent initiates cleaning on tables with confidence below 0.7 or when placeholders (e.g., "N/A") exceed 10% of entries.

The sustainability metric being considered in this example is CO_2 emissions computed per API call and compute duration.

Evaluation. In the **Static approach**, all 500 tables are cleaned prior to indexing, with $DQImprovementCost_{\text{static}} = 500 \times 400 \times (0.01 + 0.0005 + 0.002) = €2{,}500$ and $DQWaste_{\text{static}} = (1 - perc_{i,j,k}) \times DQImprovementCost_{\text{static}} = 0.90 \times €2{,}560 = €2{,}304$.

In the **progressive agentic approach**, only 50 relevant tables are cleaned based on agent-driven retrieval, with $DQImprovementCost_{\text{progressive}} = 50 \times 400 \times (0.01 + 0.0005 + 0.002) = €250$.

Initial partial results are returned within 1.5 seconds, with three progressive cleaning passes converging within 8 minutes, significantly reducing costs compared to the static approach, which requires cleaning all tables upfront to produce the same results.

Discussion. The progressive agentic RAG approach achieves a 90% reduction in data cleaning cost (from €2,560 to €256) and minimizes *DQWaste* by cleaning only what's necessary to answer the query. This approach also enables early access to intermediate insights (e.g., top states with low conversion despite high impressions), helping analysts steer the process. In contrast, static pre-cleaning guarantees consistent data quality and faster generation at query time, but incurs high upfront cost and cleaning waste. The effectiveness of the progressive approach depends on query selectivity, table reuse potential, and the confidence thresholds used by the agents. In exploratory scenarios or continuously updated datasets, the progressive method offers a sustainable and responsive alternative.

5.5 Data Preparation Pipelines Improvement

In this section, we discuss sustainability challenges in data preparation pipelines, which consist of sequences of tasks that transform and filter data from a source dataset. Sustainability is crucial for pipelines that are **reused** and executed periodically. Several factors impact sustainability, including (i) task execution order, (ii) the input dataset, and (iii) the target downstream analysis.

Regarding (i), optimizing the execution order supports a **Reduce** strategy, while tracking pipeline metrics (e.g., via a data provenance model) enables future **Reuse** of filters and configurations, extending pipeline life cycle and applicability. Suboptimal task ordering can escalate resource consumption by directly impacting $DQImprovementCost_i$ and $DQWaste_i$. For example, unnecessary elements may be processed, as seen in filtering in filtering unstructured social media data [22]. Moreover, adjusting confidence thresholds in filters directly impacts DQ dimensions, balancing accuracy, completeness, and dimensionality against $DQImprovementCost_i$ and $DQWaste_i$.

Regarding (ii) and (iii), depending on the dataset characteristics and the downstream task, specific data quality dimensions may have varying impacts on the resulting performance [96]. Thus, a possible **reduce** strategy is to prioritize data preparation actions that enhance DQ aspects that most affect such performance, leaving less impactful dimensions unaddressed. The goal is to maximize model performance while minimizing the $DQImprovementCost_i$, as only selected data preparation actions are applied, reducing the cardinality of the K applied methods [96].

Moreover, **reduce** strategies can optimize resource usage in the pipeline when relying on external sources, such as geospatial identifiers, linked data, weather, and population data. Techniques like caching and pre-downloading reduce resource consumption by shifting data closer to execution [33], and improve efficiency by minimizing access time. For instance, storing yearly weather data locally avoids frequent online queries, lowering processing costs at the expense of storage and also aligning with **reuse**. The tradeoff between caching and pre-downloading depends on factors like analysis frequency, data size, and volatility, impacting *DQImprovementCost*.

5.5.1 Example.

Design alternatives. We consider a ML-based regression task to predict the number of *impressions* associated with certain conditions, such as choice of keywords and allocated budget, using the dataset of our scenario. We focus on *Step 4*, which involves imputing missing values and identifying and correcting outliers, thus considering two DQ dimensions, completeness and accuracy. We assume that by providing the regression task and dataset characteristics as input, the system proposed in [96] orders the dimensions according to their impact, concluding that completeness issues affect the results more significantly than accuracy issues. We consider two alternative pipelines:

(i) $\mathcal{P}$ is the **complete pipeline**, in which incomplete tuples and outliers are addressed by applying imputation and outlier detection/correction techniques in the suggested order (completeness → accuracy);

(ii) $\mathcal{P}^*$ is the **partial pipeline** in which, following a **reduce** strategy, only the most important DQ dimension—completeness in our scenario—is improved [96].

Parameters. We assume the size of the dataset with impressions over 50 states is $n_i = 200{,}000$ and that approximately 20% of records contain missing or ambiguous fields. All the tuples have to be improved ($e_{i,j} = 1$).

Cleaning costs for automatic missing value imputation and for outlier detection and correction are €0.002 per row.

We consider the effectiveness $p_{i,1}$ of the partial pipeline and $p_{i,1} + p_{i,2}$ of the complete pipeline. Such effectiveness can be estimated using previous experiments, in which several pipelines were executed on diverse datasets. We compute the achieved effectiveness $p_{i,j}$ as the ratio of the ML performance (e.g., F1-score) achieved after improving one or two dimensions compared to the performance achieved by running the total pipeline. Averaging all results obtained from previous work [96]: $p_{i,1} + p_{i,2} = 0.8937$ and $p_{i,1} = 0.8943$. Results highlight that, when the contribution of improving further dimensions is marginal, completing the pipeline can worsen performance since applying additional improvements introduces approximated values.

Evaluation. For the original pipeline $\mathcal{P}$, $DQImprovement_{i,\mathcal{P}} = 200{,}000 \times 0.2 \times 0.8937 = 35{,}748$ and $DQImprovementCost_{i,\mathcal{P}} = (200{,}000 \times 0.2 \times 0.002) + (200{,}000 \times 0.2 \times 0.002) = €160$

For the reduced pipeline $\mathcal{P}^*$, we have $DQImprovement_{i,\mathcal{P}^*} = 200{,}000 \times 0.2 \times 0.8943 = 35{,}772$ and $DQImprovementCost_{i,\mathcal{P}^*} = 200,000 \times 0,2 \times 0.002 = €80$.

