DOI: 10.1111/jiec.70010

JOURNAL OF INDUSTRIAL ECOLOGY WILEY

RESEARCH ARTICLE

# Shedding light on the shadows

## Transparency challenge in background life cycle inventory data

Jing Guo[1,2] | Ruiqiao Li[2] | Ruirui Zhang[2] | Jianchuan Qi[2] | Nan Li[2] | Changqing Xu[3] | Anthony S. F. Chiu[4] | Yutao Wang[5] | Hiroki Tanikawa[6] | Ming Xu[2]

[1] School of Management Science and Engineering, Beijing Information Science & Technology University, Beijing, China

[2] School of Environment, Tsinghua University, Beijing, China

[3] School of Economics, Beijing Institute of Technology, Beijing, China

[4] Center for Engineering and Sustainable Development Research (CESDR), De La Salle University, Manila, Philippines

[5] Department of Environmental Science and Engineering, Fudan University, Shanghai, China

[6] Graduate School of Environmental Studies, Nagoya University, Nagoya, Japan

**Correspondence**
Ming Xu, School of Environment, Tsinghua University, Beijing 100084, China. Email: xu-ming@tsinghua.edu.cn

Editor Managing Review: Alexis Laurent

**Funding information**
National Key Research and Development Program of China, Grant/Award Number: 2023YFE0122400; National Natural Science Foundation of China, Grant/Award Numbers: 52293445, 52300234

## Abstract

Life cycle assessment (LCA) hinges on the transparency and reliability of inventory data. However, the transparency of background life cycle inventory (LCI) data sources remains unexamined. This research assesses data transparency in mainstream LCI databases using a two-step examination system based on source findability and accessibility. Six major databases (ecoinvent, GaBi, U.S. Life Cycle Inventory Database, European Life Cycle Database, Inventory Database for Environmental Analysis, and Chinese Life Cycle Database) were analyzed by sampling processes and tracing their sources. The results reveal widespread transparency issues, with only 40%–60% of sampled processes having findable sources and <5% being fully accessible in certain databases. Incomplete documentation and complex cross-referencing between processes and sources posed key barriers. The lack of transparency undermines LCA credibility and necessitates reconstructing databases for enhanced traceability. Although a preliminary study, these findings highlight the challenge of data transparency and provide a methodology to evaluate databases. This drives collective action to uphold transparency standards, restoring trust in LCA as a sustainability decision-making tool.

KEYWORDS
background database, data accessibility, data findability, data transparency, life cycle assessment, life cycle inventory

## 1 | INTRODUCTION

Life cycle assessment (LCA) is pivotal in environmental decision-making and policy formulation, offering a comprehensive framework to assess the environmental impacts of products and services throughout their life cycle (Finnveden et al., 2009). While LCA is a powerful tool for holistic decision-making encompassing factors from ecological consequences to economic implications (Hellweg & Milà i Canals, 2014; Reale et al., 2017), its effectiveness hinges critically on the transparency and reliability of its data (Guinée et al., 2011).

Life cycle inventory (LCI) data is the cornerstone of LCA. LCI data, forming the backbone of LCA studies, includes both foreground data obtained from the processes directly associated with the target product, and background data providing industry-representative information on upstream and downstream processes (Ciroth & Burhan, 2021). The majority of unit processes in a product system rely on this background data (Kalverkamp et al., 2020; Wernet et al., 2016) typically sourced from dedicated LCI databases (e.g., ecoinvent and GaBi). Therefore, these background LCI databases considerably determine the outcomes of LCA results.

ISO 14044:2006 underscores the need for clear disclosure in LCI data to enhance the credibility and reproducibility of LCA results. This emphasis has sparked a broader discussion about data transparency in general. Data transparency is broadly defined as "*the ability of subjects to effectively gain access to all information related to data used in the processes and decisions that affect them*" (Bertino et al., 2019). The lack of transparency not only undermines the trustworthiness of the LCA results but also hampers their utility (Astudillo et al., 2017; Guinée et al., 2011). Transparency in these data therefore becomes a growing concern as highlighted in recent studies (Hertwich et al., 2018; Joyce & Björklund, 2022; Pauliuk et al., 2015; Wu & Wang, 2022). This dialogue encompasses various aspects (Kuczenski, 2019; Saade et al., 2019). Particularly, data source disclosure is fundamental to LCI data transparency. Although the selection of databases is often disclosed in LCA studies, the sources of data in those databases have not yet been fully examined.

This study assesses the data transparency in mainstream LCI databases (ecoinvent, GaBi, USLCI, ELCD, IDEA, and CLCD), aiming to examine how transparent the data in mainstream LCI databases are and what impedes data transparency. Our results show only 40%–60% of sampled processes having fully findable sources in the six databases examined, and three databases have significant issues in source accessibility. The findings highlight the gap of data transparency in LCI databases, calling for a transformative effort to empower stakeholders across sectors to make LCI data more transparent.

## 2 | METHODOLOGY

### 2.1 | Process–source cross-citation in LCI databases

In LCI databases, a single process often has multiple sources, while a particular source can be referenced by many processes. For example, over one third of the processes in the European reference life cycle database (ELCD) use more than 50 sources each. Meanwhile, nearly a quarter of the ELCD sources have been cited for over 50 times. This intricate cross-referencing between processes and sources poses challenges for data transparency because the specific data point (i.e., flows) within a unit process cannot be traced back to their original sources, as sources are typically only indicated at the unit process level. Furthermore, when a single source, such as a book or academic paper, is cited by multiple processes, it is unclear which specific part of that source is being referenced, making it difficult to assess data origins and credibility when sources are shared across processes.

