Universiteit Leiden
The Netherlands

Bachelor Computer Science & Economics

Data Quality in

Supply Chain Mapping

Emma Glorie

First supervisor: Drs. J.B. Kruiswijk - Leiden University
Second supervisor: E. Rietveld, MSc. - TNO

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

15/01/2026

Acknowledgements

Special thanks go to my supervisors, Bas Kruiswijk and Elmer Rietveld, for their guidance, constructive feedback, and support throughout the research process, which were invaluable to the development of this thesis.

I would also like to express my gratitude to TNO for providing the opportunity to conduct this research as part of an internship. The professional environment and support offered during this period were essential to the successful completion of this thesis and allowed me to gain valuable practical experience alongside my academic work.

Finally, I would like to thank the experts who participated in the interviews, Arnoud van Bree, Dejan Zamurović, Jasper van Kempen, Mechteld Bakkenes, and Sena Yaman, for generously sharing their time, expertise, and perspectives. Their contributions provided important practical insights and significantly enriched this research.

Abstract

In an era of heightened geopolitical disruptions and environmental challenges, supply chain mapping has become increasingly important for organisations and policymakers seeking to manage risks and improve supply chain resilience. However, the quality of datasets used for supply chain mapping varies widely and often lacks a structured approach for assessing their quality, making it difficult to evaluate their reliability and suitability for specific analytical purposes.

This thesis addresses this challenge by answering the main research question of how the data quality of supply chain mapping datasets can be systematically assessed. First, it examines how datasets used for supply chain mapping can be categorised. Based on a literature review, a typology of supply chain mapping datasets is developed, distinguishing between different dataset characteristics, mapping method, scope, resolution and data sources. Next, the research investigates which data quality criteria are relevant for assessing such datasets. Drawing on established data quality literature, a set of nine key criteria is identified that reflects the specific data quality challenges in supply chain mapping: data fit, accuracy, completeness, consistency, timeliness and temporal consistency, representativeness, transparency and documentation, accessibility and comparability, and validation and verification. Then, the study focuses on how these criteria can be operationalised into a practical assessment approach. To this end, the criteria are translated into a Data Quality Assessment Framework explicitly grounded in the fitness-for-use principle, acknowledging that data quality depends on the intended purpose and context of use rather than on absolute standards.

The framework is applied to a case study on wind turbine production using the AI-Generated Production Network (AIPNET) dataset and is further evaluated through semi-structured expert interviews with domain experts in supply chain analysis and data science. There were no major obstacles in applying the framework to the AIPNET dataset in the case study. The feedback from the experts further confirms the framework's practical feasibility and its alignment with existing analytical practices. The experts suggest improvements, share how they would apply the framework in practice, and identify aspects that may be challenging or open to discussion. The results from the case study and the feedback from the expert validation are evaluated and incorporated into the final Data Quality Assessment Framework.

In conclusion, the framework contributes to greater transparency and reliability in supply chain data use by providing a systematic yet flexible tool for data quality assessment, applicable to both policy analysis and academic research. Rather than serving as a rigid checklist, it guides users in making context-sensitive, transparent judgements about data fitness. While the framework enhances consistency in evaluating data, it does not fully eliminate subjectivity. Future research is recommended to extend and standardise this approach across different contexts, further strengthening its utility and credibility.

Contents

Chapter 1. Introduction

1.1 Problem statement

In 2023, the European Commission wrote a joint communication to the European Parliament, the European Council and the Council on "European Economic Security Strategy", addressing the growing vulnerabilities in Europe's economies and supply chains (European Commission, 2023a). The global pandemic, Russia's war in Ukraine, hostile economic actions, cyber and infrastructure attacks, foreign interference and disinformation, and a global increase in geopolitical tensions have exposed new risks and vulnerabilities in European societies, economies, and industries. These risks include "risks of price surges, the unavailability or scarcity of critical products, or inputs in the EU, including but not limited to those linked to the Green Transition, those needed for a stable and diversified energy supply and pharmaceuticals."

The Joint Communication stresses the need for the EU to reduce vulnerabilities, diversify supply chains, and secure access to critical resources. To achieve these goals and support further decision making, global supply chains need to be mapped accurately. Such mapping contributes to the need to increase transparency in supply chains. This transparency supports sustainability goals that align with the People & Planet dimensions of sustainable development (Elkington, 1999) and the goals of the European Green Deal (European Commission, 2019). Supply chain mapping refers to the systematic identification of linkages between firms, products, and processes, often using trade or customs data, to make interdependencies and vulnerabilities visible (Mubarik et al., 2021).

In the Netherlands, this issue is also increasingly recognised. Statistics Netherlands (CBS) recently analysed the country's dependence on critical raw materials (Statistics Netherlands (CBS), 2025). The report shows that the Netherlands is one of the largest importers of raw materials in the EU, primarily due to its role as a European logistics hub. In 2024, the total import value of critical materials amounted to €22.3 billion, with 99 percent sourced from non-EU countries and one-quarter coming directly from China. This strong external dependence exposes the Dutch economy to strategic supply risks.

These observations, both at European and national levels, highlight the importance of accurate supply chain mapping. However, the reliability of such mappings ultimately depends on the quality of the underlying data. For example, if the data is incomplete, a supply chain map may misrepresent the real world. In practice, companies and policymakers often rely on commercial databases for supply chain mapping, yet it is often unclear which databases provide reliable and up-to-date information. Before making substantial investments or strategic decisions, organisations need to know whether a database can be trusted to accurately represent supply chain relationships. The absence of clear criteria for assessing data quality therefore creates both financial and operational risks. This thesis examines how data quality in supply chain mapping methods can be systematically assessed.

1.2 Research questions

To address the challenges described in the previous section, this study aims to develop a structured framework for systematically assessing data quality in supply chain mapping methods.

The central research question of this thesis is:

How can the data quality of datasets for supply chain mapping be systematically assessed?

This research question is examined through three subquestions:

RQ1. What types of datasets are used for supply chain mapping, and what are their main characteristics?

RQ2. Which criteria can be applied to assess the data quality of supply chain mapping datasets?

RQ3. For each criterion, how can the pass conditions and importance for the evaluation of the different datasets be defined?

1.3 Link with Computer Science & Economics

This research directly relates to the field of Computer Science and Economics, as it examines a core computer science topic, namely data quality assessment, in an economic and organisational context. By analysing the quality of datasets used for supply chain mapping, the thesis shows how methods from data science influence the economic interpretation and reliability of the information that decision-makers rely on. This connection aligns with the programme's emphasis on understanding how computational tools gain meaning when applied to real-world economic environments. The study illustrates how the properties of data affect the insights that organisations can derive about supply chains, linking technical analysis with economic reasoning.

1.4 Context of the thesis: TNO

This thesis is conducted as a graduation internship at TNO (the Netherlands Organisation for Applied Scientific Research). TNO is an independent research organisation whose mission is to connect people and knowledge to develop innovations that strengthen the competitiveness of businesses and enhance the sustainable well-being of society. By translating scientific insights into practical applications, TNO bridges the gap between academic research and real-world implementation.
The internship takes place within the Business & Finance Models (BFM) department. BFM supports decision-makers in both the private and public domain by analysing the value, risks and strategic implications of new technologies and organisational innovations. The department contributes to major societal transitions by developing business models, strategic analyses and collaborative frameworks that link technological developments with economic and societal impact. Another department involved at TNO is the Netherlands Materials Observatory (NMO). The NMO collects, manages, and provides data, information, and knowledge about raw materials and supply chains that are essential to the Dutch economy.
This setting provides a relevant context for the research. Working within BFM allows the thesis to benefit from both academic and practical perspectives. Access to internal TNO resources and

the opportunity to consult experts further strengthens the research and places it in a real-world context. Altogether, this creates a unique environment for conducting research on data quality in datasets used for supply chain mapping.

1.5 Thesis overview

This thesis is structured into several chapters. Chapter 2 provides the theoretical background by reviewing literature on supply chain mapping datasets and data quality, distinguishing between different dataset types, and identifying the research gap. Chapter 3 describes the methodology, including the research design and expert validation approach. Chapter 4 gives an overview of the most important sources of data quality criteria in the literature. Chapter 5 presents the set of relevant criteria and introduces the initial Data Quality Assessment Framework. Chapter 6 applies the framework to a case study, while Chapter 7 presents the expert validation. Chapter 8 integrates the findings from the case study and expert validation and presents the final Data Quality Assessment Framework. Chapter 9 discusses the findings and their implications. Finally, Chapter 10 concludes the thesis.

Chapter 2. Background and related work

2.1 Types of datasets

The datasets used for supply chain mapping do not fall into one fixed category. They vary widely in structure, source, and purpose. To accurately evaluate data quality and understand the strengths and limitations of each approach, it is important to first distinguish between the main dataset types encountered in this field.
In this study, we distinguish datasets along three key dimensions:

- First, by the method used to map the supply chain. Some datasets are empirically collected, while others are AI-enhanced or fully AI-generated. In practice, many datasets are a combination of the different methods.
- Second, by the scope and resolution at which the mapping is made. For example, product-to-product mappings versus business-to-business mappings.
- Finally, by the source of the data. Datasets can be based on a range of sources, such as customs data, company reports, or large-scale data extracted from online and big data platforms.

These distinctions form the foundation for the following subsections, where each category will be discussed in more detail.

2.1.1 Mapping methods

The choice of method determines both the structure and the quality of the resulting dataset. Broadly, supply chain mappings can be derived from empirically collected data, enhanced through AI, or fully generated using AI-based models. Understanding these methods is essential to evaluate the origins and reliability of datasets used in supply chain research.

Empirically collected datasets
Empirically collected datasets are built on data obtained directly from real-world sources such as company disclosures, surveys, interviews, or internal company systems. This traditional approach relies on verified, observable information, allowing for relatively high accuracy and interpretability (MacCarthy et al., 2022). Such data often capture the actual relationships and flows within or between firms, which also makes them valuable for validating AI-driven methods.
However, the empirical approach also faces limitations. Data collection can be time-consuming and costly, and access is frequently restricted due to confidentiality, as well as competitive or commercial sensitivity. Furthermore, the available data is often incomplete and biased toward large, publicly listed firms (Culot et al., 2022). Smaller suppliers and firms from developing economies are frequently underrepresented, which reduces the representativeness of these datasets (Dong et al., 2024). As a result, empirically collected datasets tend to be narrower in scope, focusing on specific industries or case studies rather than global networks (MacCarthy et al., 2022). Despite these limitations, they remain the foundation of reliable supply chain mapping, offering a direct and verifiable representation of real-world supply chain relationships.

AI-enhanced datasets

AI-enhanced datasets build upon empirically collected data by integrating AI and data-mining techniques to expand and refine supply-chain mappings. These hybrid datasets start from verified information, such as firm disclosures, trade statistics, or logistics data, and use algorithms to identify missing connections or update existing relationships dynamically. For example, this can involve techniques such as association-rule mining, backward-forward algorithms, and probabilistic estimation to detect supply chain relationships when direct data is unavailable (Mau and Vicencio, 2025).

Commercial and research initiatives such as Sourcemap, EarlyBirds, and Altana AI have already applied these techniques to provide a more comprehensive and up-to-date view of supply chain networks. However, because parts of these datasets are algorithmically derived rather than directly observed, such datasets may contain uncertain or misleading connections (Dong et al., 2024). Moreover, methodological differences in data construction can substantially affect the structure and coverage of resulting networks (Dong et al., 2024). Transparency and clear documentation are therefore essential when combining empirically collected data with AI-based estimations (Culot et al., 2022).

While AI-enhanced datasets offer an effective balance between scale and detail, their reliability ultimately depends on how thoroughly they are validated and on the transparency of the underlying assumptions.

AI-generated datasets

AI-generated datasets represent the newest category, where supply chain mappings are created almost entirely by AI to create independent, large-scale representations of supply chain networks. Instead of relying directly on empirically collected data, these datasets are typically constructed using large language-models or embedding-based algorithms that identify relationships between products based on textual, semantic, or technological similarities (Fetzer et al., 2024).

A key example is the AI-Generated Production Network (AIPNET), developed by Fetzer et al. (2024). The model uses Harmonized System (HS) product codes and their textual descriptions as its only empirical input. The Harmonized System is "a nomenclature of product codes maintained by the World Customs Organization (WCO) and used universally by national trade authorities" (Fetzer et al., 2024). GPT-4o was prompted to analyse these HS descriptions and to determine which products are likely connected in production processes, either as inputs or supporting goods. Based on this analysis, the model generates a product-to-product network that represents potential production linkages between goods. In this way, all connections in AIPNET are the result of AI reasoning rather than observed trade or firm-level data.

AI-generated datasets such as AIPNET offer highly granular representations of supply chains and allow for scalable insights, revealing interdependencies that are difficult to observe in empirically collected or AI-enhanced data.

However, these methods introduce challenges related to transparency, interpretability, and validation, as the generated links depend on algorithmic reasoning rather than direct observation. As emphasized for AI-enhanced datasets, responsible use therefore requires clear documentation of modelling assumptions and comparison with empirically verified benchmarks (Culot et al., 2022).

Together, these methods illustrate the variety of supply chain mapping methods. Each method offers different advantages regarding accuracy, scale, and accessibility, making it crucial to understand their methodological foundations when assessing dataset quality.

2.1.2 Scope and resolution

Beyond the method used to construct them, supply chain mapping datasets also differ in their scope and resolution. These two characteristics determine the breadth and depth of the representation of the networks and are therefore crucial for interpreting the type of insight each dataset can provide. Scope refers to the geographical and sectoral coverage of the dataset. It concerns whether the data captures linkages within a single country, across regions, or on a global scale, and whether it focuses on one industry or spans multiple sectors. Resolution reflects the degree of detail within that scope, such as whether relationships are represented at the company, sector, or even national level (MacCarthy et al., 2022).