Discussion. We can observe that, in specific contexts, improving only the most important dimension can roughly achieve the same average *DQImprovement* as the complete pipeline. As a result, when the two pipelines share a similar improvement, the pipeline $\mathcal{P}^*$ can save up to 50% of the total *DQImprovementCost*. The ability to rank the importance of the dimensions is based on the availability of a knowledge base related to historical experiments [96].

6 Limitations and Threats

The main goal of this article is to propose a systematic framework to assess alternative data preparation methods considering strategies for smarter product use and manufacture and to extend lifespan of product and its parts within a circular economy approach.

In general, assessment presents several challenges. As discussed in [26], the proposed FAIR assessment metrics present heterogeneity and subjectivity, with discrepancies in their intended use and context dependency. Also sustainability evaluation of IT systems presents several issues, as discussed in detail, for instance, in [67], as some factors can vary between operational settings, for example, local energy mixes, human labor costs and effort evaluation, and hardware efficiency. This limitation is partially mitigated by the contextual nature of the proposed approach.

Concerning data types, while this work has focused on structured, table-based data, other data types, such as text, images, or sensor streams, could need further metrics to assess different sustainability challenges, e.g., related to storage formats, processing complexity, and annotation requirements. While some challenges posed by dynamic data have been introduced in Section 5, data modality, velocity, and infrastructure variability could hinder the applicability of the proposed framework to large-scale, streaming, or high-frequency scenarios, which are linked to dynamic resource usage and time-sensitivity, as opposed to batch processing, and require further research.

As organizations could prioritize accuracy and speed over sustainability, practical adoption challenges arise. The proposed framework accounts for this possibility by explicitly modeling the tradeoff between sustainability, quality, and operational efficiency metrics. These metrics could be directly integrated into organizational KPIs and regulatory reporting frameworks, as forms of incentives towards sustainable data practices, aligning them with regulatory requirements or financial benefits (e.g., carbon credit).

Among the possible threats, changes in service availability and pricing models could not only disrupt pipelines, but also their sustainability. In general, pipeline robustness should also be evaluated, and possible mitigation strategies, such as redundancy and fallback services, should be considered.

7 Concluding Remarks

This article offers a novel view on data preparation issues by highlighting how a sustainability-based strategy is crucial to balance quality, efficiency, and environmental impact. Indeed, through the adoption of strategies inspired by the circular economy, it is possible to improve data management while reducing waste and computational costs. The integration of these concepts with FAIR principles ensures a more responsible and long-lasting approach to the use of digital resources. A key contribution of the article is a framework that adopts circular economy principles to extend the data life cycle, reduce waste and improve the data quality through specific strategies such as Rethink, to create reusable datasets and sharing tools; Reduce, for minimizing resources usage by eliminating redundant or useless data; Reuse, to exploit existing datasets and data preparation pipelines; Repair, for continuously cleaning and validating data to keep them usable; Refurbish, to update and enrich datasets to ensure their relevance over time. The article also presents the practical application of the general framework, discussing how specific methods and techniques could be the concrete counterparts of the principles. Examples of the evaluation of the impact of the context of the use of the methods for different strategies are presented to illustrate the application of the method and discussed as a basis for future work. Although complex to assess in detail and context-dependent, these evaluations can lead to the development of methods and tools to optimize the choice of actions for data preparation taking sustainability constraints into consideration in the development of data preparation support systems. The ability of the proposed framework to quantify sustainability dimensions, linked to circular economy strategies, can serve as the foundation for developing data-driven decision support tools. By integrating the proposed framework, these tools could dynamically recommend whether to clean, transform, or discard data in light of context-specific tradeoffs. In complex scenarios, such as streaming analytics, learning to optimize pipeline configurations in real time might support the balancing of timeliness and resource use (e.g., through the prioritization of on-demand cleaning).

Applying the framework in the domain of IoT and edge computing is a critical direction for future work, since Reuse and Reduce strategies in resource-constrained scenarios could significantly lower energy demand and processing latency. A comprehensive case study—e.g., real-time environmental monitoring—would provide a challenging and informative benchmark for the proposed framework. Additionally, applying the framework to optimize dynamic resource allocation in cloud environments could represent another valuable benchmark. The sustainability metrics proposed in the framework can also be the basis to recommend reusable pipeline fragments from a shared pool, in which the actors share metadata about pipelines, maintaining data sovereignty, reducing possible biases and curation efforts needed in data preparation to prevent privacy violations.

References

[1] Haseeb Ali, M. N. Mohd Salleh, Rohmat Saedudin, Kashif Hussain, and Muhammad Faheem Mushtaq. 2019. Imbalance class problems in data mining: A review. *Indonesian Journal of Electrical Engineering and Computer Science* 14, 3 (2019), 1560–1571.

[2] Alhanoof Althnian, Duaa AlSaeed, Heyam Al-Baity, Amani Samha, Alanoud Bin Dris, Najla Alzakari, Afnan Abou Elwafa, and Heba Kurdi. 2021. Impact of dataset size on classification performance: An empirical evaluation in the medical domain. *Applied Sciences* 11, 2 (2021), 796.

[3] Hotham Altwaijry, Dmitri V. Kalashnikov, and Sharad Mehrotra. 2013. Query-driven approach to entity resolution. *Proc. VLDB Endow.* 6, 14 (2013), 1846–1857. DOI: https://doi.org/10.14778/2556549.2556567

[4] Hotham Altwaijry, Sharad Mehrotra, and Dmitri V. Kalashnikov. 2015. QuERy: A Framework for Integrating Entity Resolution with Query Processing. *Proc. VLDB Endow.* 9, 3 (2015), 120–131. DOI: https://doi.org/10.14778/2850583.2850587

[5] Marco Angelini, Giuseppe Santucci, Heidrun Schumann, and Hans-Jörg Schulz. 2018. A review and characterization of progressive visual analytics. *Informatics* 5, 3 (2018), 31. DOI: https://doi.org/10.3390/INFORMATICS5030031

[6] Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *Proceedings of the 12th International Conference on*

Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024. OpenReview.net, 1–30. Retrieved from https://openreview.net/forum?id=hSyW5go0v8. Accessed: 2025-06-10.