Cross-citations in LCI databases reflect the complex relationships between processes and their data sources. To quantify the complexity of process–source cross-citations, we use the measure of bipartite density. The cross-citation network forms a bipartite graph, with processes and sources as distinct node sets and citations as edges. Bipartite density is the ratio of actual edges to possible edges, which enables direct comparisons of networks of different sizes to reflect the complexity of cross-citations between processes and sources. The formula for measuring the complexity is:

$$\text{Complexity of cross citations} = \frac{E}{n \times m} \times 100 \tag{1}$$

where $E$ is the number of edges in the graph, $n$ is the number of nodes in one partition of the bipartite graph (e.g., "process"), and $m$ is the number of nodes in the other partition (e.g., "source"). A higher value indicates a more complex network of citations, where processes tend to cite multiple sources and sources tend to be cited by multiple processes.

### 2.2 | Hierarchical data transparency metrics

The FAIR principle (findable, accessible, interoperable, and reusable) is a widely accepted guideline for the provision and management of open data (Hertwich et al., 2018). Particularly, the FAIR principle emphasizes machine-actionability, that is, the capacity of computational systems to find, access, interoperate, and reuse data with none or minimal human intervention. However, transparency in LCI databases aims to assist users to assess data quality and suitability and choose reliable data for their studies. In this study, we follow the F and A components of the FAIR principle to propose a two-step hierarchical criterion to evaluate the transparency of LCI databases in terms of where the data is collected and whether the source data is accessible:

**TABLE 1** Required information for findability of various source types.

| Source type | Required info for findability |
|---|---|
| Article in periodical | 1) Article title and author(s), or<br>2) DOI number |
| Chapter in anthology | Chapter title, author(s), book title, year of publication or edition version |
| Monograph | Title, author(s), year of publication or edition version |
| Industry report | Title, authoring organization, year of publication |
| Standard/directive | 1) Standard number, or<br>2) Title, issuing organization, publication time |
| Environmental Product Declaration (EPD) | 1) Declaration number, or<br>2) Product name, declaration holder, publisher, issue time |
| Patent | 1) Patent number, or<br>2) Title, author(s), time |
| Statistical documents | Document title, authoring organization, year of publication |
| Software or database | Name, version or release year |
| Personal communication | Correspondent's name, date of communication |
| Direct measurement | Date of measurement, location of measurement, entity |
| Website | URL and retrieved date |

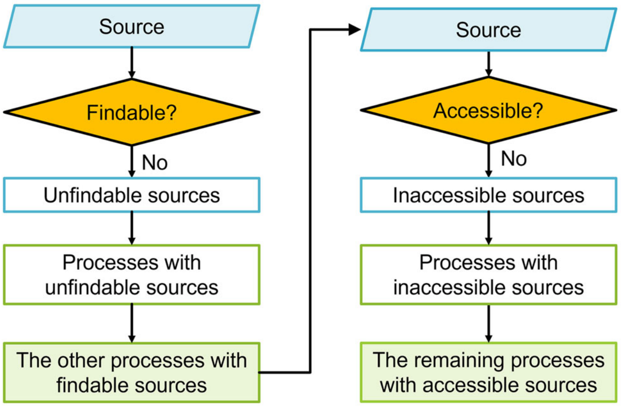
Source
Findable?
No
Unfindable sources
Processes with unfindable sources
The other processes with findable sources
Source
Accessible?
No
Inaccessible sources
Processes with inaccessible sources
The remaining processes with accessible sources

**FIGURE 1** Framework to examine the data transparency for life cycle inventory databases.

1. Findable: sources are documented with complete citation information including organization or author name, title, year, and so on, ensuring they are findable.
2. Accessible: sources are accessible by their reference information.

Based on this criterion, we thoroughly examine the sources referenced by the processes within a database to assess the data transparency of each process by the following steps:

1. **Process sampling**: We first randomly sample a set of processes from the LCI database of investigation. The source information provided in the database for these sampled processes is also obtained.
2. **Findability evaluation of sources**: In this step, we evaluate whether each source cited by the sampled processes has all the necessary information. LCI databases include various types of sources each of which has its own set of required information (Table 1). Once sources with incomplete information are identified, we proceed to scrutinize the processes that cite these sources. Any process that depends on such incomplete sources is considered not transparent and is consequently excluded from further analysis. The remaining processes, along with their sources that are complete and verifiable, are carried forward to the next step of our investigation (Figures 1 and 2b).
3. **Examination of source accessibility**: The next step examines the accessibility of sources that have complete information. We begin by organizing these sources in the order of their citation frequency, starting with the most frequently cited. Each source is then rigorously checked for its

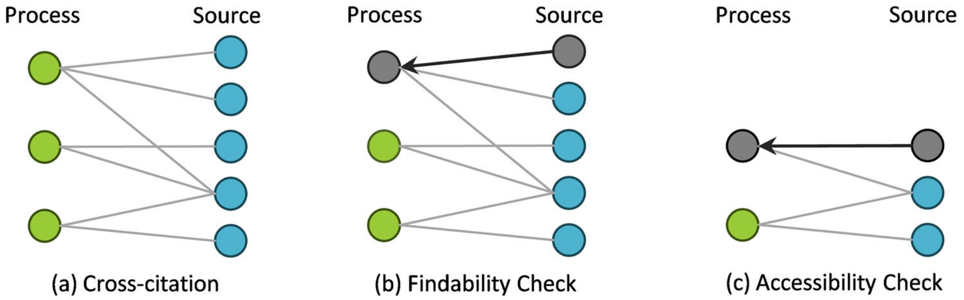

**FIGURE 2** Workflow of process transparency examination (green color indicates the sampled processes, blue indicates the sources cited by these processes, and gray indicates the sources that cannot be found or accessed and the resulting not transparent processes).

availability individually. If a source cannot be found, all processes citing that source are considered to be not transparent, and their sources are excluded from further accessibility checking as demonstrated in Figure 2c. This procedure then continues with the next most-cited source that requires examination, and so on. This iterative approach continues until every source that needs to be checked for accessibility is examined.