In terms of scope, datasets that cover multiple countries or sectors are typically designed to visualise trade or production relationships at a global level. Their resolution is usually coarse, as they combine transactions across firms and product types (Gardner et al., 2019). This enables macro-level analyses such as assessing exposure to geopolitical risks or identifying trade dependencies, but limits the ability to capture detailed supplier relationships or variations within countries (MacCarthy et al., 2022).

By contrast, datasets with a narrower scope, such as those focusing on a single country or sector, tend to achieve higher resolution. They often record relationships between individual firms and their direct suppliers or customers within the same network. This finer detail allows a more complete view of supply networks within a defined region or industry and supports operational goals such as risk management and traceability (Montecchi et al., 2021). However, because the scope is limited, such datasets are often less comparable across countries and risk overlooking international linkages that extend beyond the defined boundary of analysis (Gardner et al., 2019).

Overall, the distinction between scope and resolution highlights a fundamental trade-off in dataset design. Broad scope offers comprehensive coverage of international or cross-sector linkages but typically comes with low resolution. Narrow scope allows for high-resolution mapping of specific networks but limits the generalisability of insights. Understanding these differences is important when assessing the quality of supply chain mapping datasets, as they determine the kind of analyses that can be reliably conducted.

2.1.3 Data sources

The origin of the underlying data is another key dimension when studying data quality in supply chain mapping. In practice, most datasets are not based on a single source type but combine information from multiple origins. Understanding these sources is essential to assess the strengths and limitations of existing mapping datasets and to evaluate their overall data quality. Broadly, we distinguish the sources used in supply chain mapping as primary, secondary, and big data.

Primary data refers to information collected firsthand from supply chain actors or through direct observation. In the context of supply chain mapping, this may include surveys, interviews, or supplier lists provided by the company. Such data is often highly specific and verifiable, offering accurate insight into actual supplier relationships (MacCarthy et al., 2022). However, the collection process is time-consuming, access is frequently restricted due to confidentiality, and coverage is typically limited to large or cooperative firms (Culot et al., 2022). As a result, primary sources provide strong internal validity, making them most effective for focused analyses but limited for comprehensive supply chain overviews.

Secondary data is information originally collected for another purpose and later used for another goal such as supply chain mapping. Common examples include customs and trade statistics, company reports, sustainability disclosures, and government or statistical databases (Calantone and Vickery, 2010). These sources form the foundation of many large-scale or cross-country mappings because they provide standardized, comparable information and have a broad coverage. Customs data, for example, capture trade flows between countries and industries, enabling the identification of international dependencies (Gardner et al., 2019). However, the sources may exclude domestic transactions and they often combine data across firms and products, reducing resolution. Similarly, company reports and sustainability disclosures can provide firm-level detail but vary widely in reporting standards and quality, leading to gaps and inconsistencies (Culot et al., 2022). Secondary data therefore offers broad coverage but raises challenges in completeness, comparability, and transparency.

Big data has become an increasingly important source for constructing and updating supply chain mappings. It is a type of secondary data, but can be distinguished by its scale, structure and collection method. Unlike traditional datasets, big data is continuously generated from digital activities and often takes the form of large, unstructured information that must be processed and verified using automated techniques. Typical examples include web-scraped supplier lists, online product registries, patent data, and shipping and logistics records. These sources enable researchers and firms to identify supply chain connections that are not visible in traditional datasets (Dong et al., 2024). Qualitative big data, such as textual information from online and media sources, can also provide contextual insights into supply chain relationships that traditional quantitative data overlook (Bansal et al., 2020). While these sources improve scope and timeliness, they also introduce challenges in validation and transparency, as the origin and reliability of the information are not always clear (Cook et al., 2012).

Each of these categories contributes differently to supply chain mapping. Primary data offers precision, secondary data provides comparability, and big data enables scale. Most modern datasets combine these categories to balance their respective strengths and weaknesses. Recognising these differences is essential when evaluating the quality of supply chain mapping datasets, as the type and origin of data fundamentally determine their reliability.

2.2 Data quality challenges

Assessing the quality of supply chain mapping datasets is crucial for determining how accurately they represent real-world relationships. Their reliability determines the validity of the analyses, insights, and policy implications derived from them. However, data quality is a multidimensional and context-dependent concept, and assessing it in the context of supply chain mapping presents unique challenges. This section discusses how data quality can be understood conceptually and examines the main challenges associated with empirically collected, AI-enhanced, and AI-generated datasets.

2.2.1 Conceptual foundations of data quality

In this study, data quality is defined according to the classical "fitness-for-use" perspective commonly accepted in data quality research (Wang and Strong, 1996). This perspective conceptualizes high-quality data as data that is suitable for its intended analytical purpose. In other words, the

usefulness of data depends not only on its accuracy but also on how well it supports the specific task or analysis for which it is applied.
Within the literature, the interpretation of what constitutes "fitness-for-use" data varies across studies. Wang and Strong (1996) introduced the idea that quality should be evaluated according to its intended purpose rather than intrinsic accuracy alone. Later research has expanded on this by identifying various dimensions through which "fitness" can be assessed. Culot et al. (2022) emphasise challenges such as data fit, temporal consistency, and representativeness in large supply chain databases, while MacCarthy et al. (2022) highlight the importance of transparency and documentation for ensuring interpretability. The study *Data Quality Problems Troubling Business and Financial Researchers* (Liu, 2020) further points to recurring issues such as missing values, coding errors, and biases, which demonstrate that data quality is a complex and context-dependent construct. Taken together, these studies suggest that data quality cannot be reduced to a fixed set of dimensions or criteria but instead reflects a broader concern with whether data is appropriate, consistent, and reliable for its intended analytical purpose.
While this study applies the classical fitness-for-use definition as its conceptual foundation, it is important to note that recent literature also argues that this framework may require adaptation in specific contexts such as machine learning. Priestley et al. (2023) argues that data quality in machine learning should not be viewed as a single static property but rather as stage-specific, varying across the phases of data collection, cleaning, model training, deployment, and monitoring. Her perspective highlights that the traditional fitness-for-use framework may need reinterpretation when evaluating AI-enhanced and AI-generated datasets, where quality must be monitored continuously throughout the data lifecycle.

2.2.2 Data quality in empirically collected datasets

Empirically collected datasets form the foundation of most supply chain mapping research because they provide verifiable representations of production and trade relationships. However, despite their apparent credibility, prior studies have shown that such datasets are affected by persistent quality challenges that limit their analytical reliability and generalisability.
A recurring issue concerns representativeness. Data collection processes often rely on publicly available or voluntarily disclosed information, resulting in a bias toward large and listed firms. Smaller suppliers and firms located in developing economies are frequently excluded, leading to incomplete representations of global production networks (Culot et al., 2022). This systematic bias can misrepresent the structure of supply chains and create misleading impressions of connectivity and risk exposure.
Another key challenge relates to completeness and comparability. Even when data is collected from official or commercial sources, they differ substantially in coverage, structure, and data treatment. Dong et al. (2024) compares four major supply chain databases, Bloomberg SPLC, FactSet, Compustat, and Mergent. The study finds that differences in coverage and methodology across these databases can change the number of observed supplier-buyer links and even alter empirical findings. This highlights that database selection is not a neutral methodological choice but a potential source of systematic variation in research outcomes, complicating comparability across studies and limiting the generalisation of findings.
Temporal consistency is also a frequent concern. Firm-level datasets are often updated irregularly, contain overlapping reporting periods, or fail to document changes such as mergers and name

variations. These inconsistencies create difficulties in replicating analyses and in tracing how supply chain structures evolve over time (Liu, 2020). When databases are backfilled or rely on historical corporate records, inaccuracies in start and end dates can further reduce their suitability for longitudinal research.

A related issue concerns transparency and documentation. Many commercial databases provide little insight into how relationships are defined, verified, or updated. This lack of methodological transparency can reduce trust in widely used databases (Cook et al., 2012). When key information about data construction, verification, and maintenance remains undisclosed, users cannot judge whether the dataset meets quality criteria such as accuracy or completeness. As a result, assessments of reliability depend largely on assumptions rather than evidence, limiting the extent to which the data can be considered fit for analytical use.

To strengthen validity, several studies suggest combining secondary with primary data. For example, integrating firm-level surveys or interviews with secondary databases can confirm key relationships and reduce classification errors (Ellram and Tate, 2016). However, such triangulation efforts are generally confined to specific case studies due to time constraints, data accessibility, and the sensitivity of firm-level information.

Overall, empirically collected datasets provide verifiable information and are often perceived as the most reliable basis for supply chain analysis. Yet, their quality is not guaranteed and therefore requires a critical assessment.

2.2.3 Data quality in AI-enhanced and AI-generated datasets

The limitations of traditional datasets have encouraged the use of AI-enhanced or AI-generated datasets. These datasets use algorithmic inference and machine learning to expand, refine, or even fully construct supply chain networks. While they increase scalability and timeliness, they also introduce distinct challenges for assessing data quality, as their structure and reasoning differ fundamentally from empirically observed data.

AI-enhanced datasets build upon existing empirical data but use algorithms to infer missing connections or predict supplier relationships. This hybrid structure allows for broader coverage of global networks, yet it introduces uncertainty about the validity of inferred links. When training data contain structural biases or when inference rules are insufficiently documented, the resulting information may appear plausible but lack factual grounding (Culot et al., 2022). Such uncertainty undermines interpretability and complicates quality assessment, as users cannot easily separate verified relationships from algorithmically constructed ones.

In contrast, AI-generated datasets rely almost entirely on computational reasoning. The AI-Generated Production Network (AIPNET) illustrates this approach by using large language models to predict input-output linkages between products based on textual descriptions (Fetzer et al., 2024). While this method identifies previously invisible interdependencies, it also produces outputs that cannot be validated against observable data. The absence of a ground-truth benchmark makes traditional validation techniques difficult to apply and introduces the risk of algorithmic hallucination, which is the creation of connections that are statistically inferred but do not exist in reality (Bender et al., 2021).

A central issue across both AI-enhanced and AI-generated datasets is transparency and interpretability. Most AI models operate as “black boxes”, meaning that their internal decision processes are not visible or easily understandable to the user. Without disclosure of the data sources, parameter

settings, or model logic that generated each connection, researchers cannot determine how or why certain relationships were inferred (MacCarthy et al., 2022). This lack of interpretability reduces trust and limits reproducibility, as identical analyses can give different results when models are retrained or parameters change (Mitchell et al., 2019).

Efforts to validate AI-generated mappings remain limited and inconsistent. For example, AIPNET was partially validated by comparing its network structure to national Input-Output and Supply-Use tables, as well as trade volume data. These comparisons confirm alignment with global trade patterns but do not verify the accuracy of individual linkages (Fetzer et al., 2024). Such partial validation highlights a broader methodological gap: while empirically collected data can be checked against observable evidence, AI-generated data require alternative evaluation criteria that account for coherence, interpretability, and internal consistency rather than direct verification.

For AI-enhanced and AI-generated datasets, fitness-for-use depends less on factual accuracy and more on explainability, traceability, and consistency of the algorithms and their outputs. These criteria determine whether the data can be interpreted, replicated, and used responsibly in analytical or policy contexts.

In summary, AI-based datasets represent a transformative development in supply chain mapping, offering greater scope and real-time insights but at the cost of interpretability and transparency. While they expand analytical possibilities, their reliability cannot be assumed.

2.3 Research gap

Existing literature has provided valuable insights into data quality in supply chain mapping, but most studies address only specific aspects of the problem. Research has traditionally focused on empirically collected datasets, examining issues such as accuracy, completeness, representativeness, and transparency (MacCarthy et al., 2022; Culot et al., 2022). While this work has improved understanding of data reliability, it remains fragmented. Studies often analyse individual databases or isolated quality dimensions rather than developing a comprehensive framework for systematic assessment.

Culot et al. (2022) propose one of the most structured approaches for evaluating the reliability of supply chain databases. However, their framework is tailored to empirical data and primarily guides the selection and cleaning of specific datasets. It does not generalize to other data construction methods or offer a broader conceptual model that can compare quality across datasets. As new forms of data collection and inference are increasingly used, this absence of a unified framework limits both academic comparability and practical evaluation.

This study addresses this gap by developing a structured framework for systematically assessing data quality in supply chain mapping. It integrates existing quality criteria identified in prior research within a coherent, multi-dimensional model that can be applied across different dataset types. The framework aims to establish a transparent and generalisable approach for evaluating dataset reliability and for supporting both academic research and practical data use.

Chapter 3. Methodology

3.1 Research design

The study adopts a research design combining a literature review with framework development, complemented by an illustrative case study and expert interviews. The overall aim is to develop a structured framework for assessing data quality in datasets for supply chain mapping. The research design is structured into five main phases, each addressing a specific research objective and corresponding research question, as summarised in Table 1.

The first phase focuses on identifying the types of datasets used for supply chain mapping. This phase is addressed through a literature review and results in an overview of dataset characteristics relevant for supply chain mapping. These characteristics provide the contextual background for the study and form the basis for answering RQ1. This phase is presented in Chapter 2.

Building on this literature review, the second phase of the research aims to identify relevant data quality criteria for assessing supply chain datasets. The literature review is extended to include studies on data quality in general, as well as research specifically related to supply chain mapping. The most important sources of data quality criteria in the literature are presented in Chapter 4.

In the third phase, a set of clearly defined criteria relevant for supply chain mapping are selected from the overview of the second phase. This set of criteria forms the basis for the initial Data Quality Assessment Framework, which addresses RQ2 and is discussed in Chapter 5. The set of data quality criteria is translated into a structured and practical Data Quality Assessment Framework. The result of this phase is the initial Data Quality Assessment Framework, which addresses RQ3 and is presented in Chapter 5.