[7] Manlio Bacco, Alexander Kocian, Stefano Chessa, Antonino Crivello, and Paolo Barsocchi. 2024. What are data spaces? systematic survey and future outlook. *Data in Brief* 57, 110969 (2024), 1–23. DOI: https://doi.org/10.1016/j.dib.2024.110969

[8] Alexandre Bailly, Corentin Blanc, Élie Francis, Thierry Guillotin, Fadi Jamal, Béchara Wakim, and Pascal Roy. 2022. Effects of dataset size and interactions on the prediction performance of logistic regression and deep learning models. *Comput. Methods Programs Biomed.* 213, 106504 (2022), 1–7. DOI: https://doi.org/10.1016/J.CMPB.2021.106504

[9] Donald Ballou, Richard Wang, Harold Pazer, and Giri Kumar Tayi. 1998. Modeling information manufacturing systems to determine information product quality. *Management Science* 44, 4 (1998), 462–484.

[10] Donald P. Ballou and Giri Kumar Tayi. 1989. Methodology for allocating resources for data quality enhancement. *Commun. ACM* 32, 3 (1989), 320–329.

[11] Carlo Batini, Cinzia Cappiello, Chiara Francalanci, and Andrea Maurino. 2009. Methodologies for data quality assessment and improvement. *ACM Comput. Surv.* 41, 3 (2009), 16:1–16:52.

[12] Carlo Batini and Monica Scannapieco. 2016. *Data and Information Quality—Dimensions, Principles and Techniques*. Springer, Cham. DOI: https://doi.org/10.1007/978-3-319-24106-7

[13] Leilani Battle, Remco Chang, and Michael Stonebraker. 2016. Dynamic prefetching of data tiles for interactive visualization. In *Proceedings of the 2016 International Conference on Management of Data (SIGMOD'16)*. ACM, New York, NY, USA, 1363–1375. DOI: https://doi.org/10.1145/2882903.2882919

[14] Leilani Battle, Philipp Eichmann, Marco Angelini, Tiziana Catarci, Giuseppe Santucci, Yukun Zheng, Carsten Binnig, Jean-Daniel Fekete, and Dominik Moritz. 2020. Database benchmarking for supporting real-time interactive querying of large data. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*. ACM, New York, NY, USA, 1571–1587.

[15] Leilani Battle and Jeffrey Heer. 2019. Characterizing exploratory visual analysis: A literature review and evaluation of analytic provenance in tableau. *Comput. Graph. Forum* 38, 3 (2019), 145–159.

[16] Federico Belotti, Fabio Dadda, Marco Cremaschi, Roberto Avogadro, and Matteo Palmonari. 2024. Evaluating LLMs on Entity Disambiguation in Tables. *arXiv:2408.06423*. Retrieved from https://arxiv.org/abs/2408.06423

[17] Dario Benvenuti, Matteo Filosa, Tiziana Catarci, and Marco Angelini. 2023. Modeling and assessing user interaction in big data visualization systems. In *Proceedings of the IFIP Conference on Human-Computer Interaction*. Springer, Cham, 86–109.

[18] Dario Benvenuti, Matteo Filosa, Tiziana Catarci, and Marco Angelini. 2023. Modeling and assessing user interaction in big data visualization systems. In *Proceedings of the Human-Computer Interaction—INTERACT 2023, 19th IFIP TC13 International Conference, York, UK*. Springer, Cham, 86–109. DOI: https://doi.org/10.1007/978-3-031-42283-6_5

[19] Laure Berti-Équille. 2019. Learn2Clean: Optimizing the sequence of tasks for web data preparation. In *Proceedings of the World Wide Web Conference, WWW 2019*. ACM, New York, NY, 2580–2586. DOI: https://doi.org/10.1145/3308558.3313602

[20] Laure Berti-Equille. 2020. Active reinforcement learning for data preparation: Learn2Clean with human-in-the-loop. In *Proceedings of the CIDR*. CIDR, Amsterdam, The Netherlands, 2. Retrieved from https://www.cidrdb.org/cidr2020/gongshow2020/gongshow/abstracts/cidr2020_abstract59.pdf

[21] Frank Biermann, Norichika Kanie, and Rakhyun E. Kim. 2017. Global governance by goal-setting: the novel approach of the UN Sustainable Development Goals. *Current Opinion in Environmental Sustainability* 26-27 (2017), 26–31.

[22] Carlo A. Bono, Cinzia Cappiello, Barbara Pernici, Edoardo Ramalli, and Monica Vitali. 2023. Pipeline design for data preparation for social media analysis. *ACM Journal of Data and Information Quality* 15, 4 (2023), 1–25.

[23] Kevin Bradley. 2007. Defining digital sustainability. *Library Trends* 56, 1 (2007), 148–163.

[24] John Brooke. 1996. SUS: A quick and dirty usability scale. In *Proceedings of the Usability Evaluation in Industry*. CRC Press, London, 189–194. DOI: https://doi.org/10.1201/9781498710411-35

[25] Rina Diane Caballar. 2024. We need to decarbonize software: The way we write software has unappreciated environmental impacts. *IEEE Spectrum* 61, 4 (2024), 26–31.

[26] Leonardo Candela, Dario Mangione, and Gina Pavone. 2024. The FAIR assessment conundrum: Reflections on tools and metrics. *Data Science Journal* 23, 1 (2024), 21.

[27] Joel Castaño, Silverio Martínez-Fernández, Xavier Franch, and Justus Bogner. 2023. Exploring the carbon footprint of hugging face's ML models: A repository mining study. In *Proceedings of the 2023 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*. IEEE, Piscataway, NJ, 1–12.

[28] Vassilis Christophides, Vasilis Efthymiou, Themis Palpanas, George Papadakis, and Kostas Stefanidis. 2021. An overview of end-to-end entity resolution for big data. *ACM Comput. Surveys* 53, 6, Article 127 (2021), 42 pages. DOI: https://doi.org/10.1145/3418896

[29] Xu Chu, John Morcos, Ihab F. Ilyas, Mourad Ouzzani, Paolo Papotti, Nan Tang, and Yin Ye. 2015. KATARA: A data cleaning system powered by knowledge bases and crowdsourcing. In *Proceedings of the 2015 ACM SIGMOD*. ACM, New York, NY, 1247–1261. DOI: https://doi.org/10.1145/2723372.2749431

[30] World Wide Web Consortium. 2025. W3C. Retrieved from https://www.w3.org/. (2025). Accessed: 2025-06-10.

[31] Wikidata contributors. 2025. Wikidata. Retrieved from https://www.wikidata.org/. (2025). Accessed: 2025-06-10.