This approach allows us to systematically evaluate the transparency of individual processes within the LCI database. It is important to note that the implications of data transparency extend beyond individual processes. In addition, when a process uses a source that lacks transparency (i.e., is not findable or accessible), the transparency of that process itself becomes questionable. This issue can indeed propagate when such a process is used within other processes, creating a chain of reduced transparency.

### 2.3 | Sampling and data processing

To obtain a reasonable and appropriate sample pool, a sampling procedure is performed. Sample size for database is determined by Equation (1),

$$\text{Sample size} = \frac{NZ^2 p\,(1-p)}{e^2\,(N-1) + Z^2 p\,(1-p)} \tag{2}$$

where $N$ is the population size, $Z$ is the level of confidence (95% leading to an Z-score of 1.96), $p$ is the expected proportion (0.5 is used as it is the most conservative assumption and applicable when previous knowledge about the population is lacking), and $e$ is the margin of error (0.05, based on a confidence level of 95%) (Saavedra-Rubio et al., 2022).

Six popular databases are selected to investigate the data transparency, including ecoinvent, GaBi, ELCD, USLCI, CLCD, and IDEA. Among these, ELCD and USLCI are open source and can be accessed at https://eplca.jrc.ec.europa.eu/ELCD3/ and https://www.nrel.gov/analysis/lci.html, respectively. The other databases require a purchase or subscription for access. Process samples are randomly extracted according to the sample size of each database. In particular, ecoinvent adheres to the EcoSpold format, with source information including First Author, Additional Author(s), Title, Year, Journal, Volume Number, Issue Number, and so on. It is worth noting that in the ecoinvent database, approximately 20% of the processes with sources contain identical information but describe different countries or markets, often originating from Swiss datasets. While this could potentially be seen as repetitive, these entries were included in our analysis to maintain the integrity of our random sampling approach and to accurately represent the database as it is used by LCA practitioners. In contrast, GaBi, ELCD, and USLCI follow the ILCD format with source information including "Short name," "Source citation," "Publication type," "Source description or comment," "Reference to contact," and "Reference to digital file." The CLCD database organizes its data sources into two key categories: references and data sources, which provide almost the same information for assessing transparency. The IDEA database provides name, general comment, and information source for each process. The "general comment" field is the most useful for investigating transparency, as it contains detailed sources of the processes. However, those sources are described extensively in the text and are not structured. These data cover a wide range of information including succinct identifiers, source references, document classifications, descriptions, contacts for further information, and links to digital files, which are used for the transparency investigation.

The multifaceted nature of those source information presents a unique challenge for efficient source review because they are often in multiple languages and in unstructured formats. For example, the reference information in sources is often disorganized, with details such as title, author, and institution appearing in different orders within the same database for different sources, making manual review extremely difficult. This study leverages Large Language Model (LLM), specifically GPT 4 by OpenAI, to systematically structure source information from "Source citation" for ILCD formatted database at scale, transforming unstructured information into structured tables, thus enabling a more streamlined, efficient, and easier approach to data transparency examination.

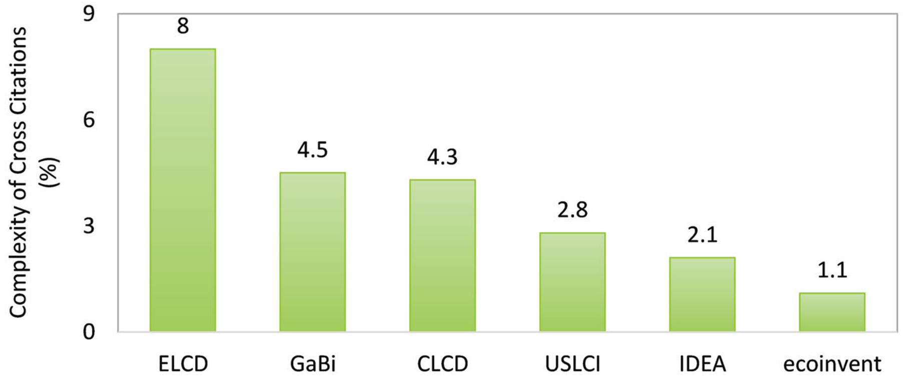

**FIGURE 3** Complexity of cross-citation between process and source of databases. Underlying data for this figure are available in Table S1 of Supporting Information S1.