The fourth phase focuses on the application and validation of the framework. The framework is applied in a focused case study to illustrate how the framework is used in practice and to identify possible ambiguities or limitations. In parallel, interviews are conducted with experts in supply chain data and analytics. Through structured feedback and discussion, experts assess the relevance, clarity, and practical feasibility of the framework. The outcome of this phase is an illustrated framework derived from the case study (presented in Chapter 6) and complemented by an assessment of the practical relevance through expert interviews (presented in Chapter 7).

In the fifth phase, insights from both the case study and the expert interviews are used to refine the framework where necessary. The outcome of this phase is the final Data Quality Assessment Framework, addressing the main RQ and presented in Chapter 8.

Research phase	Objective	Method	Result	Research question
Phase 1: Dataset typology	Identify types of datasets used for supply chain mapping	Literature review	Overview of dataset characteristics relevant for supply chain mapping	RQ1
Phase 2: Data quality criteria	Identify data quality criteria for assessing supply chain datasets	Literature review	Overview of the most important sources of data quality criteria	
Phase 3: Initial framework development	Translate relevant data quality criteria into a practical assessment framework	Criteria selection and translation	Initial Data Quality Assessment Framework	RQ2 and RQ3
Phase 4: Framework application and validation	Illustrate the application of the framework and assess its practical relevance	Case study and expert interviews	Illustrated framework and framework assessed for practical relevance	
Phase 5: Final framework development	Process the feedback from case study and expert interviews	Feedback evaluation	Final Data Quality Assessment Framework	Main RQ

Table 1: Overview of the research design

3.2 Expert validation (semi-structured interviews)

To validate and where necessary improve the developed Data Quality Assessment Framework, interviews will be conducted with experts in relevant academic and applied research fields. The main goal is to assess the relevance, clarity, and practical feasibility of the framework, as well as to verify whether its underlying literature-based foundation aligns with real-world experience.
Each expert will participate in a semi-structured interview of approximately 30 minutes. This interview format provides consistency across conversations, which makes responses comparable and enables the collection of structured input. At the same time, it still allows experts to elaborate on their reasoning or point out elements that are unclear, incomplete or misaligned with practice. This combination of structured and open-ended elements is important, as the framework aims not only to be theoretically sound but also applicable in real-world dataset evaluations. The interview protocol is included in Appendix D. After each interview, a short summary of the expert's input is shared with them, so they can review what has been discussed and comment if anything needs

further clarification or correction.
To ensure that the interviews make efficient use of time and to allow experts to reflect on their responses beforehand, all interviewees receive a briefing document in advance. This document includes a shortened version of the research context, the dataset characteristics, the nine diagnostic criteria, the Data Quality Assessment Framework and the structure and questions of the interview. The briefing document is also provided in Appendix E.

Chapter 4. Literature on data quality criteria

A targeted literature review is conducted to identify recurring criteria used to assess the quality and suitability of datasets in academic research, with a particular focus on supply chain data. The review focuses on both general data quality frameworks and domain-specific studies, to help find criteria that are applicable to supply chain analysis. The sources listed below emerge as the most relevant:

Wang and Strong (1996)
"Beyond Accuracy: What Data Quality Means to Data Consumers"
This paper describes dimensions of data quality that are important to data consumers. It contains an extensive list of 179 potential data quality attributes, including accuracy, completeness, consistency, and accessibility.

Culot et al. (2022)
"Using supply chain databases in academic research: A methodological critique"
This article describes the main methodological implications of using Bloomberg SPLC, FactSet Supply Chain Relationships, and Mergent Supply Chain for academic purposes. It introduces key dimensions such as data fit, accuracy, representativeness, temporal consistency, and accessibility.

Liu (2020)
"Business and Finance Data Quality Problems: What Economic Researchers Need to Know"
Offers a comprehensive description of data quality problems in business and financial research, including issues of comparability, consistency, validation, and documentation across secondary datasets.

Van Schilt et al. (2024)
"Dimensions of data sparseness and their effect on supply chain visibility"
The study analyses how different forms of data sparseness affect the reliability and interpretability of supply chain visibility data. It highlights issues of completeness, consistency, and representativeness in supply chain visibility datasets.

Dong et al. (2024)
"Supply Chain Data and Analysis: The Case of Supply Chain Innovation"
This paper reviews challenges and opportunities in using supply chain data for innovation research. It addresses representativeness and selection bias in supply chain data and emphasizes aligning dataset construction with research questions to ensure generalisability.

Lee and Yoon (2025)
"The Development of a Methodology for Assessing Data Value Through the Identification of Key Determinants"
This paper describes a methodology for assessing data value by identifying the key determinants that influence it. It discusses timeliness and temporal consistency as determinants of data value, particularly in dynamic analytical settings.

Bans-Akutey and Tiimub (2021)
"Triangulation in Research"
This work explains how triangulation across data sources, methods, or perspectives enhances research validity and credibility.

Ellram and Tate (2016)
"The use of secondary data in purchasing and supply management (P/SM) research"
Examines the use of secondary data in purchasing and supply management research, highlighting data fit, transparency, and validation challenges.

Cook et al. (2012)
"An issue of trust: Are commercial databases really reliable?"
This article critically evaluates the trustworthiness of commercial databases used in business research, reinforcing the importance of transparency, documentation, and external validation.

Fetzer et al. (2024)
"AI-generated production networks: Measurement and applications to global trade"
This paper describes the methods used to generate the AIPNET dataset and specifically discusses accuracy, validation, and interpretability.

In the next phase, the most relevant and best usable quality assessment criteria are selected to form the foundation for the Data Quality Assessment Framework.

Chapter 5. Initial Data Quality Assessment Framework

This section translates the insights from the literature review into an initial Data Quality Assessment Framework. It applies the fitness-for-use perspective (Wang and Strong, 1996), which holds that data quality is not an absolute property but depends on whether data is suitable for its intended analytical purpose. Accordingly, each criterion is assessed in relation to the research question or decision context for which the dataset is used. The same criterion may therefore be more or less critical depending on the mapping method, scope, resolution and data sources of the dataset.
The sections that follow present the nine diagnostic criteria identified, describe their pass conditions and importance in different contexts, and provide the Step-by-Step Guide to support its practical application.

5.1 Diagnostic criteria

The literature review reveals several recurring diagnostic criteria that determine the reliability and usefulness of datasets. For the Data Quality Assessment Framework, the following nine criteria have been selected:

1. Data fit
Data fit refers to the extent to which the dataset's structure, coverage, and granularity are aligned with the intended analytical or policy goal. This criterion operationalises the broader fitness-for-use principle by testing whether the dataset is appropriate for its specific purpose. In supply chain research, Culot et al. (2022) note that one must assess the fit between the data, the supply chain constructs, and the research questions to avoid drawing invalid conclusions.
Example: suppose policymakers want to analyze multi-tier supplier risks for critical products. A dataset that only contains immediate supplier links would be unfit for this purpose, because it lacks the depth (tiers) needed to map cascading risks.

2. Accuracy
Accuracy is the degree to which the recorded information truthfully represents real-world supply-chain relationships. It focuses on the correctness of existing data entries, whether supplier links, firm identifiers, and classifications accurately describe actual relationships. Culot et al. (2022) highlight "errors caused by the databases' classifications and assumptions" as a key issue, referring to data accuracy problems.
Example: a supply chain mapping dataset for a major automaker lists companies as suppliers that, in reality, never supplied that automaker. Such inaccuracies could obviously lead the automaker to the wrong conclusions.

3. Completeness
Completeness measures the extent to which the dataset contains all relevant entities, tiers, and attributes. Missing records, unreported supplier links, or omitted small firms can distort analyses and bias results (Van Schilt et al., 2024).
Example: consider a global electronics manufacturer trying to map its supply chain. An incomplete dataset might only list that manufacturer's top 5 suppliers (perhaps derived from a few public disclosures), omitting smaller or second-tier suppliers. As a result, any analysis of supply risk or

diversity would be skewed.

4. Consistency
Consistency refers to the uniform application of definitions, classifications, and identifiers within the dataset and the alignment of those with other sources (Van Schilt et al., 2024). Inconsistent terminology, duplicate entries, or varying aggregation rules reduce comparability.
Example: if a company appears under slightly different names or codes in different parts of the dataset, it might be treated as two different suppliers incorrectly. In supply chain mapping, consistency issues often arise with firm identifiers (e.g. one record uses a legal name, another uses a common name for the same entity) or with classification of industries and locations. Such inconsistencies make it hard to merge or compare datasets and can lead to double counting or omissions.

5. Timeliness and temporal consistency
Timeliness ensures data is up-to-date and reflects the current state of supply chain relationships (Lee and Yoon, 2025). In fast-moving supply chains, outdated data (even by months) can mislead decisions, particularly in real-time operational use.
When the focus is on analyzing historical data, temporal consistency is important. This means the data is recorded in a time-consistent manner: reporting periods are aligned, changes over time are tracked without unexplained gaps, and historical data is not retroactively altered without documentation. Culot et al. (2022) observe, for instance, that many commercial supply chain databases had methodological adjustments over the years in sources, languages, industry coverage, etc., which complicates longitudinal analysis.
Example of timeliness: in a low-quality dataset, a supplier that has already been replaced still appears in the data.
Example of temporal consistency: imagine tracking a company's supplier count over a 5-year period using a supply chain database. In a poor-quality scenario, the database changed its inclusion criteria in year 3 (say, it started including smaller suppliers or added new industries). If this change is not transparent, an analyst might see a spike in year 3's supplier count and conclude the company dramatically expanded its supply base.

6. Representativeness
Representativeness measures the extent to which the dataset captures the full and balanced population of firms, sectors, regions, and supply-chain tiers relevant to the analysis. From a scientific standpoint, representativeness is related to sampling bias: if the data sample is not representative of the population, conclusions may not generalize. Recent studies (such as Dong et al. (2024)) highlight issues like selection bias in supply chain data and emphasize aligning data choices with research questions to improve generalizability.
Example: in a mapping of the global textile supply chain, a non-representative (biased) dataset might only include suppliers that are in developed countries or only those suppliers that voluntarily report to a certain platform. Such a sample would not be representative of the real situation.

7. Transparency and documentation
Transparency refers to how openly the data source discloses its collection methods, definitions, and any cleaning or verification procedures. Documentation is the detailed information about how data was gathered, what each field means, and how often it is updated. Transparency and documentation

are crucial for evaluating whether a dataset can be trusted, replicated, and correctly interpreted (Culot et al., 2022).
Example: a particular supply chain mapping platform provides a public dataset of supplier relationships but gives no information on how it collects these links (are they from surveys, from purchase orders, from web scraping news?). It also does not explain its company matching process or how often data is refreshed. This lack of transparency/documentation is risky: if an error or odd pattern appears, users cannot tell if it is a real supply chain insight or a data error.

8. Accessibility and comparability
This criterion addresses whether the data is easily obtainable and usable, and whether it can be readily compared or integrated with other data sources. Accessibility refers to the extent to which data can be accessed by those who need it, considering technical format, cost, and access rights. Data that cannot be accessed or practically used has limited value regardless of its intrinsic accuracy or completeness (Wang and Strong, 1996). Expensive subscriptions or delivery in highly specialized software environments may therefore limit both the dataset's impact and the level of external scrutiny it receives. Comparability concerns the alignment of definitions, classifications, and formats across datasets, enabling meaningful comparison or integration. A lack of comparability complicates data integration and increases the risk of mismatches or biased results. Limited comparability is a recurring challenge in business and supply chain research and poses a barrier to replication and cumulative knowledge building (Liu, 2020).
Example of accessibility: a dataset of supplier relationships that is only available through a paid subscription costing thousands of dollars and delivered in a very specialized software format would be a quality concern, as ideally data should be as accessible as possible (without compromising necessary confidentiality).
Example of comparability: suppose one supply chain dataset identifies companies by their full legal names, while another dataset uses standardized business identifiers. If you try to combine them, you face a big reconciliation task, potentially matching names one by one and risking mismatches. This is an example of poor comparability.

9. Validation and verification
Validation and verification refer to the processes that confirm the dataset's accuracy and reliability, both through internal checks and external comparisons. Essentially, this criterion asks: has the data been cross-checked and quality-assured from multiple angles? Scientific arguments for this criterion come from the principle of triangulation: by comparing data with independent sources, one can significantly increase confidence in its correctness (Bans-Akutey and Tiimub, 2021).
Example: imagine a supply chain mapping dataset that lists Supplier X as providing components to Manufacturer Y. If this dataset is unverified, it is possible this link came from a single press article or a one-time survey response. In a high-quality scenario, the data provider or user would verify this relationship. For instance, they might check Manufacturer Y's annual report or Supplier X's customer list (if available), or see if trade data shows commerce between the two. If no independent confirmation is found, the link could be flagged as low confidence or removed.

There are some additional criteria that are often recognised in general data quality literature, but that have not been included in the nine criteria for the Data Quality Assessment Framework. The main ones are:

- Uniqueness (lack of duplicates), which demands that each real-world entity or relationship is only recorded once. This prevents double-counting and confusion. Uniqueness is recognised in data quality literature as a core criterion for avoiding redundancy and ensuring that each entity is represented only once, but consistency largely covers this. Ensuring consistent identifiers inherently prevents unintentional duplicates. Liu (2020) discusses duplicate records, redundancy, and inconsistent identifiers as data quality problems in secondary datasets, and treats them as symptoms of broader consistency and integration issues, not as isolated quality dimensions.

- Integrity (referential integrity), which ensures that relationships between data elements are valid. Referential integrity implies that if one entity references another, both must exist and be logically linked. In a supply chain network, referential integrity could mean that if Company A is listed as supplying Company B, both A and B are present in the dataset's firm list. In the supply chain mapping framework this aspect is covered under consistency and completeness. Culot et al. (2022) discuss logical validity of relationships, entity matching, and errors arising from missing or inconsistent firm identifiers in commercial supply chain databases. They treat these issues as consequences of incomplete firm lists and inconsistent identifiers and classifications, which maps directly onto completeness and consistency.