[32] Marco Cremaschi, Blerina Spahiu, Matteo Palmonari, and Ernesto Jimenez-Ruiz. 2024. Survey on Semantic Interpretation of Tabular Data: Challenges and Directions. *arXiv:2411.11891*. Retrieved from https://arxiv.org/abs/2411.11891

[33] Vincenzo Cutrona, Flavio De Paoli, Aljaž Košmerlj, Nikolay Nikolov, Matteo Palmonari, Fernando Perales, and Dumitru Roman. 2019. Semantically-enabled optimization of digital marketing campaigns. In *Proceedings of the Semantic Web—ISWC 2019*. Ghidini, C. (Ed.), Springer, Cham, 345–362.

[34] Lisa Ehrlinger and Wolfram Wöß. 2022. A survey of data quality measurement and monitoring tools. *Frontiers Big Data* 5, 850611 (2022), 1–30. DOI: https://doi.org/10.3389/FDATA.2022.850611

[35] Frank Emmert-Streib, Zhen Yang, Han Feng, Shailesh Tripathi, and Matthias Dehmer. 2020. An introductory review of deep learning for prediction models with big data. *Frontiers in Artificial Intelligence* 3:4 (2020), 1-23. DOI: 10.3389/frai.2020.00004

[36] Ju Fan, Jianhong Tu, Guoliang Li, Peng Wang, Xiaoyong Du, Xiaofeng Jia, Song Gao, and Nan Tang. 2024. Unicorn: A unified multi-tasking matching model. *ACM SIGMOD Record* 53, 1 (2024), 44–53.

[37] Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on RAG meeting LLMs: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2024, Barcelona, Spain, August 25-29, 2024*. Ricardo Baeza-Yates and Francesco Bonchi (Eds.), ACM, New York, NY, USA, 6491–6501. DOI: https://doi.org/10.1145/3637528.3671470

[38] Jean-Daniel Fekete, Danyel Fisher, and Michael Sedlmair. 2024. *Progressive Data Analysis*. Eurographics Association, Eindhoven, The Netherlands.

[39] Alvaro A. A. Fernandes, Martin Koehler, Nikolaos Konstantinou, Pavel Pankin, Norman W. Paton, and Rizos Sakellariou. 2023. Data preparation: A technological perspective and review. *SN Comput. Sci.* 4, 4 (2023), 425. DOI: https://doi.org/10.1007/S42979-023-01828-8

[40] Matthias Feurer, Katharina Eggensperger, Stefan Falkner, Marius Lindauer, and Frank Hutter. 2022. Auto-sklearn 2.0: Hands-free AutoML via meta-learning. *J. Mach. Learn. Res.* 23, 261 (2022), 261:1–261:61. Retrieved from http://jmlr.org/papers/v23/21-0992.html

[41] Matteo Filosa, Alexandra Plexousaki, Dario Benvenuti, Tiziana Catarci, and Marco Angelini. 2024. InterView: A system to support interaction-driven visualization systems design. In *Proceedings of the International Conference on Human-Centred Software Engineering*. Springer, Cham, 321–329.

[42] Donatella Firmani, Barna Saha, and Divesh Srivastava. 2016. Online entity resolution using an oracle. *Proc. VLDB Endow.* 9, 5 (2016), 384–395. DOI: https://doi.org/10.14778/2876473.2876474

[43] Donatella Firmani, Letizia Tanca, and Riccardo Torlone. 2020. Ethical dimensions for data quality. *ACM J. Data Inf. Qual.* 12, 1 (2020), 2:1–2:5.

[44] International Organization for Standardization. 2025. ISO. Retrieved from https://www.iso.org/home.html. (2025). Accessed: 2025-06-10.

[45] Daniele Foroni, Matteo Lissandrini, and Yannis Velegrakis. 2021. Estimating the extent of the effects of data quality through observations. In *Proceedings of the ICDE'21*. IEEE, Washington, DC, 1913–1918. DOI: https://doi.org/10.1109/ICDE51399.2021.00176

[46] Gorka Fraga-González, Hester van de Wiel, Francesco Garassino, Willy Kuo, Diane de Zélicourt, Vartan Kurtcuoglu, Leonhard Held, and Eva Furrer. 2025. Affording reusable data: Recommendations for researchers from a data-intensive project. *Scientific Data* 12, 1 (2025), 258.

[47] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv:2312.10997*. Retrieved from https://arxiv.org/abs/2312.10997

[48] Shivani Gupta and Atul Gupta. 2019. Dealing with noise problem in machine learning data-sets: A systematic review. *Procedia Computer Science* 161 (2019), 466–474. DOI: https:doi.org/10.1016/j.procs.2019.11.146

[49] Elizabeth Hetzler and Alan Turner. 2004. Analysis experiences using information visualization. *IEEE Computer Graphics and Applications* 24, 5 (2004), 22–26.

[50] International Organization for Standardization. 2018. *ISO 9241-11:2018 - Ergonomics of human-system interaction—Part 11: Usability: Definitions and concepts*. International Organization for Standardization (ISO), Geneva, Switzerland. Retrieved from https://www.iso.org/standard/63500.html

[51] Ousmane Issa, Angela Bonifati, and Farouk Toumani. 2021. INCA: Inconsistency-aware data profiling and querying. In *Proceedings of the SIGMOD'21*. ACM, New York, NY, 2745–2749. DOI: https://doi.org/10.1145/3448016.3452760

[52] Annika Jacobsen, Ricardo de Miranda Azevedo, Nick Juty, Dominique Batista, Simon Coles, Ronald Cornet, Mélanie Courtot, Mercè Crosas, Michel Dumontier, Chris T. Evelo, et al. 2020. FAIR principles: Interpretations and implementation considerations. *Data Intelligence* 2, 1-2 (2020), 10–29. DOI: https://doi.org/10.1162/dint_r_00024

[53] Annika Jacobsen, Rajaram Kaliyaperumal, Luiz Olavo Bonino da Silva Santos, Barend Mons, Erik Schultes, Marco Roos, and Mark Thompson. 2020. A generic workflow for the data FAIRification process. *Data Intelligence* 2, 1-2 (2020), 56–65.

[54] Sebastian Jäger and Felix Biessmann. 2024. From data imputation to data cleaning - automated cleaning of tabular data improves downstream predictive performance. In *Proceedings of the AISTATS (Proceedings of Machine Learning Research)*. PMLR, proceedings.mlr.press, 3394–3402.