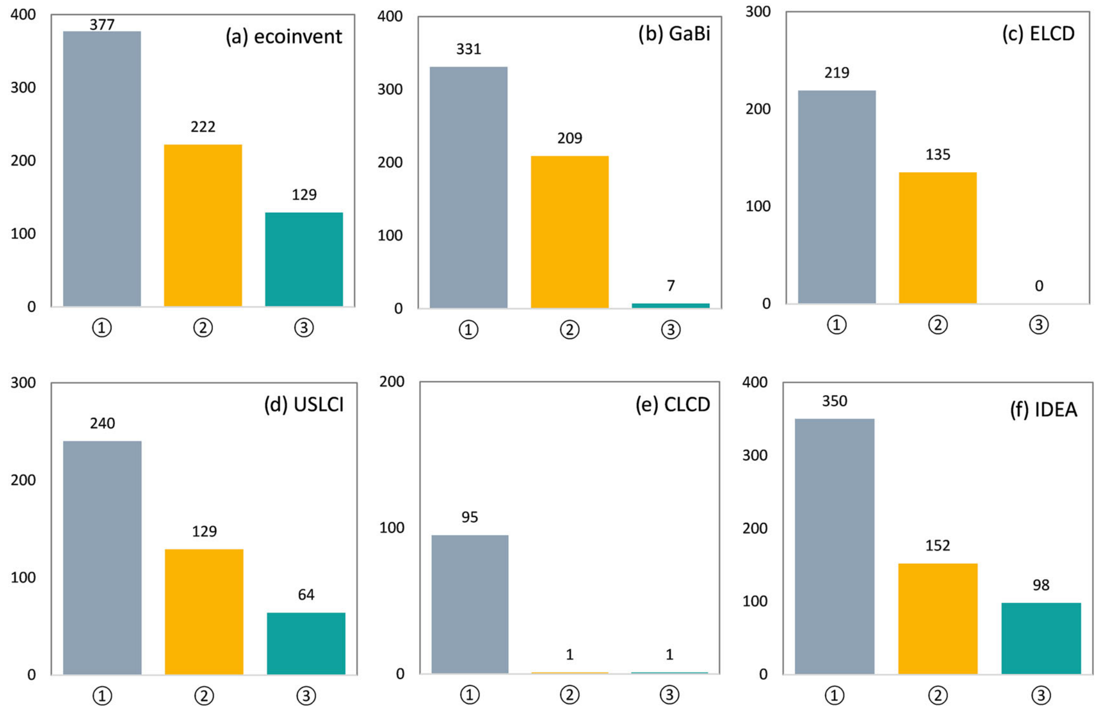

**FIGURE 4** Process transparency. Bar ① illustrates the total number of sampled processes, bar ② depicts the number of processes with findable sources, and bar ③ shows the number of processes with sources that are fully accessible. Underlying data for this figure are available in Table S2 of Supporting Information S1.

To ensure the accuracy and reproducibility of the LLM-structured outputs, we conduct a manual validation process. A representative subset of the structured tables generated by the LLM is reviewed, comparing the outputs to the original source information to verify correctness and consistency. The results of this validation confirmed the high stability and reliability of the structured outputs; repeated runs with identical prompts and datasets consistently produced minimal variability in the results. All sampled data and prompts to query the LLM are available in the Supporting Information.

## 3 | RESULTS

Figure 3 illustrates the complexity of cross-citation between processes and sources, while Figure 4 shows the results of process transparency. Detailed analysis for each database is provided below.

## 3.1 | Ecoinvent

We examined the random sample of 377 processes and found that over 40% do not provide any source information. The other 222 processes refer to a total of 156 unique sources. Approximately 20% of processes were identical except for regional differences. Compared with other LCI databases, the complexity of cross-citations in the processes examined is remarkably low, at just 1.1% (Figure 3). This indicates a low level of interlinking between sources across various processes. On average, each process referenced only 1.7 sources, underscoring a relatively simple sourcing structure.

Most of these sources are reports, journal articles, and books, offering a structured form of information. Importantly, they are relatively complete, consistently providing key details such as title, author, and year, indicating high findability.

Upon a detailed accessibility review, it is found that 129 processes (representing 34.2% of the sampled processes) are fully source accessible. These sources, totaling 68, included not only publicly available literature but also ecoinvent LCI reports, showcasing ecoinvent's commendable transparency. One notable example is a source entitled "Life Cycle Inventories of Chemicals," a 957-page report that detailed inventories for approximately 100 chemicals, along with information on production technology, data sources, and data quality considerations. However, there are issues with the accessibility of some sources. Some could not be found online at all, while others are located but their files are not accessible.

## 3.2 | GaBi

Among the 331 process samples analyzed, only 9 of them lacks source documentation. Ninety-seven percent processes are cross-cited with a total of 797 distinctive sources, including primarily reports, datasets, and Environmental Product Declarations (EPDs), and so on. The complexity of cross-citation is 4.63%. Each process involves an extensive array of sources, averaging to about 37 different sources per process. A significant factor contributing to this complexity is that 85% of the processes within the database are LCI results derived from multiple sources.

Upon closer inspection of the details for each source, there are only 49 sources (around 6%) that we are not able to find, mainly due to the absence of necessary information such as year, author, or even title. These unfindable sources relate to 113 processes, making those processes prone to transparency issues. As a result, those processes are excluded from the accessibility check, leaving 209 processes and their corresponding 289 sources.

Through thorough review, 46 processes are identified with sources considered fully accessible. However, 39 of those processes exclusively utilize the "GaBi databases" as their source of primary data and are thus excluded from our study. This leaves only 7 processes (2%) with 10 distinct sources. The challenges in accessing those sources are multifaceted and not unique to Gabi. Some sources are retrievable but only lead to general or homepage links rather than the specific data source. Access restrictions, including login requirements, are common across multiple databases and reflect a broader challenge in balancing data protection with transparency. Book sources, while not inherently inaccessible, present challenges for immediate verification in digital workflows. In some instances, sources remain untraceable. These findings suggest that a significant portion of processes across various databases, including GaBi, face substantial accessibility issues even when sources are findable.

## 3.3 | ELCD

A sample of 219 processes are analyzed for the ELCD database, involving a total of 453 sources. It exhibits the highest level of cross-citation complexity, estimated at 8.02%. Similar to GaBi, the majority of processes (over 80%) are LCI outcomes. On average, each process encompasses 36 sources, with each source being cited approximately 17 times on average. This complexity undoubtedly poses challenges to the transparency of sources for each process.

Upon conducting a findability check, it is discovered that 21 sources are unfindable, which affect 84 processes. For the remaining 135 processes with findable sources, they are associated with 208 distinct sources. However, it is regrettable to note that during the assessment of accessibility, not a single process had sources that are completely accessible. The primary reason for this is that, among the 208 sources, the most frequently cited one is the GaBi database, which is referenced by 116 processes. Given that the GaBi database is behind paywall, users are unable to directly verify the accuracy of the data. Furthermore, as previously discussed, the process transparency of the GaBi database is not entirely satisfactory. This lack of transparency, coupled with the restricted accessibility due to the paywall, means that the sources for the 116 processes cannot be deemed fully accessible.