- Cost and Effort to Use: even though cost is a criterion for its quality in real-world terms, it is outside the strict definition of data quality. The practicality of using the data is related to accessibility. Culot et al. (2022) clearly distinguish between intrinsic data quality dimensions and practical constraints related to data usage, such as licensing costs and access restrictions.

To ensure broader applicability, the diagnostic criteria are designed to remain flexible rather than prescriptive. Because each dataset differs in construction and purpose, the criteria must be interpreted in light of the dataset's intended use and characteristics discussed in Chapter 2. These characteristics, such as mapping method, scope, and resolution, influence how the fitness-for-use principle is applied and what constitutes acceptable quality.
The framework therefore allows users to define their own "pass conditions" for the quality criteria, enabling adaptation to different dataset types and analytical purposes. In addition, users assign relative importance to the criteria, reflecting their relevance for the specific use case. To support consistent application, the framework also provides guidance on how such conditions can be formulated and how relative importance can be assigned in a transparent and systematic manner.

5.2 Pass conditions and importance

To apply the framework in practice, each diagnostic criterion must be assessed against clear pass conditions and assigned a relative importance based on the dataset's intended use and characteristics. Pass conditions specify what qualifies as acceptable quality. After assessing the criteria we assign one of the following decisions to the criteria:

- Passed: sufficient quality

- Not passed: insufficient quality

- Unknown quality: when information is missing

Importance reflects the weight each criterion carries in the overall assessment. Together, these elements operationalise the fitness-for-use principle by translating conceptual quality criteria into measurable evaluation standards.

Formulating and documenting pass conditions and importance
Pass conditions describe, in clear terms, what constitutes acceptable quality for each criterion within the specific analytical or policy context. They can be described either quantitatively or qualitatively.
Quantitative descriptions may include, for example, minimum coverage levels, update frequencies, or accuracy rates that indicate whether the dataset meets the expected standard.
Qualitative descriptions may refer to aspects such as the availability of documentation on data collection methods, the clarity of definitions and classifications, or the presence of validation procedures.
While quantitative specifications are generally preferred because they allow for more objective comparison and replication, qualitative descriptions remain appropriate when numerical information is unavailable, for simple yes/no conditions, or when evaluative judgement is required.
This flexible approach aims not to impose universal criteria, but to make explicit how "acceptable quality" is defined in the specific context.
The relative importance of criteria is written down using the following five-level verbal scale:

- Required
- Very important
- Important
- Somewhat important
- Not required

This format ensures clarity while avoiding false precision and enables consistent interpretation across users. The scale can be translated into numeric weights (5-1) if a numerical comparative analysis is desired. However, this study does not recommend such conversion, as it may oversimplify the qualitative nuances that the verbal scale is intended to preserve.
Each dataset characteristic influences the definition of pass conditions and the relative importance of criteria in its own way. Because the analytical contexts in which datasets are used vary greatly, it is not possible to identify a single criterion per characteristic that is universally most important. However, it is possible to provide guidance on how certain criteria become more or less critical depending on specific contexts, and to explain the underlying reasons for these differences. The following subsection outlines how the criteria importance changes in different contexts.

Criteria importance
Before presenting a schematic overview, this subsection first discusses the relative importance of each criterion. The narrative explanation below clarifies how the relevance of each criterion varies across different contexts. After this textual discussion, a summarising table is provided that captures the conceptual extremes of all nine criteria in a concise and systematic format.

1. Data fit
Data fit is most important when the dataset is used for targeted analytical or policy purposes, as the structure and level of detail must align with the research goal. It becomes less critical when data is explored in a more general or descriptive way, where precision in structure is not essential.

2. Accuracy
Accuracy carries the most weight for empirical datasets, which rely on factual information to represent supply-chain relationships. In AI-enhanced or AI-generated datasets, its importance depends on how strongly the data is grounded in observed reality, making it slightly less decisive for overall quality.

3. Completeness
Completeness is particularly important when datasets aim to provide a full overview of supply-chain networks or when quantitative measures depend on coverage. It is less crucial when the dataset serves exploratory or illustrative purposes, where partial information can still be informative.

4. Consistency
Consistency is key for ensuring that variables, classifications, and identifiers are applied uniformly within the dataset. It is especially relevant when information is collected from multiple sources or compiled over several stages. Minor inconsistencies are less problematic when the dataset is used for descriptive or one-time analysis.

5. Timeliness and temporal consistency
Timeliness is most important when analyzing volatile, rapidly changing phenomena, where it is critical that the dataset is up to date. Temporal consistency is critical for longitudinal analyses, as stable definitions and measurement methods are required to ensure comparability over time.

6. Representativeness
Representativeness is most relevant when findings are intended to reflect broader sectors, countries, or populations. For narrowly defined or case-specific analyses, it is of lower importance, as detailed accuracy within the selected boundary is the main concern.

7. Transparency and documentation
Transparency is crucial to understanding how data were obtained and processed. It is especially important in AI-enhanced or AI-generated datasets, where users must be able to interpret how relationships were inferred and validated.

8. Accessibility and comparability
Accessibility becomes more critical when the project constraints are very tight, e.g. when budgets, expertise, the available tools, time, etc. are limited. Comparability is most relevant when datasets are used across studies, sectors, or organisations. It is less critical for internal or single-purpose datasets, where external replication is not required.

9. Validation and verification
Validation gains importance when datasets include inferred or automatically collected information, as it confirms that the results are reliable. For empirical data, validation is often more straightforward but remains important to ensure credibility.

To provide a concise overview of how each criterion conceptually ranges between two extremes, Table 2 summarises the findings above.

Criterion	**Extreme 1: When the criterion becomes highly important**	**Extreme 2: When the criterion becomes less important**
1) Data fit	Targeted analytical or policy use	Explorative or descriptive use
2) Accuracy	Empirically collected, fact-based datasets	AI-inferred datasets
3) Completeness	Full network coverage required	Exploratory or illustrative purposes
4) Consistency	Multi-source or longitudinal analysis	One-time or descriptive analysis
5) Timeliness & Temporal consistency	Timeliness: volatile, dynamic data Temporal consistency: longer time period, more or varying data sources	Timeliness: stable data Temporal consistency: shorter time period, fewer or consistent data sources
6) Representativeness	Broad scope (generalisable)	Narrow scope (case-specific)
7) Transparency & Documentation	To interpret modelling assumptions (e.g. AI inferred datasets)	Internal or exploratory use
8) Accessibility & Comparability	Accessibility: limited budget, expertise, tools available Comparability: cross-study use or data integration	Accessibility: ample availability of budget, expertise, tools Comparability: internal or single-purpose use
9) Validation & Verification	AI-inferred datasets	Empirically collected datasets

Table 2: Conceptual extremes of the nine diagnostic criteria

In summary, the context determines how data quality should be interpreted and prioritised. The mapping method shifts the focus from factual accuracy to methodological transparency, while scope and resolution balance representativeness against precision. The data sources influence what level of validation and documentation is feasible. Together, these characteristics guide how the framework should be applied in practice and ensure that data quality is always assessed in relation to the dataset's intended use.

5.3 Step-by-Step Guide

To support its practical application, the framework includes a clear step-by-step guide.

Step 1: Define the use
Clarify the analytical or policy objective of the study.
As the framework is based on the fitness-for-use principle, the research goal helps determine which criteria are most relevant and under what conditions they should be applied.

Step 2: Specify the dataset characteristics
Identify the key characteristics of the dataset.
In addition to the research goal, the type of dataset also helps to determine which criteria are most relevant and under what conditions they should be applied. We ask the user to specify the mapping method, scope, resolution and the data sources.

Step 3: Define pass conditions
For each diagnostic criterion, determine what constitutes acceptable quality in the given context. Specify when a criterion is considered sufficient to "pass". And assign the weight each criterion carries in the overall assessment: Required, Very important, Important, Somewhat important, or Not required.

Step 4: Assess the criteria
Evaluate the dataset against each diagnostic criterion. Report the results and assign one of the following outcomes: passed, not passed, or unknown (when information is insufficient).

Step 5: Interpret and conclude
Combine the individual results and their relative importance to derive an overall evaluation. Summarise the findings and provide a conclusion or recommendation on the dataset's suitability for the defined use.

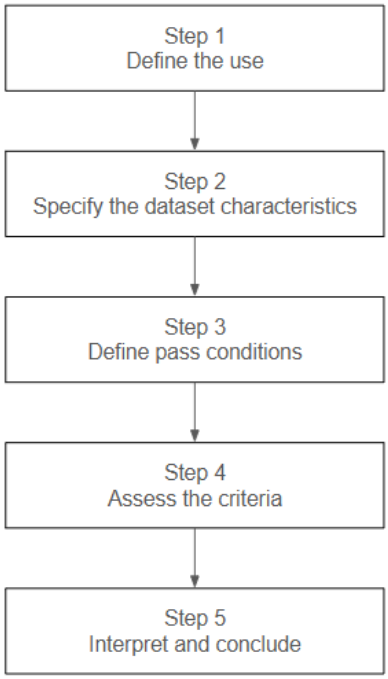


Figure 1: Overview of the Step-by-Step Guide of the initial Data Quality Assessment Framework

To facilitate a structured format to document these steps, an evaluation form is developed and a compromised example is shown in Appendix A.

In the following chapters, the proposed framework is applied in a focused case study to illustrate its use in practice and subsequently evaluated through expert validation to assess its practical usability and identify potential ambiguities or limitations.

Chapter 6. Case study

This chapter presents the case study selection and describes the step-by-step application of the framework in the case study.

6.1 Case study selection

The selection of a suitable case follows several practical and methodological considerations. The aim is to identify an analytical context, also referred to as a Use Case, and a Dataset that allows for an in-depth yet feasible examination of data quality within the scope of this thesis.

6.1.1 Use case: Energy capital (wind turbines)

To apply the Data Quality Assessment Framework, we have defined a use case around one of the topics TNO is currently working on, and further described and contextualised it using additional information from recent reports and news articles. The resulting use case is not an actually used real-world case, but it is strongly based on a real topic and closely reflects the type of situations in which supply chain mapping datasets are typically applied.
The use case selected for this thesis focuses on the supply chain of wind turbine manufacturing, a technology that plays a central role in the European energy transition. The energy transition refers to the systemic shift from fossil-fuel-based energy systems to low-carbon and renewable energy sources, driven by climate objectives, industrial competitiveness and long-term energy security (International Energy Agency, 2021).
Wind energy is identified by the European Commission as one of the strategic net zero industries necessary for meeting climate and industrial objectives (European Commission, 2023b). Recent analyses also highlight that wind deployment faces supply chain challenges, including delays and increased equipment costs, which contribute to broader supply chain vulnerabilities (International Energy Agency, 2022). In particular, several key components in modern turbines, such as permanent magnet generators, rely on materials that are classified as critical for the European economy, including neodymium and dysprosium (European Commission Joint Research Centre, 2020).
Because of these challenges, supply chain analysts and policymakers frequently seek insight into the structure of wind turbine production. To illustrate the type of analytical questions for which datasets such as AIPNET are used, the following example question is introduced.

Illustrative question:

> *Which product categories serve as direct inputs in the production of wind turbines?*

In practice, policy-relevant analyses often aim to capture complex, multi-tier production structures, including indirect inputs, upstream dependencies, and cross-border linkages. Such analyses typically go beyond direct input relationships and require detailed modelling of production networks across multiple stages. However, addressing these broader questions falls outside the scope of this illustrative case. For the purpose of demonstrating the Data Quality Assessment Framework, a simplified and narrowly defined question is sufficient and appropriate.

6.1.2 Dataset: AI-Generated Production Network (AIPNET)

The dataset selection process starts with a broad review of datasets that exist in the field of supply chain mapping and are commonly used in research. This results in an overview of the relevant datasets. Next, the identified datasets are examined to determine whether they are appropriate candidates for this study.
First, the dataset has to be accessible for research purposes. Many commercially used datasets require separate, often paid, licences that are not available through the university or TNO. As a result, only datasets for which access can be arranged within the timeframe of the project are considered, for example because they are covered by existing TNO licences or are openly available for research.
Second, feasibility and analytical depth are important considerations. The dataset needs to be manageable in size and structure, yet contain sufficient detail to apply the Data Quality Assessment Framework and provide meaningful insights.
Based on these requirements, several datasets were excluded and a final choice had to be made between the AI-Generated Production Network (AIPNET) (Fetzer et al., 2024) and the FactSet Supply Chain Relationships dataset (FactSet, 2025). AIPNET was selected as the dataset for this study because applying the Data Quality Assessment Framework to this dataset is less straightforward than for more conventional supply chain datasets. The inferred nature of the relationships and the reliance on modelling assumptions introduce additional ambiguity when evaluating several data quality criteria. This makes AIPNET a suitable case for illustrating how the framework can be applied in practice and for identifying limitations and interpretational challenges that may arise when assessing emerging types of supply chain mapping datasets.
AIPNET is an AI-generated dataset that represents product-to-product relationships. AIPNET is constructed using a large language model that infers production linkages between products based on HS codes and their textual descriptions, rather than observed trade flows or firm data. As such, it provides a representative example of a model-based, AI-generated dataset used to study production structures. For this study, the dataset is analysed in a tabular (Excel) format, where spreadsheet-based filtering and selection functions are used to examine relevant relationships. A detailed description of the dataset is included in Appendix B.

6.2 Case study analysis

This section applies the initial Data Quality Assessment Framework to the previously described case. The results of the Step-by-Step Guide of the framework are documented in the same format as the Data Quality Assessment Form shown in Appendix A.

Step 1: Define the use

Goal:
To answer the following research question: “Which product categories serve as direct inputs in the production of wind turbines?”