[55] Johannes Jakubik, Michael Vössing, Niklas Kühl, Jannis Walk, and Gerhard Satzger. 2024. Data-centric artificial intelligence. *Business and Information Systems Engineering* 66, 4 (2024), 507–515. DOI: https://doi.org/10.1007/s12599-024-00857-8

[56] Artjom Janssen, Wouter Botzen, Justin Dijk, and Patty Duijm. 2022. Overcoming misleading carbon footprints in the financial sector. *Climate Policy* 22, 6 (2022), 817–822.

[57] Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Houda Bouamor, Juan Pino, and Kalika Bali (Eds.), Association for Computational Linguistics, Singapore, 7969–7992. DOI: https://doi.org/10.18653/v1/2023.emnlp-main.495

[58] Brian R. Keeble. 1988. The brundtland report:'Our common future'. *Medicine and War* 4, 1 (1988), 17–25.

[59] Daniel A. Keim, Florian Mansmann, Jörn Schneidewind, Jim Thomas, and Hartmut Ziegler. 2008. Visual analytics: Scope and challenges. In *Proceedings of the Visual Data Mining: Theory, Techniques and Tools for Visual Analytics*. Simeon J. Simoff, Michael H. Böhlen, and Arturas Mazeika (Eds.), Springer, Berlin, 76–90. DOI: https://doi.org/10.1007/978-3-540-71080-6_6

[60] Gangwoo Kim, Sungdong Kim, Byeongguk Jeon, Joonsuk Park, and Jaewoo Kang. 2023. Tree of clarifications: Answering ambiguous questions with retrieval-augmented large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Houda Bouamor, Juan Pino, and Kalika Bali (Eds.), Association for Computational Linguistics, Singapore, 996–1009. DOI: https://doi.org/10.18653/v1/2023.emnlp-main.63

[61] Julian Kirchherr, Denise Reike, and Marko Hekkert. 2017. Conceptualizing the circular economy: An analysis of 114 definitions. *Resources, Conservation and Recycling* 127 (2017), 221–232. DOI: http://dx.doi.org/10.1016/j.resconrec.2017.09.005

[62] Sanjay Krishnan, Jiannan Wang, Eugene Wu, Michael J. Franklin, and Ken Goldberg. 2016. ActiveClean: Interactive data cleaning for statistical modeling. *Proc. VLDB Endow.* 9, 12 (2016), 948–959. DOI: https://doi.org/10.14778/2994509.2994514

[63] Peng Li, Xi Rao, Jennifer Blase, Yue Zhang, Xu Chu, and Ce Zhang. 2021. CleanML: A study for evaluating the impact of data cleaning on ML classification tasks. In *Proceedings of the ICDE 2021*. IEEE, Washington, DC, 13–24. DOI: https://doi.org/10.1109/ICDE51399.2021.00009

[64] Zhicheng Liu and J. Heer. 2014. The effects of interactive latency on exploratory visual analysis. *IEEE Transactions on Visualization and Computer Graphics* 20, 12 (2014), 2122–2131. DOI: https://doi.org/10.1109/TVCG.2014.2346452

[65] Kadan Lottick, Silvia Susai, Sorelle A. Friedler, and Jonathan P. Wilson. 2019. Energy usage reports: Environmental awareness as part of algorithmic accountability. In *Proceedings of the Workshop on Tackling Climate Change with Machine Learning at NeurIPS 2019*. Vancouver Convention Center, British Columbia, Canada, 1–12.

[66] Andreea Roxana Luca, Tudor Florin Ursuleanu, Liliana Gheorghe, Roxana Grigorovici, Stefan Iancu, Maria Hlusneac, and Alexandru Grigorovici. 2022. Impact of quality, type and volume of data used by deep learning models in the analysis of medical images. *Informatics in Medicine Unlocked* 29, 100911 (2022), 1–7. DOI: https://doi.org/10.1016/j.imu.2022.100911

[67] Alexandra Sasha Luccioni, Sylvain Viguier, and Anne-Laure Ligozat. 2023. Estimating the carbon footprint of BLOOM, a 176B parameter language model. *Journal of Machine Learning Research* 24, 253 (2023), 1–15.

[68] Yuyu Luo, Chengliang Chai, Xuedi Qin, Nan Tang, and Guoliang Li. 2020. Interactive cleaning for progressive visualization through composite questions. In *Proceedings of the ICDE'20*. IEEE, Piscataway, NJ, 733–744. DOI: https://doi.org/10.1109/ICDE48307.2020.00069

[69] Gustavo V. Machado, Ítalo Cunha, Adriano C. M. Pereira, and Leonardo B. Oliveira. 2019. DOD-ETL: Distributed on-demand ETL for near real-time business intelligence. *Journal of Internet Services and Applications* 10, 1 (2019), 21.

[70] Mohammad Mahdavi and Ziawasch Abedjan. 2021. Semi-supervised data cleaning with Raha and Baran. In *Proceedings of the 11th Conference on Innovative Data Systems Research, CIDR 2021*. www.cidrdb.org, online, 7. Retrieved from http://cidrdb.org/cidr2021/papers/cidr2021_paper14.pdf

[71] Mohammad Mahdavi, Felix Neutatz, Larysa Visengeriyeva, and Ziawasch Abedjan. 2019. Towards automated data cleaning workflows. In *Proceedings of the Conference on "Lernen, Wissen, Daten, Analysen" (CEUR Workshop Proceedings)*. CEUR-WS.org, Aachen, Germany, 10–19. Retrieved from https://ceur-ws.org/Vol-2454/paper_8.pdf

[72] Niels Martin, Antonio Martinez-Millana, Bernardo Valdivieso, and Carlos Fernández-Llatas. 2019. Interactive data cleaning for process mining: A case study of an outpatient clinic's appointment system. In *Proceedings of the BPM 2019 International Workshops (LNBIP)*. Springer, Cham, 532–544. DOI: https://doi.org/10.1007/978-3-030-37453-2_43

[73] Toni Mastelic, Ariel Oleksiak, Holger Claussen, Ivona Brandic, Jean-Marc Pierson, and Athanasios V. Vasilakos. 2014. Cloud computing: Survey on energy efficiency. *ACM Comput. Surveys* 47, 2 (2014), 1–36.

[74] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2022. A survey on bias and fairness in machine learning. *ACM Comput. Surv.* 54, 6 (2022), 115:1–115:35.