## 3.4 | USLCI

We sample 240 processes from the USLCI database, which encompass a total of 253 distinct sources. The USLCI database showcases a lower cross-citation complexity at 2.82%. Typically, each process is linked to around seven sources. This simplicity in the sourcing of processes stems from the

**TABLE 2** Sampling size for each database examined.

| | ecoinvent | GaBi | USLCI | ELCD | IDEA | CLCD |
|---|---|---|---|---|---|---|
| Version | v3.8 | Edition 2022 | Fall 2022 Release | v3.2 | v2.3 | CLCD-China 0.9 |
| Process count | 19,565 | 15,936 | 628 | 509 | 3876 | 404 |
| Process sample size | 377 | 331[a] | 240 | 219 | 350 | 95[b] |
| Source count | 156 | 797 | 253 | 454 | 180 | 146 |

[a] Based on Equation (1), 375 processes were randomly sampled, 331 of them are owned by Sphera Solutions GmbH, and others are originated from other databases, such as US LCI and World steel, which were eliminated from the sample pool.

[b] The CLCD database encompasses additional datasets, such as CLCD-China-ECER 0.8. Yet, despite having purchased the CLCD-China 0.9 database which contains a total of 404 processes, only 95 processes are accessible for specific details.

fact that the vast majority of processes drawn from the USLCI database are identified as unit processes. This classification tends to increase their transparency, thereby rendering the information more direct and user friendly.

Upon conducting a thorough examination of source findability, it is found that 20 sources lack the necessary information to be deemed findable. This lack of findability has significant implications, as it affects a total of 111 processes.

The remaining 129 processes are associated with 121 distinct sources. It is noteworthy that 67 of these processes have sources that are fully accessible, encompassing a total of 44 unique sources. The accessible sources are diverse, encompassing reputable governmental sources such as the US Energy Information Administration (EIA) and the US Environmental Protection Agency (EPA), as well as academic publications. Industry reports and studies also play a significant role, with contributions from Franklin Associates on various LCAs and inventories for plastics and resins, and the American Chemistry Council on chemicals and plastics.

## 3.5 | CLCD

We look at 95 processes in the CLCD database due to limited access (Table 2), among which 25 processes are not compiled with any references. The remaining 70 processes are linked to 146 different sources and show a cross-reference complexity of about 6.80%. Out of these sources, 101 only provide title information without essential information, which compromise their findability and affect the transparency of 69 processes. This leaves only one process (melamine production) with findable sources. Although the source associated with this process is accessible, significant transparency issues exist across nearly all sampled processes in the CLCD database.

## 3.6 | IDEA

A total of 350 processes are sampled from the IDEA database. One hundred and eighty sources are extracted from the general comment sections of 287 processes, while 18% of the sampled processes are without any source information. The cross-reference complexity is 2.63%. Additionally, each process involved in the database typically references an average of four different sources.

Among the 181 sources identified, 61 are unfindable, which impact 135 processes. The rest of the 152 findable processes have 64 sources, in which 50 are accessible. These sources constitute 98 processes (28%) that completely pass the two-step transparency examination.

Many of the transparent data sources in the IDEA database primarily come from Japanese government agencies, including the Ministry of Economy, Trade and Industry; the Ministry of Agriculture, Forestry and Fisheries; and the Ministry of Land, Infrastructure, Transport and Tourism. These agencies consistently maintain an extensive collection of historical statistical documents and industry reports, ensuring public access and transparency.

## 3.7 | Comparison

The comparative analysis across all databases reveals notable variations in cross-citation complexity, source findability, and accessibility, shedding light on the strengths and challenges of each database. ELCD exhibits the highest cross-citation complexity, followed by GaBi, while ecoinvent shows the lowest complexity. This suggests that databases like ELCD and GaBi have more intricate interconnections between sources, potentially complicating transparency and traceability.

In terms of source findability and accessibility, ecoinvent and USLCI demonstrate relatively higher transparency. In contrast, ELCD and CLCD face significant challenges. The main reasons for unfindable sources include missing key information (such as author, year, or title), incomplete citations, and in some cases, a complete lack of source documentation for certain processes. Recurring issues across databases include inconsistent citation styles (e.g., IDEA's unstructured "General Comment" field and GaBi's varying citation formats), which make it difficult to locate key information.

**TABLE 3** Main issues restricting data transparency and possible solutions.

| Aspect | Main issues restricting data transparency | Possible solutions and suggestions |
|---|---|---|
| Integrity | Missing or incomplete source information makes it difficult for users to understand the process data and determine its suitability. | Implement detailed documentation of source information from the very beginning of data compilation, emphasizing thorough record-keeping and annotations to ensure the integrity. |
| Consistency | The lack of consistent source information, along with the unstructured and varied citation styles, further complicates the identification of key information. | Adopt consistent citation formats across all databases by using a uniform schema for documenting sources, ensuring that all essential details (e.g., author, year, and title) are consistently included. |
| Openness | The source information may be inaccessible due to a variety of factors, such as outdated information, changing URLs, paywalls, lack of electronic formats, regional or institutional restrictions. | Maximize open access to publications such as reports and papers, and ensure periodic review and updating of source information to maintain accessibility. |
| Traceability | Complexity of cross-citations makes it difficult to trace sources and even a small percentage of unfindable sources can affect a large number of processes due to interconnectivity. | Reduce complexity of cross-citations by providing simple unit processes and openly sharing these unit processes. |
| Granularity | Aggregation of LCI results makes it challenging to understand the connection between specific inputs/outputs and their sources. | Instead of aggregating sources, incorporate source details directly into the description of each process flow when publishing an LCI result. |

Common issues for inaccessible sources across multiple databases include paywalls, login requirements, links to general pages instead of specific data, book sources not readily available digitally, and reliance on proprietary databases that are not openly accessible. These barriers significantly reduce the usability of the data for external users.