Policymakers need to analyze the supply chains in wind turbine manufacturing to identify dependencies and vulnerabilities in order to support strategic policy decisions.

Note: as described in Section 6.1 this is a simplified example of a real-world use case.

Step 2: Specify the dataset characteristics

Mapping method:
AI-generated dataset

AIPNET is constructed using large language models to infer production relationships between products based on their Harmonized System (HS) descriptions.

Scope:
Global and multi-industry coverage

The dataset covers the full universe of internationally traded goods defined in the Harmonized System. It is not limited to specific countries, firms, or industries, but represents global production structures across all major sectors of the economy.

Resolution:
Product-to-product relationships

The dataset maps relationships at the product level, with each observation identifying a directed upstream-downstream relationship between two product categories classified under the Harmonized System. AIPNET is provided at both the HS 6-digit level and an aggregated HS 4-digit level. The dataset does not include firm-level information and does not identify individual companies involved in production relationships.

Data sources:
Secondary data

The dataset is constructed using secondary data sources. The primary inputs consist of official HS product codes and their textual descriptions, which serve as prompts for the language model. Additional secondary sources, including national input-output tables and trade statistics, are used for validation and pruning of inferred relationships, but not for the direct construction of individual links.

Step 3: Define pass conditions
The pass conditions and importance of each criterion are first defined in the text and subsequently documented in Table 3. Note that the proposed pass conditions are only meant to illustrate the principle and are formulated based on assumptions about the selected case and its analytical context. In a real-life application, the conditions depend on the demands of the specific project. For example, a lower accuracy may be acceptable for a quick scan, or for a relative comparison, but a higher accuracy would be needed for a more precise analysis.

1. Data fit
Pass conditions: The dataset must contain the relevant first-tier input-output links needed to answer the question. In practice, this means AIPNET should include the wind turbine product category as an output, and list its immediate input product categories. All major components or materials that

go directly into manufacturing a wind turbine should appear as linked input categories and the data should be detailed enough to distinguish key component types. Additionally, the scope must be global: inputs should be captured regardless of their country of origin. In summary, AIPNET passes this criterion if it explicitly represents "wind turbines" and maps out the direct input categories with sufficient clarity to single out each key component class on a global basis.
Importance: Required - In Table 2 we see that the criterion becomes highly important when the dataset is for policy use. If the data does not contain the wind turbine production inputs at the appropriate level (e.g. wrong tier, missing context), the core question cannot be answered.

2. Accuracy
Pass conditions: The product-to-product relationships in AIPNET should truthfully reflect real-world input linkages to a reasonable degree. The listed direct inputs of wind turbine production should actually represent the product categories. As the number of nodes in our case study is limited and the data is easily readable, we can manually evaluate the accuracy. We demand a high level of accuracy, minimum of 95%, due to the risk of missing relevant information. For more extensive studies, expressing the required accuracy in a numeric score or percentage of a sample instead of the whole dataset may be necessary.
Importance: Important - Representing the real-world product categories is of clear relevance, although it is secondary to fitness-for-use. As long as the dataset supports the intended analysis, some degree of inaccuracy is accepted. This is in line with Table 2, which shows that accuracy is usually less important for AI-inferred datasets.

3. Completeness
Pass conditions: The dataset should capture all or nearly all relevant direct input product categories for wind turbine production. In other words, for a wind turbine, the important component categories are included. The completeness for the relevant nodes should ideally be 100%. If it is less, the missing data should have minimal impact on the outcome of the analysis (fitness-for-use principle). What constitutes minimal impact in this context is determined by the expert judgement of the assessor.
Importance: Important - The analysis focuses on specific branches of the supply network. For these branches, completeness is important, but full network coverage is not required. As Table 2 shows, the completeness criterion only becomes highly important when full network coverage is required, which is not the case here.

4. Consistency
Pass conditions: Data entries must be internally consistent in definitions and identifiers. The same product category should not appear under multiple names or codes. A pass means the number of classification errors in the tested sample is zero or minimal.
Importance: Important - AI and trade information are integrated, so we need consistent usage of HS codes and relationship definitions to confidently trust and interpret the results. Without consistent naming and classification, aggregating or interpreting the supplier network could introduce errors.

5. Timeliness and temporal consistency
Pass conditions: The objective of the analysis is to identify which product categories structurally serve as direct inputs into wind turbine production. This does not change much over time (it is

much more stable than, say, a year-on-year trade analysis). But using very old data could lead to the wrong conclusions and wind turbine technology progresses continuously, so we specify that the data may not be older than five (5) years.
Importance: Somewhat important - This is what Table 2 specifies for a single point-in-time analysis. Temporal consistency is not required, because the case study does not analyze changes over time.

6. Representativeness
Pass conditions: The analysis aims to identify which product categories structurally function as direct inputs for wind turbine production, rather than to make claims about the relative importance or prevalence of these inputs across regions or firms. This means we do not need to apply any specific pass conditions.
Importance: Not required.

7. Transparency and documentation
Pass conditions: The dataset should be accompanied by sufficient documentation to allow correct interpretation of the inferred product-to-product relationships. This includes clear descriptions of the data sources used, the AI-based mapping methodology, the approach used to validate inferred relationships, and the main limitations of the dataset. Transparency is achieved if users can understand how relationships are generated, how they are evaluated, and under which assumptions and constraints they should be interpreted.
Importance: Very important - Transparency and documentation are essential for the responsible use of AI-generated datasets. Since production relationships in AIPNET are inferred rather than directly observed, users must be able to assess how the mapping was constructed and validated in order to interpret the results correctly. According to Table 2, this makes transparency and documentation highly important.

8. Accessibility and comparability
Pass conditions: The dataset should be reasonably accessible to the research team and structured in a way that allows comparison with other relevant datasets. Accessibility requires that the data can be obtained and used in practice. Comparability requires that product categories are defined using standardized classifications, allowing alignment with external data sources such as trade statistics or input-output tables. The criterion is met if the dataset can be accessed and product identifiers are compatible with commonly used classification systems.
Importance: Somewhat important - This criterion is relatively less critical for the research question, because the analysis can be performed with the AIPNET dataset alone once it is obtained. We are not necessarily integrating multiple datasets or conducting cross-study comparisons in this specific use case. It supports the analysis (for instance, using HS codes does make it easier to validate with trade data or explain results to others in standard terms), but it is not a decisive factor in answering the core question.

9. Validation and verification
Pass conditions: A substantial share of inferred product-to-product relationships should be supported by external evidence. Given the AI-generated nature of the dataset, validation does not require transaction-level confirmation but should rely on systematic comparison with independent sources such as trade data, input-output tables, or established production knowledge. The criterion is met

if the supplier of the dataset has provided convincing documentation of the triangulation techniques applied.
Importance: Very important - For this use case, validation and verification are necessary to ensure that inferred input relationships are credible and suitable for interpretation. As indicated in Table 2, validation is particularly important for AI-generated datasets, where relationships are inferred rather than directly observed.

Criteria	**Pass conditions**	**Importance**
1) Data fit	→ Presence of wind turbine product category → Coverage of key component categories → Sufficient product-level granularity → Global scope	Required
2) Accuracy	→ Minimum of 95% accuracy	Important
3) Completeness	→ Ideally 100% completeness If <100%: missing data has minimal impact	Important
4) Consistency	→ Zero or minimal classification errors	Important
5) Timeliness and temporal consistency	→ Data not older than 5 years	Somewhat important
6) Representativeness		Not required
7) Transparency and documentation	→ Data sources documented → Mapping methodology described → Validation approach documented → Dataset limitations stated	Very important
8) Accessibility and comparability	→ Dataset is accessible to researchers → Product identifiers and dataset structure allow comparison with external datasets	Somewhat important
9) Validation and verification	→ Convincing documentation of triangulation techniques	Very important

Table 3: Pass conditions and importance case study

Step 4: Assess the criteria
In this step, the dataset is evaluated against the diagnostic criteria. For each criterion, the pass conditions specified in Step 3 are applied by executing the corresponding assessment steps. Table 4 documents this assessment process, including the evaluation steps for each subpart and the resulting interim decisions. Based on these interim results, overall assessment outcomes are derived for each criterion and summarized in Table 5. The detailed execution of the assessment steps and where needed underlying justifications are reported in Appendix C.

Criterion	**Subpart pass conditions**	**Assessment**	**Interim decision**
1) Data fit	Presence of the wind turbine product category	1. Determine HS code for wind turbine category 2. Check if dataset contains HS code	Passed
	Coverage of key component categories	1. Research and make list of key components 2. Determine HS codes for key components 3. Check if dataset contains HS codes	Passed
	Sufficient product-level granularity	1. Inspect HS level of product categories 2. Evaluate whether key component types can be distinguished	Passed
	Global scope	1. Review dataset methodology and documentation 2. Verify that product relationships are inferred independently of country or region	Passed
2) Accuracy	Minimum of 95% accuracy	1. List direct input relationships for wind turbine HS category 2. Manually verify whether each input represents a plausible real-world input 3. Count correct relationships 4. Check whether at least 95% of relationships are correct	Passed
3) Completeness	Ideally 100% completeness If <100%: missing data has minimal impact	1. Compile list of key direct input product categories for wind turbine production 2. Check whether each key category appears as a direct input in the dataset 3. Identify missing categories 4. Assess whether missing categories materially affect the analysis	Passed

Continued on next page

Criterion	**Subpart pass conditions**	**Assessment**	**Interim decision**
4) Consistency	Zero or minimal classification errors	1. Check whether each product category is mapped to a unique HS code 2. Verify that the same product does not appear under multiple codes or names 3. Identify any classification or definition inconsistencies	Passed
5) Timeliness and temporal consistency	Data not older than 5 years	Identify the construction and release period of the dataset	Passed
6) Representa-tiveness	-	-	-
7) Transparency and documentation	Data sources documented	Review dataset documentation and verify that the data sources are described	Passed
	Mapping methodology described	Review dataset documentation and verify that the mapping methodology is described	Passed
	Validation approach documented	Review dataset documentation and verify that the validation methods are described	Passed
	Dataset limitations stated	Review dataset documentation and verify that constraints related to inference, scope, or validation are stated	Passed
8) Accessibility and comparability	Dataset is accessible to researchers	1. Verify that the dataset is available to the research team 2. Confirm that the data can be accessed and used in practice	Passed
	Product identifiers and dataset structure allow comparison with external datasets	1. Check whether product categories are defined using standardized identifiers (HS codes) 2. Verify that the dataset structure enables alignment with external datasets (with similar dataset characteristics)	Passed

Continued on next page

Criterion	**Subpart pass conditions**	**Assessment**	**Interim decision**
9) Validation and verification	Convincing documentation of triangulation techniques	1. Review dataset documentation describing applied triangulation techniques 2. Verify that validation methods used are convincing and clearly explained	Passed

Table 4: Assessment procedure and interim results per criterion

The interim decisions reported in Table 4 are aggregated to obtain an overall assessment outcome for each criterion. These final outcomes and decisions are summarized in Table 5 as the user would do in the Data Quality Assessment Form.

Criteria	Result	Decision
1) Data fit	The wind turbine product category and its key component categories are identifiable at the HS level and are present in the dataset. The dataset provides sufficient product-level granularity to analyze wind turbine-related production linkages and has a global scope.	Passed
2) Accuracy	Direct upstream input relationships for the wind turbine product category are identified and manually verified for technological plausibility using HS headings. All identified upstream product categories represent plausible real-world inputs, resulting in an accuracy rate of 100% which is above the 95% threshold.	Passed
3) Completeness	All key direct input product categories required for the production of wind turbines are identified at the 6-digit HS level and are present as direct upstream inputs to HS 850231 in the dataset. Each key input category is represented by at least one corresponding HS code among the identified upstream inputs, resulting in 100% completeness.	Passed
4) Consistency	Product categories are consistently defined using the HS 2002 6-digit classification, ensuring unique and standardized product definitions. Repetition in short product descriptions does not indicate misclassification, as each category is uniquely identified by its HS code. A small number of HS codes (7) appear in the edge list but not in the node list, affecting 493 out of 429,871 links (approximately 0.115%). These inconsistencies are limited in scope and do not indicate systematic classification errors.	Passed
5) Timeliness and temporal consistency	The dataset was constructed in 2024 and meets the five-year timeliness requirement. The use of HS 2002 codes as a stable product taxonomy does not affect this assessment.	Passed
6) Representa-tiveness	-	-
7) Transparency and documentation	The AIPNET dataset is well documented. Data sources, mapping methodology, validation procedures, and key limitations are clearly described in the accompanying paper.	Passed
8) Accessibility and comparability	The AIPNET dataset is accessible and can be used in practice without technical or licensing barriers. The use of standardized HS product codes and a clear product-level network structure allows the dataset to be mapped and aligned with external relevant datasets, such as trade statistics and Input-Output tables.	Passed
9) Validation and verification	The dataset documentation provides clear and detailed descriptions of multiple triangulation and validation techniques. These methods collectively support the credibility of the production network at the aggregate and structural level.	Passed

Table 5: Final assessment outcomes for the case study

Step 5: Interpret and conclude

Conclusion:
Suitable

The assessment results indicate that the AIPNET dataset is suitable for the intended use: to identify direct input product categories in wind turbine production within the defined scope of this case study. The dataset provides a clear and interpretable set of upstream product categories linked to the wind turbine HS code, and these categories align well with established knowledge of turbine production structures.
A wind turbine is a complex product that may involve a large number of distinct inputs. Nevertheless, within the scope of this assessment and under the adopted definition of completeness, no key direct upstream input categories were found to be missing. Furthermore, no critically implausible production relationships were identified in the examined sample. All identified direct inputs correspond to electrical components, mechanical parts, control systems, or material inputs that are consistent with established wind turbine manufacturing processes.
Some components used in wind turbine production may not be represented as separate product categories with distinct HS codes. This limitation is inherent to the structure and granularity of the Harmonized System classification rather than to the AIPNET dataset itself.
Finally, minor internal inconsistencies are identified at the network level, including a small number of missing nodes in the product list. These issues affect only a very small share of links and do not materially influence the wind turbine input structure examined in this case, but they underline the importance of basic preprocessing and consistency checks when applying the dataset in practice.