[75] Leonel Aguilar Melgar, David Dao, Shaoduo Gan, Nezihe Merve Gürel, Nora Hollenstein, Jiawei Jiang, Bojan Karlas, Thomas Lemmin, Tian Li, Yang Li, Susie Xi Rao, Johannes Rausch, Cédric Renggli, Luka Rimanic, Maurice Weber, Shuai Zhang, Zhikuan Zhao, Kevin Schawinski, Wentao Wu, Ce Zhang. 2021. Ease.ML: A lifecycle management system for machine learning. In *Proceedings of the CIDR. www.cidrdb.org, online*, 7. Retrieved from https://www.cidrdb.org/cidr2021/papers/cidr2021_paper26.pdf

[76] Gaurav Menghani. 2023. Efficient deep learning: A survey on making deep learning models smaller, faster, and better. *ACM Comput. Surveys* 55, 12 (2023), 1–37.

[77] Angelo Mozzillo, Luca Zecchini, Luca Gagliardelli, Adeel Aslam, Sonia Bergamaschi, and Giovanni Simonini. 2025. Evaluation of dataframe libraries for data preparation on a single machine. In *Proceedings of the 28th International Conference on Extending Database Technology, EDBT 2025, Barcelona, Spain, March 25-28, 2025*. 337–349. DOI: https://doi.org/10.48786/EDBT.2025.27

[78] Felix Naumann, Johann Christoph Freytag, and Ulf Leser. 2004. Completeness of integrated information sources. *Inf. Syst.* 29, 7 (2004), 583–615.

[79] Felix Neutatz, Binger Chen, Yazan Alkhatib, Jingwen Ye, and Ziawasch Abedjan. 2022. Data cleaning and AutoML: Would an optimizer choose to clean? *Datenbank-Spektrum* 22, 2 (2022), 121–130. DOI: https://doi.org/10.1007/S13222-022-00413-2

[80] Jakob Nielsen. 1996. Usability metrics: Tracking interface improvements. *IEEE Software* 13, 6 (1996), 1–2.

[81] Folorunso Y. Osisanwo, Joseph E.T. Akinsola, Oludele Awodele, John O. Hinmikaiye, Oluwole Olakanmi, and Joseph Akinjobi. 2017. Supervised machine learning algorithms: Classification and comparison. *International Journal of Computer Trends and Technology* 48, 2 (2017), 128–138.

[82] Kim Ouwehand, Avalon van der Kroef, Jacqueline Wong, and Fred Paas. 2021. Measuring cognitive load: Are there more valid alternatives to Likert rating scales?. In *Proceedings of the Frontiers in Education*. Frontiers Media SA, Lausanne, Switzerland, 702616.

[83] Alessandro Palma and Marco Angelini. 2024. It is time to steer: A scalable framework for analysis-driven attack graph generation. In *Proceedings of the Computer Security—ESORICS 2024*. Joaquin Garcia-Alfaro, Rafał Kozik, Michał Choraś, and Sokratis Katsikas (Eds.), Springer Nature Switzerland, Cham, 229–250.

[84] Thorsten Papenbrock, Arvid Heise, and Felix Naumann. 2015. Progressive duplicate detection. *IEEE Transactions on Knowledge and Data Engineering* 27, 5 (2015), 1316–1329. DOI: https://doi.org/10.1109/TKDE.2014.2359666

[85] Hima Patel, Nitin Gupta, Naveen Panwar, Ruhi Sharma Mittal, Sameep Mehta, Shanmukha C. Guttula, Shashank Mujumdar, Shazia Afzal, Srikanta Bedathur, and Vitobha Munigala. 2022. Automatic assessment of quality of your data for AI. In *Proceedings of the CODS-COMAD 2022: 5th Joint International Conference on Data Science and Management of Data (9th ACM IKDD CODS and 27th COMAD), Bangalore, India*. Gargi Dasgupta, Yogesh Simmhan, Balaji Vasan Srinivasan, Sourav Bhowmick, Amith Singhee, Maya Ramanath, Nipun Batra, and Abhinandan S. Prasad (Eds.), ACM, New York, NY, USA, 354–357. DOI: https://doi.org/10.1145/3493700.3493774

[86] Hima Patel, Shanmukha C. Guttula, Nitin Gupta, Sandeep Hans, Ruhi Sharma Mittal, and Lokesh Nagalapatti. 2023. A data-centric AI framework for automating exploratory data analysis and data quality tasks. *ACM J. Data Inf. Qual.* 15, 4 (2023), 44:1–44:26. DOI: https://doi.org/10.1145/3603709

[87] Ralph Peeters, Aaron Steiner, and Christian Bizer. 2025. Entity matching using large language models. *Proceedings of the 28th International Conference on Extending Database Technology (EDBT)*. Retrieved from https://arxiv.org/abs/2310.11244

[88] Javier Perera-Lago, Victor Toscano-Duran, Eduardo Paluzo-Hidalgo, Rocio Gonzalez-Diaz, Miguel A. Gutiérrez-Naranjo, and Matteo Rucco. 2024. An in-depth analysis of data reduction methods for sustainable deep learning. *Open Research Europe* 4, 101 (2024), 101.

[89] José Potting, Marko P. Hekkert, Ernst Worrell, and Aldert Hanemaaije. 2017. Circular economy: Measuring innovation in the product chain. *PBL Netherlands Environmental Assessment Agency* 2544 (2017), 46.