The types of sources cited significantly influence the transparency of each database. ecoinvent primarily relies on structured references like reports and journal articles, which often include complete metadata, enhancing findability. GaBi relies heavily on datasets and EPDs but often links to general pages or requires logins, creating transparency barriers. USLCI predominantly cites accessible government and industry reports, contributing to its higher transparency. In contrast, ELCD and CLCD frequently reference proprietary databases like GaBi, restricted by paywalls, limiting transparency. IDEA mainly uses Japanese government publications, which are generally accessible but occasionally lack detailed metadata, hindering findability and accessibility.

Overall, across all databases, common barriers to findability and accessibility include missing metadata (e.g., author, year, and title), inconsistent citation styles, incomplete documentation, and reliance on inaccessible or proprietary sources. These recurring issues highlight the need for standardized citation formats, structured metadata, and open access to improve transparency and usability.

## 4 | DISCUSSION

Table 3 provides a comprehensive overview of main issues in various aspects that restrict data transparency in LCI databases, along with corresponding possible solutions and suggestions for improvement. The following subsections will elaborate discussion in depth.

### 4.1 | Complete and consistent source for findability

We find that many processes in various databases are recorded with missing source information or without any source information at all. Even when sources are listed, often the details are not thorough enough to ensure transparency. Missing or incomplete source information directly impacts the ability of practitioners to assess the quality and suitability of process data for their specific case studies. Without sufficient metadata, practitioners may face challenges in selecting appropriate datasets or verifying the context in which the data was collected, leading to potential inaccuracies in the results.

In addition, an issue we have noticed is that source information often is not placed in a consistent structure among various databases. Take the IDEA database as an example. It lists all sources in a "General Comment" field. This approach allows for explaining the connection between process data and sources, but the unstructured long text increases the difficulty for users to locate specific information. On the other hand, the GaBi database provides detailed citations for its data sources, which is commendable. However, the citation style lacks consistency, as demonstrated by the following examples. Specifically, the ordering of author, date, title, and other information vary a lot, making it challenging to identify key information for sources.

"Pereira, T. C.; Seabra, T.; Maciel, H.; Torres, P., Portuguese Environmental Agency, Portuguese Informative Inventory 1990–2008, Amadora, Portugal, 2010"

"Assessment and Recommendations for Improving the Performance of Waste Containment Systems, David A. Carson, U.S. EPA Office of Research and Development, National Risk Management Research Laboratory, Cincinnati, OH, December 2002."

To address these challenges, database providers need to consider two pivotal points. First, a detailed documentation of source information from the start of data compilation is essential. This is not just best practice; it is fundamental for upholding the integrity of the database. The importance of detailed record-keeping and annotations is emphasized for comparability and reliability of data (Khirfan et al., 2020; Trautwein, 2021). Second, standardizing the structure of source data is crucial. A uniform framework would not only make documenting and retrieving information more efficient but also ensure the interoperability and accuracy of databases (Fritter et al., 2020; Zanghelini et al., 2016). This standardization should extend beyond mere formatting to encompass the use of structured data formats and a comprehensive source schema. Data formats such as XML used by LCI databases allow for the creation of nested, hierarchical structures that can capture the complexity of bibliographic data more effectively than a single, unstructured "source" field. Furthermore, implementing a standardized source schema within these structured formats can address the inconsistency issues observed in databases like GaBi. A well-defined schema could include fields such as title, authors (potentially further subdivided into first name, last name, and affiliation), year, type of source, DOI, URL, publisher, and location. This level of granularity not only facilitates more precise searching and filtering of sources but also enables automated generation of properly formatted citations, regardless of the original input format.

Of course, this approach is not without its own challenges, including the costs of implementation, issues surrounding data privacy, and the technical and resource burdens of retrofitting existing databases. Nevertheless, there is a compelling argument for constructing standards to guide the development of future databases. By establishing a set of well-defined norms and protocols, we can ensure that upcoming databases are built on a foundation of interoperability and transparency. This forward-looking approach not only streamlines data management but also paves the way for more robust and versatile data ecosystems. As new databases emerge, adhering to these standards from the outset can significantly reduce the complexity and cost associated with data integration and analysis, fostering a more cohesive and efficient research environment for LCA.

## 4.2 | Open and updating source for accessibility

In our quest for the accessibility of sources, we navigate through a diverse landscape of scenarios. We encounter instances where the provided source information leads to a dead end on the Internet, with absolutely no related information in sight. There are sources that pointed us to websites, often just a company's main page, which unfortunately offer no clear connection to the intended information. Some sources are indeed traceable online, but barriers such as paywalls or the lack of an electronic format—think of publications that require a visit to a specific library—made them inaccessible. Notably, a minority of sources are either primary data or personal communications. We exclude these scarce but problematic sources and their corresponding processes from our study. However, it is important to clarify that this exclusion does not inherently label these sources as inaccessible. Particularly, the processes linked to sources that are either behind paywalls or involve primary data present a unique challenge, raising questions about the methods we use to verify its credibility, which is beyond the scope of this study.

In cases where the information from sources appears complete yet remains inaccessible, several factors might be at play. Outdated information and changing URLs often make data unreachable. Access to certain materials may be regionally or institutionally restricted, and some publications might be absent from main academic databases. Limited digitization and legal restrictions also hinder online access. However, it is also possible that the information is incorrect or not authentic. Regardless of the reason, the inability to locate data essentially equates to not providing a valid source, posing significant challenges to the transparency of the data. This lack of accessibility creates a barrier to compliance with ISO 14040/14044 standards, which emphasize transparency and reproducibility in LCA studies.