6.3 Case study evaluation

In general, there were no major procedural obstacles in applying the Data Quality Assessment Framework to the AIPNET dataset in the case study.
Implementing Steps 1 (define the use) and 2 (specify the dataset characteristics) was straightforward. There were no ambiguities and the framework provided clear guidance.
Step 3 (define the pass conditions) required more careful consideration. Table 2 was very helpful in assigning a relative importance to each quality criterion, making it easy to execute this step. But formulating explicit pass conditions was somewhat more challenging. Translating the quality criteria into concrete and testable conditions required interpretative choices, which were not easy to make based on the limited information available for the use case. The framework recommends the use of quantitative and objectively testable pass conditions where possible. The case study demonstrated that the justification for the choices made may often be equally valuable as the exact threshold value chosen.
A specific challenge in Step 3 was the perceived ambiguity between the validation and accuracy criteria. The validation criterion concerns the assessment of how the dataset has been validated by its creators, whereas the accuracy criterion involves evaluating the correctness of the data. Although this distinction is described in the framework documentation, the conceptual overlap between the two criteria can make it less immediately apparent during application.
Step 4 (assess the criteria) could largely be executed as the framework prescribes. There were some criteria, such as completeness and representativeness, that were more difficult to assess in a precise

manner than others. But a combination of available evidence in the dataset and reasoned judgment helped assess these criteria. What proved really helpful was that full access to the dataset was available in the case study.
Step 5 (interpret and conclude) involved consolidating the results of the preceding steps into an overall assessment. This step could be completed without any additional methodological challenges. Overall, the framework proved to be usable and no missing steps or other serious issues were identified during the case study.

Chapter 7. Expert validation

This chapter presents the expert validation of the proposed framework. Through expert interviews, feedback is collected on the clarity, relevance, and practical feasibility of the framework and its underlying concepts.

7.1 Expert selection

The pool of consulted experts consists of five employees from TNO. All experts have experience in the supply chain and/or data science domain, while differing in their roles, professional backgrounds, and years of experience. The experts were selected using purposive sampling, as the study requires informed input from individuals with specific domain expertise rather than statistical representativeness. The selection was guided by theoretical considerations regarding which perspectives are most relevant for validating the Data Quality Assessment Framework. Within this purposively defined group, variation was sought to capture complementary viewpoints on technical data handling, practical dataset use, analytical decision-making, and policy relevance. An anonymised overview of the experts and the rationale for their selection is provided in Table 6.

Identifier	**Role**	**Description**	**Selection rationale**
A	Consultant / data-analyst	Consultant and data analyst at TNO with experience in data-driven supply chain and international trade analyses.	Selected for their practical perspective on applied data analysis in supply chain and trade-related projects, contributing insight into the practical use and relevance of data quality assessments in policy-oriented contexts.
B	Team lead Data Innovation team	Team lead of a Data Innovation team at TNO, responsible for developing AI- and data-driven analyses for materials and supply chain projects across multiple TNO units.	Selected for their combined expertise as a practicing data scientist and team lead, with direct involvement in developing data infrastructures and AI-based analyses for supply chain applications. This dual perspective supports validation of the framework from both a technical data science standpoint and an implementation perspective in collaborative, multi-source projects.

Continued on next page

Identifier	Role	Description	Selection rationale
C	Data Scientist Data Innovation team	Data scientist at TNO within the Data Innovation team. Works on projects with a focus on supply chain analysis using datasets such as FactSet, recycling data, and SNP data, as well as on data visualisation and data enrichment.	Selected for their hands-on experience with supply chain datasets in practice. Involved in the analysis and evaluation of datasets used within projects, including datasets considered for continued use or renewal. Their perspective supports an assessment of the framework from a practical, user-oriented viewpoint.
D	Economist	Economist at TNO with a background in economics, working on innovation and R&D policy and increasingly focusing on data driven analyses of supply chain disruptions, critical materials, and policy relevant market dynamics.	Selected for their experience with applied economic analyses, contributing to an evaluation of whether the framework is practically feasible and usable within real-world economic and policy research settings.
E	Data Analyst Supply Chains	Data analyst at TNO with a multidisciplinary background in econometrics, policy analysis, and data science, working on the analysis of supply chains of strategic raw materials and technologies.	Selected for their hands-on experience with supply chain datasets and analytical methods, enabling assessment of the relevance, clarity, and practical feasibility of the data quality framework in real-world supply chain analysis.

Table 6: Anonymized overview of experts

7.2 Expert interviews

This section analyses the conducted expert interviews. The focus is on the substantive feedback provided by each expert, rather than on a quantitative comparison of responses across experts. The transcripts per expert are included in Appendix F. The interviews are interpreted and discussed per segment of the interview. Where relevant, interview quotes originally expressed in Dutch have been translated into English.

Dataset characteristics

The interviews indicate that the proposed dataset characteristics closely reflect how datasets are characterised in practice. Experts considered the set of characteristics to provide a comprehensive overview and recognised them as aligning with practical approaches to describing and assessing

datasets, even though such distinctions are often applied implicitly rather than explicitly articulated. Expert D and E indicated that timeliness and temporal aspects are important characteristics of datasets. In particular, D emphasised that datasets are strongly defined by their historical coverage and update frequency, as well as the extent to which data is available in near real-time. According to D, these aspects determine how datasets differ in practice and which types of analyses they can support. Similarly, E stressed that insufficient update frequency can cause datasets to lag behind the current state of supply chains, resulting in temporal bias. D noted that temporal properties could be seen as an aspect of the data resolution. E suggested that temporal properties could be seen as an aspect of the dataset scope, in a similar way as geographic coverage.

Diagnostic criteria

Across the interviews, experts indicated that the nine diagnostic criteria capture the main data quality challenges encountered in supply chain mapping and provide a sufficiently comprehensive overview. Expert D described the criteria as "a good and extensive overview" that can give a useful indication of dataset quality and usability. Similarly, expert A noted that the criteria form "a fairly complete picture," while expert B indicated that no additional criteria were missing, stating that they "could not think of a tenth".

Experts further emphasised that supply chain datasets are rarely ideal and typically contain limitations in several respects. Rather than expecting a dataset to meet every criterion, interviewees stressed the importance of being explicit about where the data falls short and what that implies for interpretation and use.

In terms of specific criteria, consistency was repeatedly highlighted as a recurring challenge in their work, for example because of changes in HS codes over time. Experts also pointed to accuracy, completeness, timeliness, transparency, and validation as important criteria. At the same time, several experts noted that the importance of individual criteria depends on the specific analytical context and purpose of the analysis.

Overall, the interviews indicate that the nine diagnostic criteria together provide a complete and coherent picture of the main data quality issues encountered in supply chain mapping.

Data Quality Assessment Framework application

All experts considered the five-step Data Quality Assessment Framework to be logically structured and, in principle, applicable in practice. In addition, several experts noted that the assessment form helps to make the evaluation process more structured and efficient. At the same time, they indicated that some steps are more straightforward to apply than others.

Defining the use upfront was widely seen as an essential and intuitive first step. At the same time, views differed on how clearly the use can be defined in advance. Expert B indicated that datasets are generally selected with a clear purpose in mind, even if this purpose may evolve over time. Expert A, by contrast, noted that research questions and data use can change as projects develop, which makes it difficult to fully define the intended use at the moment of assessment. Expert C similarly observed that evaluating a dataset for a single analysis is more straightforward than assessing its suitability across multiple or evolving applications. Identifying dataset characteristics was considered unproblematic, as experts indicated that this is something they already do implicitly in practice.

More discussion arose around the later steps of the framework. While most experts indicated that assigning importance levels was relatively straightforward and agreed that a qualitative, word-based

scale was appropriate, the formulation of pass conditions and the assessment of datasets against these conditions were identified as the most challenging elements of the framework. In particular, experts noted that certain criteria, such as completeness and representativeness, are inherently difficult to measure in practice.

Other experts were more positive about the feasibility of these steps. Expert D indicated that, when presented with the list of criteria, it is relatively easy to form an indication of when a dataset is "good enough". Expert B similarly found the formulation of pass conditions straightforward and noted that he immediately began thinking in terms of scores and importance levels for different datasets.

Overall, none of the experts considered the stepwise structure illogical or unclear. The main challenge lies in making and documenting judgments when defining pass conditions and assessing whether a dataset meets them for the intended use. Some experts proposed ways to further improve the framework's application.

Expert B suggested explicitly recording ownership of the assessment as part of the framework application. According to B, an assessment should be linked to a specific project, individual, or programme, so that it is clear who conducted the assessment, from which perspective, and for what purpose. This would allow others within the organisation to understand the context of the assessment and to consult the assessor as a domain expert on the dataset. B noted that such ownership could also support incremental improvement of assessments, as insights from different domain perspectives could be added over time, rather than requiring the assessment to be repeated from scratch.

Expert E proposed an addition to the framework involving a preliminary review of existing independent assessments in academic literature. E noted that it can provide useful insights into the quality and limitations of datasets and can support a more informed initial assessment. At the same time, E emphasised that such external assessments should be used as a complementary input rather than a substitute for use-case-specific evaluation, as dataset suitability ultimately depends on the specific use case and analytical purpose.

E also highlighted that defining absolute pass conditions can be difficult for certain criteria and suggested using benchmark datasets as a comparative reference. According to E, comparing a dataset against other available datasets can provide a more practical basis for assessment than setting fixed thresholds. E illustrated this by noting that for some pass conditions, such as country coverage, absolute thresholds can be defined, whereas for others, such as the number of supply chain relationships captured in a dataset, no clear reference point exists, making benchmarking against alternative datasets more practical.

Finally, expert D proposed a refinement to the application of the framework by explicitly incorporating an iterative element. D noted that no dataset fully satisfies all criteria and suggested adding a step in which the methodology is adapted to address criteria that are not met but remain important. This could involve supplementing the dataset with additional sources or introducing assumptions to account for known limitations. This reflects D's view that data quality assessment often involves adjusting the analytical approach rather than making a definitive accept or reject decision.

Practical usability of the Data Quality Assessment Framework

Experts were generally positive about the practical usability of the framework and described different ways in which they would use it in their work. Several experts emphasised that the

framework addresses a gap in current practice, as dataset selection and evaluation are often based on informal expert judgment rather than a consistent and comparable approach. For example, expert D indicated that the framework could be used to evaluate and compare datasets offered by different data providers. Expert B similarly noted that such a structured approach is currently missing and stated that he would "immediately use it on all sources" if the framework were available and ready to use.

At the same time, experts emphasised that data availability in supply chain analysis is often limited and that datasets are frequently used despite known shortcomings. Several experts noted that in such situations there is little room to reject a dataset outright. Expert A explained that analysts typically work with "the best possible option" rather than an ideal dataset, a situation also recognised by other experts. In these cases, the framework was described as useful for identifying and documenting dataset strengths and weaknesses, thereby supporting reflection on how these limitations affect the analysis and enabling an assessment of how different datasets may complement each other.

Expert A further noted that the framework can also be applied retrospectively, after an analysis has been completed, to reflect on the extent to which methodological choices were supported by the available data and to document and justify these choices.

While recognising the conceptual value of the framework, some experts noted limitations in its use in practice. Expert C noted that while the framework is helpful for evaluating a dataset for a specific use case, broader decisions in practice, such as whether to purchase a dataset, require considering multiple potential uses rather than a single application. In addition, expert E noted that the practical application of the framework is constrained by data accessibility, particularly for commercial supply chain datasets that require purchase before full evaluation is possible. According to E, demo sessions may offer a general impression but provide limited insight into suitability for specific use cases. E therefore identified limited access to data prior to purchase as a key challenge for applying the framework in practice.

7.3 Expert validation summary

Overall, the expert interviews indicate that the proposed framework is considered usable in practice and that the dataset characteristics and diagnostic criteria derived from the literature align well with how data quality is assessed in real-world supply chain analysis. At the same time, experts expressed differing views on how the framework would be applied in their own work and raised several points for consideration regarding its practical implementation. While no fundamental structural shortcomings were identified, certain aspects were viewed as challenging or subject to discussion in practice.

The key discussion points include:

- Timeliness and temporal aspects, such as update frequency, historical coverage, and near real-time availability, were identified as characteristics that influence dataset use in practice.
- The intended use of a dataset can be difficult to define upfront. The use may evolve over time and broader decisions, such as whether to purchase a dataset, may require considering multiple potential use cases rather than a single application.

- Explicitly recording ownership of the assessment was suggested, linking it to a specific project or individual to clarify perspective, context, and enable incremental improvement over time.
- Measuring pass conditions can be challenging, particularly for criteria such as completeness and representativeness.
- Benchmarking against other datasets was highlighted as a practical approach when absolute pass conditions are difficult to define.
- A preliminary review of independent academic assessments was suggested as a complementary input to support dataset evaluation.
- The inclusion of an iterative component in the framework was suggested, acknowledging that datasets rarely meet all criteria and may require methodological adjustments or complementary data.
- Data availability and quality is often limited in practice, meaning datasets are frequently used despite known shortcomings.
- Limited pre-purchase access to commercial datasets was identified as a key practical constraint when applying the framework.

These points reflect practical considerations raised by the experts and provide context for the application of the framework.

The results from the case study and the expert validation are presented and combined in the next chapter to refine the initial framework into the final Data Quality Assessment Framework.

Chapter 8. Results and final Data Quality Assessment Framework

This chapter first presents the results from the case study, then presents the results from the expert validation, and finally refines the initial framework into the final Data Quality Assessment Framework.