[90] Ben Purvis, Yong Mao, and Darren Robinson. 2019. Three pillars of sustainability: In search of conceptual origins. *Sustainability Science* 14 (2019), 681–695. DOI: https://doi.org/10.1007/s11625-018-0627-5

[91] I. Pérez-Messina, Marco Angelini, Davide Ceneda, Christian Tominski, and Silvia Miksch. 2025. Coupling guidance and progressiveness in visual analytics. *Computer Graphics Forum* 44, 3 (2025), 1–12. DOI: https://doi.org/10.1111/cgf.70115

[92] Zhixin Qi and Hongzhi Wang. 2021. Dirty-data impacts on regression models: An experimental evaluation. In *Proceedings of the DASFAA 2021 (LNCS)*. Springer, Cham, 88–95. DOI: https://doi.org/10.1007/978-3-030-73194-6_6

[93] Zhi-Xin Qi, Hong-Zhi Wang, and An-Jie Wang. 2021. Impacts of dirty data on classification and clustering models: An experimental evaluation. *J. Comput. Sci. Technol.* 36, 4 (2021), 806–821. DOI: https://doi.org/10.1007/S11390-021-1344-6

[94] Theodoros Rekatsinas, Xu Chu, Ihab F. Ilyas, and Christopher Ré. 2017. HoloClean: Holistic data repairs with probabilistic inference. *Proceedings of the VLDB Endowment* 10, 11 (2017), 1190–1201. DOI: https://doi.org/10.14778/3137628.3137631

[95] Kevin Rennert, Frank Errickson, Brian C. Prest, Lisa Rennels, Richard G. Newell, William Pizer, Cora Kingdon, Jordan Wingenroth, Roger Cooke, Bryan Parthum, et al. 2022. Comprehensive evidence implies a higher social cost of CO_2. *Nature* 610, 7933 (2022), 687–692.

[96] Camilla Sancricca and Cinzia Cappiello. 2025. Lightweight Pipelines: Good Enough is Sometimes Better. In *VLDB 2025 Workshop: 2nd International Workshop on Data-Centric AI (DATAI)*.

[97] Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. 2020. Green AI. *Commun. ACM* 63, 12 (2020), 54–63.

[98] Oleksandr Shchur, Ali Caner Türkmen, Nick Erickson, Huibin Shen, Alexander Shirkov, Tony Hu, and Bernie Wang. 2023. AutoGluon-timeseries: AutoML for probabilistic time series forecasting. In *Proceedings of the International Conference on Automated Machine Learning*. Proceedings of Machine Learning Research, Potsdam, Germany, 9/1–21. Retrieved from https://arxiv.org/pdf/2308.05566

[99] Donghee Shin and Emily Y. Shin. 2023. Human-centered AI: A framework for green and sustainable AI. *Computer* 56, 6 (2023), 16–25.

[100] Ben Shneiderman. 2003. The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings of the Craft of Information Visualization*. Elsevier, 364–371.

[101] Shrey Shrivastava et al. 2019. DQA: Scalable, automated and interactive data quality advisor. In *Proceedings of the 2019 (IEEE BigData)*. IEEE, Piscataway, NJ, 2913–2922. DOI: https://doi.org/10.1109/BIGDATA47090.2019.9006187

[102] Giovanni Simonini, George Papadakis, Themis Palpanas, and Sonia Bergamaschi. 2018. Schema-agnostic progressive entity resolution. In *Proceedings of the IEEE International Conference on Data Engineering (ICDE)*. IEEE, Washington, DC, 53–64. DOI: https://doi.org/10.1109/ICDE.2018.00015

[103] Giovanni Simonini, Luca Zecchini, Sonia Bergamaschi, and Felix Naumann. 2022. Entity resolution on-demand. *Proc. VLDB Endow.* 15, 7 (2022), 1506–1518. DOI: https://doi.org/10.14778/3523210.3523226

[104] Aditi Singh, Abul Ehtesham, Saket Kumar, and Tala Talaei Khoei. 2025. Agentic retrieval-augmented generation: A survey on agentic RAG. *arXiv:2501.09136*. Retrieved from https://arxiv.org/abs/2501.09136

[105] Dalwinder Singh and Birmohan Singh. 2020. Investigating the impact of data normalization on classification performance. *Appl. Soft Comput.* 97, Part B (2020), 105524. DOI: https://doi.org/10.1016/J.ASOC.2019.105524

[106] Matthias Stuermer, Gabriel Abu-Tayeh, and Thomas Myrach. 2017. Digital sustainability: Basic conditions for sustainable digital artifacts and their ecosystems. *Sustainability Science* 12 (2017), 247–262. DOI: https://doi.org/10.1007/s11625-016-0412-2

[107] John Sweller. 2017. Measuring cognitive load. *Perspectives on Medical Education* 7, 1 (2017), 1–2. DOI: https://doi.org/10.1007/s40037-017-0395-4

[108] J. Joshua Thomas and Kristin A. Cook. 2006. A visual analytics agenda. *IEEE Computer Graphics and Applications* 26, 1 (2006), 10–13. DOI: https://doi.org/10.1109/MCG.2006.5

[109] Alex Ulmer, Marco Angelini, Jean-Daniel Fekete, Jörn Kohlhammer, and Thorsten May. 2024. A survey on progressive visualization. *IEEE Transactions on Visualization and Computer Graphics* 30, 9 (2024), 6447–6467. DOI: https://doi.org/10.1109/TVCG.2023.3346641

[110] Aimee Van Wynsberghe. 2021. Sustainable AI: AI for sustainability and the sustainability of AI. *AI and Ethics* 1, 3 (2021), 213–218.

[111] Roberto Verdecchia, Luís Cruz, June Sallou, Michelle Lin, James Wickenden, and Estelle Hotellier. 2022. Data-centric green AI an exploratory empirical study. In *Proceedings of the International Conference on ICT for Sustainability (ICT4S)*. IEEE, Piscataway, NJ, 35–45.

[112] Roberto Verdecchia, Patricia Lago, Christof Ebert, and Carol De Vries. 2021. Green IT and green software. *IEEE Software* 38, 6 (2021), 7–15.

[113] Ricardo Vinuesa, Hossein Azizpour, Iolanda Leite, Madeline Balaam, Virginia Dignum, Sami Domisch, Anna Felländer, Simone Daniela Langhans, Max Tegmark, and Francesco Fuso Nerini. 2020. The role of artificial intelligence in achieving the Sustainable Development Goals. *Nature Communications* 11, 1 (2020), 1–10.

[114] Pattaramon Vuttipittayamongkol, Eyad Elyan, and Andrei Petrovski. 2021. On the class overlap problem in imbalanced data classification. *Knowl. Based Syst.* 212 (2021), 106631. DOI: https://doi.org/10.1016/j.knosys.2020.106631

[115] Usman Wajid, Cinzia Cappiello, Pierluigi Plebani, Barbara Pernici, Nikolay Mehandjiev, Monica Vitali, Michael Gienger, Kostas Kavoussanakis, David Margery, David Garcia Perez, et al. 2015. On achieving energy efficiency and reducing CO_2 footprint in cloud computing. *IEEE Transactions on Cloud Computing* 4, 2 (2015), 138–151.