In addressing the challenges of source accessibility encountered in our research, there are two promising directions worth pursuing. We advocate for the maximal open sourcing of reports and papers in LCA studies. This approach would significantly enhance the transparency and replicability of research, making it easier for others to access, verify, and build upon the existing work. Open sourcing can include making full datasets, methodologies, and results available in public repositories, where they can be freely accessed and used by the wider research community. Additionally, ensuring the authenticity of source information and its regular updating is crucial. Given the dynamic nature of data and the frequency with which websites and digital platforms are updated or become obsolete, maintaining current and accurate source information is essential. This would involve not only initial thorough examination of sources but also periodic reviews and updates to ensure continued relevance and accuracy. This practice would mitigate issues related to outdated or inaccessible data, thereby improving the overall quality and reliability of LCI database and LCA research. By implementing these strategies, the significant barriers to data transparency encountered in LCI databases may be addressed, thereby contributing to the credibility of LCA research and its broader applications in sustainable development.

## 4.3 | LCI results or unit process

Our analysis also reveals that the complexity of cross-citations significantly influences the overall transparency of LCA databases. The IDEA database, for instance, despite a considerable proportion of sources being unfindable—nearly 40%—demonstrates high transparency as over one third of sampled processes passed the two-step examination. This is attributed to its lower complexity of cross-citations, which ensures that each untraceable source compromises fewer processes. In contrast, the ELCD has a mere 4.6% of sources classified as unfindable, yet these affect up to 40% of processes, underscoring a critical issue of source–process interconnectivity. These cascading effects make it difficult to trace errors or inconsistencies, reducing the reliability of the overall LCA results.

Complexity of cross-citation between processes and sources not only heightens the challenge of locating and tracing sources, but also means that any lack of clarity in one source propagates through a larger segment of the database. A significant factor contributing to this complexity is that the bulk of the processes within databases are LCI results derived from multiple sources. GaBi, in particular, provides LCI results, upholding the principle that unit process data must be technically compatible within a single system for reliability and representativeness. They caution against random connections without verifying technical consistency, as this could yield inaccurate results even if unit processes are disclosed (Thinkstep GaBi, 2021). Regarding the transparency issue, GaBi claims to provide comprehensive documentation that encompasses all critical technical facts for aggregated processes by. However, the reality confronts practitioners with opacity. Although GaBi documents a multitude of source information, ensuring thorough transparency remains a challenge. The problem lies in the difficulty of understanding which specific inputs and outputs are connected to specific sources; simply listing multiple sources is not enough.

Simple unit processes, despite potential privacy concerns, significantly improve process transparency. Openly sharing these unit processes is recommended, as detailed information equips users to utilize them as needed. Users are then responsible for ensuring the technical consistency of the processes within their own systems. When it is necessary to publish an LCI result, we advocate for the incorporation of source details directly into the description of each process flow. Such an approach shifts from an overwhelming aggregation of sources to a more streamlined and elucidating system. Here, the provenance of each data point becomes immediately evident, ensuring that the lineage of information is transparent and easily discernible.

## 5 | CONCLUSIONS AND LIMITATIONS

This study demonstrates that mainstream LCI databases currently suffer from low transparency, which significantly impacts the reliability of LCA results. Through a systematic evaluation across major databases, we identified widespread issues related to incomplete and inaccessible data sources. Our findings highlight the need for reconstructing LCI databases to enhance transparency, tracing the lineage of each data point back to its origin.

Our research makes three key contributions. First, we developed a check system to assess LCI data transparency based on the findability and accessibility of sources. This provides a standardized methodology to evaluate database transparency. Second, by applying this system to major databases, we revealed significant shortcomings in current data transparency practices. Very few processes passed our two-step examination, underscoring the scale of improvements required. Third, our analysis of the factors driving low transparency enabled targeted recommendations to enhance traceability in future databases.

Limitations exist in our study. Our database selection did not encompass regional LCI databases from developing countries and regions like South Asia, Southeast Asia, Latin America, and Africa. This indicates possible regional bias in our sample, excluding a significant portion of the global LCI database landscape. This study assumes complete referencing between LCI result datasets and unit process datasets, which may not always hold true due to factors like data confidentiality, aggregation, or technical limitations. Therefore, the presented transparency scores should be interpreted as potential upper limits. Future research could assess the completeness of references between datasets for a more comprehensive understanding of LCA database transparency. Additionally, examining database interoperability could reveal how transparency issues compound when integrating data across multiple sources, providing insights into enhancing LCI data reliability and usability.

Nevertheless, this research highlights the crucial role of data transparency and takes an important step in constructing solutions. By emphasizing data traceability from the ground up and retrofitting current databases, the LCA community can enhance trust and enable better evidence-based decision-making. We hope our findings galvanize collective action to uplift transparency standards, propelling LCA firmly into the future as a driver for sustainability.

### ACKNOWLEDGMENTS

The authors gratefully appreciate the financial support from National Key R&D Program of China (grant number 2023YFE0122400), and National Natural Science Foundation of China (grant numbers 52293445 and 52300234).

### CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

### DATA AVAILABILITY STATEMENT

The data that supports the findings of this study are available in the supporting information of this article.