8.1 Results from case study

The application of the framework in the case study revealed a small number of ambiguities in the wording of the steps and the Data Quality Assessment Form. To address these ambiguities, several minor textual refinements were made to the framework to improve clarity and consistency, without altering its underlying structure or logic.

8.2 Results from expert validation

The experts suggested several ways to improve the framework. These suggestions have been divided into two groups:

1. Improvements that do not require further discussion and that are easy to implement. These improvements, which are mainly related to the accountability and reproducibility of the assessments, have been implemented in the final Data Quality Assessment Framework, as described in the next paragraph.

2. Improvements that require further discussion or investigation and/or that are too time-consuming to be implemented within the scope of this thesis research. These suggestions will be described in the Discussion chapter.

An example of group 1 is the explicit recording of assessment ownership. The experts have provided convincing arguments for the desirability of being able to link an assessment to a specific project and individual, because this clarifies responsibility, perspective, and purpose, and enables incremental refinement over time.
Other examples of group 1 include the benchmarking against other datasets and a preliminary review of relevant academic assessments, which have been added to the framework as optional supporting tools. The expert validation shows that defining the pass conditions and assessing the criteria (Steps 3 and 4) are the most challenging aspects of the framework, particularly for criteria such as completeness and representativeness. The new supporting tools help ground judgments without imposing fixed thresholds.
An example of group 2 is the addition of timeliness and other temporal aspects as dataset characteristics. These aspects are currently conceptualised in the framework as data quality criteria rather than descriptive properties. Treating them as separate characteristics could blur the distinction between dataset description and quality assessment without providing additional analytical clarity, so these potential changes require careful consideration and a more detailed discussion.
Another group 2 example is the possible inclusion of an explicit iterative component. The framework has been designed as a structured assessment tool rather than a procedural workflow, and prescribing

iterative cycles could potentially reduce clarity and ease of use. Finding a way to add iterations without compromising the current framework setup also requires further discussion.

8.3 Final Data Quality Assessment Framework

The improvements selected for incorporation into the Data Quality Assessment Framework have been translated into concrete modifications of the framework. These modifications are outlined below.

The Step-by-Step Guide is adjusted as follows:

- A note about benchmarking against other datasets has been incorporated in Step 3.
- A note about including preliminary academic literature review has also been incorporated in Step 3.
- Textual refinements have been made in various places, to improve clarity and avoid ambiguities.

The following fields have been added to the Data Quality Assessment Form:

- Assessment date
- Assessor name, department, contact details
- Project name and description
- Dataset name, version and provider
- Benchmark dataset references
- Academic review references

The final Data Quality Assessment Framework is included in Appendix G and the matching final Data Quality Assessment Form is included in Appendix H

Chapter 9. Discussion

9.1 Evaluation of results

9.1.1 Validity of the Data Quality Assessment Framework

The study evaluated the validity of the Data Quality Assessment Framework through two complementary forms of validation. First, the framework was applied in a case study, to illustrate the framework in practice and to identify possible ambiguities or limitations. Second, expert interviews were used to assess the relevance, clarity, and practical feasibility of the framework. Overall, both the case study and the interviews support the validity of the framework. The results indicate that the framework captures the main challenges of data quality assessment in supply chain analysis and provides a systematic structure that helps make evaluation choices explicit and reproducible.

Dataset characteristics and criteria

The inclusion of nine criteria and the dataset characteristics is grounded in the data quality literature and reflects recurring challenges in supply chain mapping. The expert interviews supported the relevance and completeness of this selection: experts recognised the criteria as covering the main data quality issues they typically encounter in practice, while also acknowledging the dataset characteristics.

The framework is designed to allow the set of data quality criteria to be adapted to what proves most effective in practical application. The number of criteria is deliberately limited to nine in order to balance analytical coverage with usability and to avoid making the assessment unnecessarily complex or time-consuming. As a result, some criteria frequently discussed in the literature, such as uniqueness (lack of duplicates) and referential integrity, are not included as explicit standalone criteria but are addressed implicitly within the existing set. These aspects could also be incorporated more explicitly, either as separate criteria or as clearly defined components of current criteria, if this improves clarity or ease of use in practice. As data practices and data sources evolve, further adjustments to the criteria may become necessary over time.

The framework models timeliness and temporal aspects, such as update frequency, historical coverage, and near real-time availability, as quality criteria. However, some experts argued that these properties could also be conceptualised as database characteristics, for example as aspects of dataset scope, similar to geographic coverage. These suggestions were not incorporated into the final Data Quality Assessment Framework, as further discussion and empirical testing would be required to determine which modelling choice best supports the framework's objectives.

Pass conditions, importance and assessment

A central practical challenge is formulating clear pass conditions and evaluating datasets against these conditions. Both the case study and the expert feedback indicate that the difficulties encountered in assessing certain criteria are inherent to real-world data quality work, rather than a weakness of the framework itself. In supply chain analysis, criteria such as completeness and representativeness are often inherently difficult to assess in a precise manner, because there is rarely an objective ground truth or a definitive benchmark. These challenges are therefore primarily related to the nature of the data and the industry context, not to the design of the framework.

In practice, data quality assessments often rely on information provided by dataset documentation,

under the assumption that reported validation steps have been performed as described. However, previous studies have shown that such validation steps are not always consistently implemented, or that their outcomes cannot be independently verified by users. This creates uncertainty about the actual quality of the data and highlights the need to treat documentation as evidence that requires critical evaluation, rather than as definitive proof of data quality.
Moreover, data quality assessment inevitably involves judgement. Even if pass conditions and importance levels are defined in detail, there is usually no single objectively correct threshold, and different reviewers may arrive at slightly different scores. This does not devalue the framework. Instead, it supports the framework's emphasis on a structured and transparent process. The framework's value lies in making the underlying reasoning explicit. By clearly documenting their reasoning, users enable readers and reviewers to understand the assumptions made and to reproduce the evaluation results.
The expert interviews further highlight why importance levels are essential. Experts differed in which criteria they considered most critical, which reflects that data quality is context-dependent. By explicitly weighting criteria, the framework enables users to align the evaluation with their analytical priorities, rather than implicitly assuming a universally applicable definition of data quality.

9.1.2 Practical application of the Data Quality Assessment Framework

The case study and the expert interviews indicate that the Data Quality Assessment Framework is usable and has clear potential for practical application.
The framework can be applied in several practical ways, depending on the analytical context and the purpose of the assessment. Its primary application is the systematic evaluation of a dataset with respect to a clearly defined use case, enabling the comparison of datasets when multiple datasets are available for a similar purpose.
In many supply chain analyses, however, data availability is limited and analysts are often required to work with imperfect datasets. In these cases, the framework does not primarily function as a strict accept-or-reject mechanism, but rather as a tool for identifying and documenting dataset strengths and weaknesses. By making limitations explicit, the framework supports reflection on how data quality constraints may affect analytical results and helps decide about possible solutions, e.g. a targeted data cleanup, or involving supplementary datasets.
The framework can also be applied retrospectively, after an analysis has been conducted. In this context, it supports reflection on whether methodological choices were adequately supported by the available data and enables transparent documentation of these choices.
The framework's practical relevance therefore extends beyond dataset selection alone, by supporting transparent and well-justified decisions about how datasets can be used, interpreted, and combined within an analytical process.
Despite its practical usefulness, the application of the framework is subject to several constraints that should be taken into account. These limitations are discussed in Section 9.3.

9.2 Significance of the Data Quality Assessment Framework

9.2.1 Contributions of this research

This research contributes to the literature on data quality in supply chain mapping by addressing the lack of a unified and structured approach for dataset evaluation. While existing studies have provided valuable insights into specific data quality criteria or individual supply chain databases, data quality assessment in supply chain mapping has remained fragmented. By developing a structured Data Quality Assessment Framework, this study offers a coherent and transparent approach that integrates multiple data quality criteria within a single assessment structure.
A key conceptual contribution of this research lies in its treatment of data quality in supply chain mapping as context-dependent rather than absolute. Instead of evaluating datasets in isolation, the framework explicitly assesses data quality with respect to a clearly defined use case. By linking dataset characteristics, criteria, pass conditions, and importance levels, the framework makes underlying assumptions explicit and requires users to articulate and justify their evaluation choices. This directly addresses limitations of existing approaches that are primarily tailored to empirically collected supply chain data and do not readily generalise to other data construction methods increasingly used in supply chain mapping.

9.2.2 Generalisability of results

The Data Quality Assessment Framework is developed and evaluated within the context of data quality assessment in supply chain mapping. The framework is developed with this scope in mind, and its applicability beyond supply chain mapping has not been explicitly considered or validated. At the same time, some underlying principles of the framework are not specific to supply chain mapping alone. Core elements, such as the explicit consideration of dataset characteristics, the use of clearly defined assessment criteria, the formulation of use-case-specific pass conditions, and the emphasis on transparency and documentation of assumptions, reflect general challenges that arise in many forms of data-driven analysis. These principles suggest potential relevance beyond the domain of supply chain mapping, but application in other domains would require careful contextualisation.

9.2.3 Practical implications of the research

In practice, insight into data quality does not necessarily translate into higher data quality. Identifying limitations in data quality does not enable users to avoid or replace problematic datasets. In many supply chain mapping contexts, analysts are required to work with datasets despite known shortcomings, simply because no better alternatives are available. Even when data quality assessments clearly reveal weaknesses such as incomplete coverage or limited transparency, these datasets often remain the only feasible basis for analysis. Awareness of poor data quality therefore does not remove reliance on such data, but instead increases the importance of cautious interpretation and explicit acknowledgement of uncertainty.
In practice, improved insight into supply chain structures also does not automatically translate into feasible policy or operational action. Even when data quality is well understood and dependencies or vulnerabilities within supply chains are clearly identified, addressing these issues often requires substantial financial resources, organisational capacity, and political coordination. Such resources are usually constrained, limiting the extent to which analytical insights can be translated into

concrete interventions. As a result, greater analytical clarity does not necessarily imply greater control over supply chain risks.
Within this context, the practical value of the Data Quality Assessment Framework does not lie in replacing imperfect datasets or directly resolving supply chain vulnerabilities, but in supporting more responsible and transparent use of available data. By clarifying which conclusions can and cannot be drawn from a given dataset, the framework helps prevent overconfidence in analytical results and supports more nuanced interpretation. This enables users to better justify analytical choices, communicate limitations more clearly to decision-makers, and focus attention on the most critical issues, even when neither data quality nor policy options can be readily improved.

9.3 Limitations

This study is subject to several limitations that should be considered when interpreting the findings.

General scope limitations
This study focuses specifically on data quality considerations and does not address several additional factors that may influence dataset suitability in practice.
First, sustainability considerations are not included in the framework. While sustainability is increasingly relevant in research and policy contexts, it is treated as outside the scope of this study. Consequently, the framework does not assess environmental or social impacts related to data collection, maintenance, or use, even though such factors may influence dataset selection in practice.
Second, the framework does not account for organisational and resource-related constraints that can affect the usability of datasets in practice. Factors such as internal data governance practices, existing data infrastructure, data integration costs, and budget limitations may significantly influence whether a dataset can be effectively adopted within an organisation, even when its assessed data quality is high. These practical considerations fall beyond the scope of the framework, which focuses on evaluating data quality rather than implementation feasibility.
Finally, legal and ethical considerations related to data use are not explicitly incorporated. Issues such as data ownership, contractual restrictions, or geopolitical considerations may limit the use of certain datasets, regardless of their assessed quality. For example, organisations may avoid using data from providers based in countries where regulatory or geopolitical uncertainty exists. While these factors can strongly affect practical applicability, they are treated as contextual constraints rather than as data quality criteria within this study.

Practical usability limitations
Despite its practical usefulness, the application of the framework is subject to several constraints that should be taken into account. One limitation arises in situations where a dataset is intended to serve multiple use cases. In practice, decisions such as whether to purchase a dataset are often based on its potential applicability across a range of analyses, rather than a single, well-defined use case. In such situations, applying the framework to only one specific application may provide an incomplete basis for decision-making. This highlights that the framework is most informative when the intended analytical purpose is clearly specified.
Another important constraint concerns the expertise required to apply the framework. Meaningful evaluation of criteria such as completeness, representativeness, or relevance depends on substantive

knowledge of the domain, but also on general data-related and analytical expertise. While the framework structures the assessment process, it does not eliminate the need for expert judgement. As a result, the quality of the assessment depends on the assessor's familiarity with both the topic of the use case and the characteristics of the data.
Another practical usability limitation is that the framework does not explicitly incorporate an iterative assessment component as one interviewed expert suggested. While datasets used for supply chain mapping often require methodological adjustments or the use of complementary data, the framework is applied as a structured, one-pass evaluation. As a result, decisions about reassessment and methodological refinement are left to the user, rather than being guided by the framework itself.
Data accessibility further constrains the practical application of the framework. In this study, full access to the dataset enabled a detailed evaluation. In many real-world settings, however, access to data is limited, particularly for commercial datasets that must be purchased before full inspection is possible. When access is restricted, the possibilities for assessing data quality are correspondingly reduced, and certain criteria can only be evaluated based on partial information or documentation. These constraints do not necessarily prevent the use of the framework, but they do require careful interpretation of the assessment outcomes. Several approaches can be used to address such limitations in practice. When the framework is applied to compare datasets, relative differences between datasets may be more informative than absolute assessment scores. In addition, uncertainty regarding data quality can be explicitly incorporated into the analysis by documenting limitations and reflecting on how these limitations influence the robustness of research conclusions. This may involve placing greater emphasis on validating analytical results or clearly indicating that certain conclusions should be interpreted with caution.
More generally, the framework allows users to determine how extensively it is applied in response to these practical constraints. Depending on the required level of reliability, the time and resources available, and the importance of the decision at hand, the assessment can range from a relatively lightweight evaluation to a more extensive application. This flexibility enables the framework to be adapted to different practical contexts while preserving its systematic structure.