[116] Anna M. Walker, Katelin Opferkuch, Erik Roos Lindgreen, Andrea Raggi, Alberto Simboli, Walter J. V. Vermeulen, Sandra Caeiro, and Roberta Salomone. 2022. What is the relation between circular economy and sustainability?

Answers from frontrunner companies engaged with circular economy practices. *Circular Economy and Sustainability* 2, 2 (2022), 731–758.

[117] Gerd Waloszek and Ulrich Kreichgauer. 2009. User-centered evaluation of the responsiveness of applications. In *Proceedings of the IFIP Conference on Human-Computer Interaction*. Springer, Cham, 239–242.

[118] Richard Y. Wang and Diane M. Strong. 1996. Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems* 12, 4 (1996), 5–33.

[119] Steven Euijong Whang, David Marmaros, and Hector Garcia-Molina. 2013. Pay-as-you-go entity resolution. *IEEE Transactions on Knowledge and Data Engineering* 25, 5 (2013), 1111–1124. DOI: https://doi.org/10.1109/TKDE.2012.43

[120] Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, et al. 2016. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data* 3, 1 (2016), 1–9.

[121] Mengwei Xu, Dongqi Cai, Wangsong Yin, Shangguang Wang, Xin Jin, and Xuanzhe Liu. 2025. Resource-efficient algorithms and systems of foundation models: A survey. *ACM Comput. Surveys* 57, 5 (2025), 1–39.

[122] Cong Yan and Yeye He. 2020. Auto-suggest: Learning-to-recommend data preparation steps using data science notebooks. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data (SIGMOD'20)*. ACM, New York, NY, USA, 1539–1554. DOI: https://doi.org/10.1145/3318464.3389738

[123] Junwen Yang, Yeye He, and Surajit Chaudhuri. 2021. Auto-pipeline: Synthesizing complex data pipelines by-target using reinforcement learning and search. *Proceedings of the VLDB Endowment* 14, 11 (2021), 2563–2575.

[124] Fernando Rezende Zagatti, Lucas Cardoso Silva, Lucas Nildaimon dos Santos Silva, Bruno Silva Sette, Helena de Medeiros Caseli, Daniel Lucrédio, and Diego Furtado Silva. 2021. MetaPrep: Data preparation pipelines recommendation via meta-learning. In *Proceedings of the ICMLA*. IEEE, Piscataway, NJ, 1197–1202.

[125] Luca Zecchini, Tobias Bleifuß, Giovanni Simonini, Sonia Bergamaschi, and Felix Naumann. 2024. Determining the largest overlap between tables. *Proc. ACM Manag. Data* 2, 1 (2024), 48:1–48:26. DOI: https://doi.org/10.1145/3639303

[126] Luca Zecchini, Giovanni Simonini, Sonia Bergamaschi, and Felix Naumann. 2023. BrewER: Entity resolution on-demand. *Proc. VLDB Endow.* 16, 12 (2023), 4026–4029. DOI: https://doi.org/10.14778/3611540.3611612

[127] Tianshu Zhang, Xiang Yue, Yifei Li, and Huan Sun. 2024. TableLlama: Towards open large generalist models for tables. In *Proceedings of the NAACL (Volume 1: Long Papers)*. Association for Computational Linguistics, Kerrville, TX, 6024–6044.

[128] Yanjie Zhao, Li Li, Haoyu Wang, Haipeng Cai, Tegawendé F. Bissyandé, Jacques Klein, and John C. Grundy. 2021. On the impact of sample duplication in machine-learning-based android malware detection. *ACM Trans. Softw. Eng. Methodol.* 30, 3 (2021), 40:1–40:38. DOI: https://doi.org/10.1145/3446905

Appendix

A Cost Table

The appendix contains the general costs used in the evaluations exemplified in Section 5. The tables illustrate the evaluation parameters considered for the case study. Costs are derived from the literature or previous research work, and the assumptions made in this context are mentioned. These costs might have different values in different contexts and are considered realistic costs in the context of the case study, enabling the comparison of sustainability aspects across alternative techniques.

A.1 Service Costs and Error Rates

The following services are used, indicating estimated costs and their expected accuracy errors.

Parameter	Value	Notes
OpenWeather service cost	0.0000016	Euros per item (Developer subscription)
HERE service cost	0.0003	Euros per item (Enterprise subscription)
LLM API cost (GPT3.5-turbo)	0.50/1M tokens	E.g., comparing a pair of records can be approximated to 1,000 tokens, which means 0.0005 euro/pair)
LLM API cost (GPT4o)	$2.50(in)-$10(out)/1M tokens	Comparing a pair of records (93 input tokens, 1 output token [87]) can be approximated to $0.0002425/pair)
LLM API cost (GPT4o-mini)	$0.15(in)-$0.6(out)/1M tokens	Comparing a pair of records (93 input tokens, 1 output token [87]) can be approximated to $0.00001455/pair)
Wikidata service	0.00	Open source

The cost estimates may change depending on scale, service configuration, market conditions, and technological advancements.

A.2 Computational Costs

We assume to have two types of machines, available as a service, a general purpose machine (c8g.16xlarge/64 cores/128 GiB) and a memory optimized machine (x2gd.16xlarge/64 cores/1,024 GiB). Costs are derived from available AWS services.[13]

Parameter	Value	Notes
General purpose machine	2.729	Euros per hour
Memory optimized machine	6.400	Euros per hour

A.3 Sustainability Costs

Sustainability costs have a large variability in terms of their context of application, since they depend on location, time of the day, energy mix, external conditions, and so on. We consider the following estimated costs in the computations given in the article:

[13] https://aws.amazon.com/ec2/pricing/, visited on 10 June 2025

Parameter	Value	Notes
HILT	40	Euros per person-hour
CO_2 footprint	160–487	Yearly Kg, per server (see https://www.goclimate.com/blog/the-carbon-footprint-of-servers/)
CO_2 equivalent	38–360	Euros per tonne [95]
CO_2 footprint	0.003–0.010	Euros per hour (min - max)
Error correction	0.50	Euros per item—as in [9] we assume constant costs, due to their variability. In the framework of this article, a standard correction cost can be applied as we assume to compare methods based on human and computational times, applying the same correction techniques
Error assessment	0.01	euros per item—as for correction costs, we assume a standard assessment cost

Received 14 February 2025; revised 19 June 2025; accepted 12 August 2025