## ORCID

*Jing Guo* https://orcid.org/0000-0003-1962-6559
*Yutao Wang* https://orcid.org/0000-0001-8297-8579

## REFERENCES

Astudillo, M. F., Treyer, K., Bauer, C., Pineau, P.-O., & Amor, M. B. (2017). Life cycle inventories of electricity supply through the lens of data quality: Exploring challenges and opportunities. *The International Journal of Life Cycle Assessment*, *22*(3), 374–386. https://doi.org/10.1007/s11367-016-1163-0

Bertino, E., Merrill, S., Nesen, A., & Utz, C. (2019). Redefining data transparency: A multidimensional approach. *Computer*, *52*(1), 16–26. https://doi.org/10.1109/MC.2018.2890190

Ciroth, A., & Burhan, S. (2021). Life cycle inventory data and databases. In A. Ciroth & R. Arvidsson (Eds.), *Life cycle inventory analysis: Methods and data* (pp. 123–147). Springer International Publishing. https://doi.org/10.1007/978-3-030-62270-1_6

Finnveden, G., Hauschild, M. Z., Ekvall, T., Guinée, J., Heijungs, R., Hellweg, S., Koehler, A., Pennington, D., & Suh, S. (2009). Recent developments in life cycle assessment. *Journal of Environmental Management*, *91*(1), 1–21. https://doi.org/10.1016/j.jenvman.2009.06.018

Fritter, M., Lawrence, R., Marcolin, B., & Pelletier, N. (2020). A survey of life cycle inventory database implementations and architectures, and recommendations for new database initiatives. *The International Journal of Life Cycle Assessment*, *25*(8), 1522–1531. https://doi.org/10.1007/s11367-020-01745-5

Guinée, J. B., Heijungs, R., Huppes, G., Zamagni, A., Masoni, P., Buonamici, R., Ekvall, T., & Rydberg, T. (2011). Life cycle assessment: Past, present, and future. *Environmental Science & Technology*, *45*(1), 90–96. https://doi.org/10.1021/es101316v

Hellweg, S., & Milà i Canals, L. (2014). Emerging approaches, challenges and opportunities in life cycle assessment. *Science*, *344*(6188), 1109–1113. https://doi.org/10.1126/science.1248361

Hertwich, E., Heeren, N., Kuczenski, B., Majeau-Bettez, G., Myers, R. J., Pauliuk, S., Stadler, K., & Lifset, R. (2018). Nullius in verba: Advancing data transparency in industrial ecology. *Journal of Industrial Ecology*, *22*(1), 6–17. https://doi.org/10.1111/jiec.12738

Joyce, P. J., & Björklund, A. (2022). Futura: A new tool for transparent and shareable scenario analysis in prospective life cycle assessment. *Journal of Industrial Ecology*, *26*(1), 134–144. https://doi.org/10.1111/jiec.13115

Kalverkamp, M., Helmers, E., & Pehlken, A. (2020). Impacts of life cycle inventory databases on life cycle assessments: A review by means of a drivetrain case study. *Journal of Cleaner Production*, *269*, 121329. https://doi.org/10.1016/j.jclepro.2020.121329

Khirfan, L., Peck, M. L., & Mohtat, N. (2020). Digging for the truth: A combined method to analyze the literature on stream daylighting. *Sustainable Cities and Society*, *59*, 102225. https://doi.org/10.1016/j.scs.2020.102225

Kuczenski, B. (2019). Disclosure of product system models in life cycle assessment: Achieving transparency and privacy. *Journal of Industrial Ecology*, *23*(3), 574–586. https://doi.org/10.1111/jiec.12810

Pauliuk, S., Majeau-Bettez, G., Mutel, C. L., Steubing, B., & Stadler, K. (2015). Lifting industrial ecology modeling to a new level of quality and transparency: A call for more transparent publications and a collaborative open source software framework. *Journal of Industrial Ecology*, *19*(6), 937–949. https://doi.org/10.1111/jiec.12316

Reale, F., Cinelli, M., & Sala, S. (2017). Towards a research agenda for the use of LCA in the impact assessment of policies. *The International Journal of Life Cycle Assessment*, *22*(9), 1477–1481. https://doi.org/10.1007/s11367-017-1320-0

Saade, M. R. M., Gomes, V., da Silva, M. G., Ugaya, C. M. L., Lasvaux, S., Passer, A., & Habert, G. (2019). Investigating transparency regarding ecoinvent users' system model choices. *The International Journal of Life Cycle Assessment*, *24*(1), 1–5. https://doi.org/10.1007/s11367-018-1509-x

Saavedra-Rubio, K., Thonemann, N., Crenna, E., Lemoine, B., Caliandro, P., & Laurent, A. (2022). Stepwise guidance for data collection in the life cycle inventory (LCI) phase: Building technology-related LCI blocks. *Journal of Cleaner Production*, *366*, 132903. https://doi.org/10.1016/j.jclepro.2022.132903

Thinkstep Gabi. (2021). GaBi Database & Modelling Principles 2017 Edition.

Trautwein, C. (2021). Sustainability impact assessment of start-ups—Key insights on relevant assessment challenges and approaches based on an inclusive, systematic literature review. *Journal of Cleaner Production*, *281*, 125330. https://doi.org/10.1016/j.jclepro.2020.125330

Wernet, G., Bauer, C., Steubing, B., Reinhard, J., Moreno-Ruiz, E., & Weidema, B. (2016). The ecoinvent database version 3 (part I): Overview and methodology. *The International Journal of Life Cycle Assessment*, *21*(9), 1218–1230. https://doi.org/10.1007/s11367-016-1087-8

Wu, S. R., & Wang, L. (2022). Higher transparency: A desideratum in environmental life cycle assessment research. *Journal of Cleaner Production*, *374*, 134074. https://doi.org/10.1016/j.jclepro.2022.134074

Zanghelini, G. M., de Souza Junior, H. R. A., Kulay, L., Cherubini, E., Ribeiro, P. T., & Soares, S. R. (2016). A bibliometric overview of Brazilian LCA research. *The International Journal of Life Cycle Assessment*, *21*(12), 1759–1775. https://doi.org/10.1007/s11367-016-1132-7

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Guo, J., Li, R., Zhang, R., Qi, J., Li, N., Xu, C., Chiu, A. S. F., Wang, Y., Tanikawa, H., & Xu, M. (2025). Shedding light on the shadows: Transparency challenge in background life cycle inventory data. *Journal of Industrial Ecology*, *29*, 766–776. https://doi.org/10.1111/jiec.70010