Case study limitations
The study uses a single case study, which restricts the extent to which the framework can be illustrated and validated across different analytical contexts and datasets. Ideally, multiple case studies involving different types of datasets and use cases would be examined to further assess the robustness and versatility of the framework. However, this was not feasible within the scope of this thesis.
In addition, the selection of the case study dataset was shaped by data availability. Several potentially relevant datasets could not be included because they require licences or access rights that were not available within the budget and timeframe of this project. As a result, the case study relies on a dataset developed by academic researchers, who are generally associated with a high level of methodological transparency and care. This may limit the extent to which the findings can be generalised to commercially developed datasets.

Expert validation limitations
The expert validation is based on interviews with five experts from a single organisation (TNO). Although these experts differ in background and expertise, their shared organisational environment

may limit the diversity of perspectives represented in the feedback. Including experts from other organisations or sectors could provide additional insights into how the framework is perceived and applied in different practical settings.

9.4 Recommendations for future research

Building upon the Data Quality Assessment Framework, several options for future research and development can be suggested.

Extending the Data Quality Assessment Framework
A first recommendation for future research concerns exploring extensions of the framework beyond data quality criteria alone. In practice, dataset selection and use in supply chain mapping are often influenced by additional considerations, such as dataset costs, sustainability concerns, organisational constraints, and legal or ethical restrictions. Organisational factors may include internal data governance practices, existing data infrastructure, or data integration costs, all of which can affect whether a dataset can be effectively used in a given context.
Incorporating these factors into extended versions of the framework, or studying how they interact with data quality assessments in practice, would provide a more comprehensive understanding of dataset suitability. Such extensions could help bridge the gap between technical data quality evaluation and real-world decision-making in supply chain mapping, where trade-offs between quality, cost, feasibility, and compliance are often unavoidable.

Improving practical usability
A second important direction for future research involves improving the practical usability of the framework, particularly with respect to consistency and efficiency. While the proposed framework provides a structured approach to data quality evaluation, its application can currently be time-consuming and requires sufficient domain-specific expertise to be applied effectively. Future research could therefore focus on developing more standardised definitions, reference values, or measurement approaches for commonly used criteria. Such standardisation could reduce the effort required to apply the framework while preserving transparency.
This can be done through comparative applications of the framework across multiple datasets, which could support the development of empirical benchmarks for typical data quality profiles in supply chain mapping. By evaluating a broad range of datasets spanning different sectors, data collection methods, and scales, future studies could identify typical ranges for criteria such as accuracy or timeliness, for example by comparing public and commercial datasets. Such benchmarking could inform more consistent and empirically grounded pass condition thresholds, reducing reliance on ad hoc judgement and increasing comparability across assessments. Over time, this could also reveal opportunities to simplify the framework by identifying criteria that are consistently decisive or rarely binding in practice.
Partial automation of the assessment process could also help reduce the assessment effort. While expert judgement remains essential, certain steps could be supported by computational tools. For example, once a user describes the dataset and the intended use case, automated support could assist in proposing preliminary pass conditions, suggesting importance weightings, or checking consistency across assessments. Such developments could substantially reduce the time required to apply the framework, lower the level of expertise required for effective use, and thereby improve its

practical usability.
Another way of improving the practical usability of the framework concerns its application in situations where a dataset is intended to support multiple use cases. In practice, datasets are often evaluated not for a single, clearly defined analysis, but for their potential applicability across a range of analytical contexts. Future research could explore how the framework can be extended to better accommodate such scenarios, for example by supporting parallel or layered assessments that document how dataset suitability varies across different use cases. This would reduce the need to perform fully separate evaluations while still preserving transparency regarding use-case-specific assumptions, pass conditions, and importance levels.
The practical usability of the framework can be further improved by supporting iterative use more explicitly. Although datasets used in supply chain mapping frequently require methodological adjustments or the combination of complementary data sources, the framework is currently applied as a structured, one-pass evaluation. Future research could investigate how iterative reassessment might be more explicitly supported, for example by providing guidance on when reassessment is appropriate or how updated information should be incorporated, while maintaining clarity and ease of use.
Finally, future research could address practical usability challenges related to data accessibility. As access to full datasets is often restricted, particularly in the case of commercial data providers, future work could explore how the framework can be applied more effectively under conditions of partial information. This may include developing guidance for assessments based on documentation, metadata, or limited samples, and clarifying how uncertainty arising from restricted access should be documented and interpreted within the evaluation. In addition, future research could explore the development of standardized tests that data vendors can execute without disclosing their full datasets, allowing certain quality aspects to be assessed while respecting confidentiality constraints.

Extending the empirical validation

Finally, future research could explore extending the empirical validation of the framework. Applying the framework to a larger number of case studies would enable systematic evaluation of its robustness and generalisability across different supply chain mapping contexts. Future studies could examine how the framework performs when applied to different types of datasets and use cases.
In addition, future research could investigate how different assessors apply the framework to the same dataset. Analysing inter-assessor variation would provide insight into the degree of subjectivity involved in the evaluation process and help evaluate whether pass conditions and importance levels are interpreted consistently.
Including experts from a wider range of organisations, such as commercial data providers and public institutions, would further strengthen the empirical validation of the framework. Broader expert involvement would help assess whether the framework is interpreted consistently across different practical contexts and would enhance the external validity of the findings.

Chapter 10. Conclusion

The main research question that this thesis examined was how the data quality of datasets for supply chain mapping can be systematically assessed. As supply chain mapping increasingly informs policy decisions, strategic investment choices, and academic research, the reliability of the underlying datasets becomes critical. The research question was addressed through the development of an integrated Data Quality Assessment Framework. Each subquestion contributed a necessary building block to this framework:

- RQ.1 examined the types of datasets used for supply chain mapping and identified their main characteristics: mapping method, scope, resolution, and data source. These characteristics provide important contextual information for understanding why data quality challenges differ across datasets.
- RQ.2 focused on identifying relevant criteria for assessing data quality in supply chain mapping. Drawing on both general data quality literature and domain-specific studies, a structured set of nine criteria was selected and tailored to the specific context of supply chain analysis.
- RQ.3 addressed how these criteria can be operationalised through explicit pass conditions and importance weightings. The framework provides guidance on how users can define and document these elements in a transparent manner, while allowing flexibility to reflect different analytical purposes.

Together, the answers to the subquestions demonstrate how the concept of data quality can be translated into a coherent and practically applicable step-by-step assessment approach.
Application of the framework in a case study on wind turbine production, using the AI-Generated Production Network (AIPNET) dataset, demonstrated that the framework can be applied without major obstacles and helps structure analytical judgment. The subsequent expert validation confirmed the framework's relevance and practical feasibility, while also identifying areas where further refinement or support could enhance usability.
The relevance of this research lies in its contribution to greater transparency and accountability in the use of supply chain mapping data. The proposed framework helps policymakers, analysts, and researchers evaluate dataset suitability and quality by explicitly documenting limitations, assumptions, and sources of uncertainty. From an academic perspective, the thesis addresses a clear gap in the academic literature by integrating fragmented data quality dimensions into a unified framework applicable across different dataset types, including empirically collected, AI-enhanced, and AI-generated datasets.
A limitation of this research is that the framework specifically focuses on data quality and does not incorporate broader considerations such as dataset costs, sustainability aspects, legal constraints, or organisational feasibility. In addition, the empirical validation is based on a single case study and expert interviews within one organisation, limiting the generalisability of the findings. Applying the framework also requires domain knowledge and analytical expertise, which may lead to variation between assessors. Future research could address these limitations by evaluating additional dimensions for the framework, applying it across a wider range of datasets and case studies, and exploring options for improving usability, robustness, and consistency.
In conclusion, this thesis demonstrates that systematic data quality assessment is both feasible and necessary in supply chain mapping. In principle, virtually any research project that relies on

data sources could benefit from a structured, reproducible, and as far as possible standardised assessment of those sources, conducted prior to or during the project. The Data Quality Assessment Framework provides an initial step towards such a systematic and structured approach. By making data quality considerations explicit rather than implicit, such assessments support more responsible analysis, more credible conclusions, and ultimately better-informed decisions in an increasingly data-driven world.

References

Bans-Akutey, A. and Tiimub, B. M. (2021). Triangulation in research. *Academia Letters*. https://doi.org/10.20935/al3392.

Bansal, P., Gualandris, J., and Kim, N. (2020). Theorizing supply chains with qualitative big data and topic modeling. *Journal of Supply Chain Management*, 56(2). https://doi.org/10.1111/jscm.12224.

Bender, E. M., Gebru, T., et al. (2021). On the dangers of stochastic parrots: Can language models be too big? FAccT '21. Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922.

Calantone, R. J. and Vickery, S. K. (2010). Introduction to the special topic forum: using archival and secondary data sources in supply chain mannagement research. *Journal of Supply Chain Management*, 46(4). https://doi.org/10.1111/j.1745-493x.2010.03202.x.

Cook, R. G., Campbell, D. K., and Kelly, C. (2012). An issue of trust: Are commercial databases really reliable? *Journal of Business Finance Librarianship*, 17(4). https://doi.org/10.1080/08963568.2012.712501.

Culot, G., Podrecca, M., Nassimbeni, G., Orzes, G., and Sartor, M. (2022). Using supply chain databases in academic research: A methodological critique. *Journal of Supply Chain Management*, 59(1). https://doi.org/10.1111/jscm.12294.

Dong, Y., Zou, F., Song, S., Peng, Y., and Xu, K. (2024). Supply Chain Data and Analysis: The Case of Supply Chain Innovation. *SSRN Electronic Journal*. http://doi.org/10.2139/ssrn.4974699.

Elkington, J. (1999). *Cannibals with Forks*. Capstone Publishing.

Ellram, L. M. and Tate, W. L. (2016). The use of secondary data in purchasing and supply management (P/SM) research. *Journal of Purchasing and Supply Management*, 22(4). https://doi.org/10.1016/j.pursup.2016.08.005.

European Commission (2019). The european green deal. `https://eur-lex.europa.eu/resource.html?uri=cellar:b828d165-1c22-11ea-8c1f-01aa75ed71a1.0002.02/DOC_1&format=PDF`.

European Commission (2023a). Joint communication on a european economic security strategy. `https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52023JC0020`.

European Commission (2023b). Net-zero industry act: Strengthening europe's net-zero technology manufacturing ecosystem. `https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52023PC0161`.

European Commission Joint Research Centre (2020). Critical materials for strategic technologies and sectors in the eu: A foresight study. `https://rmis.jrc.ec.europa.eu/uploads/CRMs_for_Strategic_Technologies_and_Sectors_in_the_EU_2020.pdf`.

FactSet (2025). Factset supply chain relationships [dataset]. Online dataset. https://www.factset.com/marketplace/catalog/product/factset-supply-chain-relationships.

Fetzer, T., Lambert, P. J., Feld, B., and Garg, P. (2024). Ai-generated production networks: Measurement and applications to global trade (aipnet v20241118). Technical report, Kiel Institute for the World Economy. https://aipnet.io/wp-content/paper/ai_generated_production_networks_aipnet_v20241118.pdf.

Gardner, T., Benzie, M., Börner, J., Dawkins, E., Fick, S., Garrett, R., Godar, J., Grimard, A., Lake, S., Larsen, R., Mardas, N., McDermott, C., Meyfroidt, P., Osbeck, M., Persson, M., Sembres, T., Suavet, C., Strassburg, B., Trevisan, A., West, C., and Wolvekamp, P. (2019). Transparency and sustainability in global commodity supply chains. *World Development*, 121. https://doi.org/10.1016/j.worlddev.2018.05.025.

International Energy Agency (2021). Net zero by 2050: A roadmap for the global energy sector. https://www.iea.org/reports/net-zero-by-2050.

International Energy Agency (2022). Renewables 2022: Analysis and forecast to 2027. https://www.iea.org/reports/renewables-2022.

Lee, D. and Yoon, B. (2025). The development of a methodology for assessing data value through the identification of key determinants. *Systems*, 13(4). https://doi.org/10.3390/systems13040305.

Liu, G. (2020). Data quality problems troubling business and financial researchers: A literature review and synthetic analysis. *Journal of Business Finance Librarianship*, 25(3-4). https://doi.org/10.1080/08963568.2020.1847555.

MacCarthy, B. L., Ahmed, W. A., and Demirel, G. (2022). Mapping the supply chain: Why, what and how? *International Journal of Production Economics*, 250. https://doi.org/10.1016/j.ijpe.2022.108688.

Mau, K. and Vicencio, A. (2025). Inferring input-output linkages with firm-level trade data. Technical report, Maastricht University.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., and Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT '19)*. Association for Computing Machinery. https://doi.org/10.1145/3287560.3287596.

Montecchi, M., Plangger, K., and West, D. C. (2021). Supply chain transparency: A bibliometric review and research agenda. *International Journal of Production Economics*, 238. https://doi.org/10.1016/j.ijpe.2021.108152.

Mubarik, M. S., Kusi-Sarpong, S., Govindan, K., Khan, S. A., and Oyedijo, A. (2021). Supply chain mapping: a proposed construct. *International Journal of Production Research*, 61(8). https://doi.org/10.1080/00207543.2021.1944390.

Priestley, M., O'donnell, F., and Simperl, E. (2023). A Survey of Data Quality Requirements That Matter in ML Development Pipelines. *Journal of Data and Information Quality*, 15(2). https://doi.org/10.1145/3592616.

Statistics Netherlands (CBS) (2025). Nederlandse afhankelijkheid van kritieke materialen. https://www.cbs.nl/nl-nl/publicatie/2025/40/nederlandse-afhankelijkheid-van-kritieke-materialen.

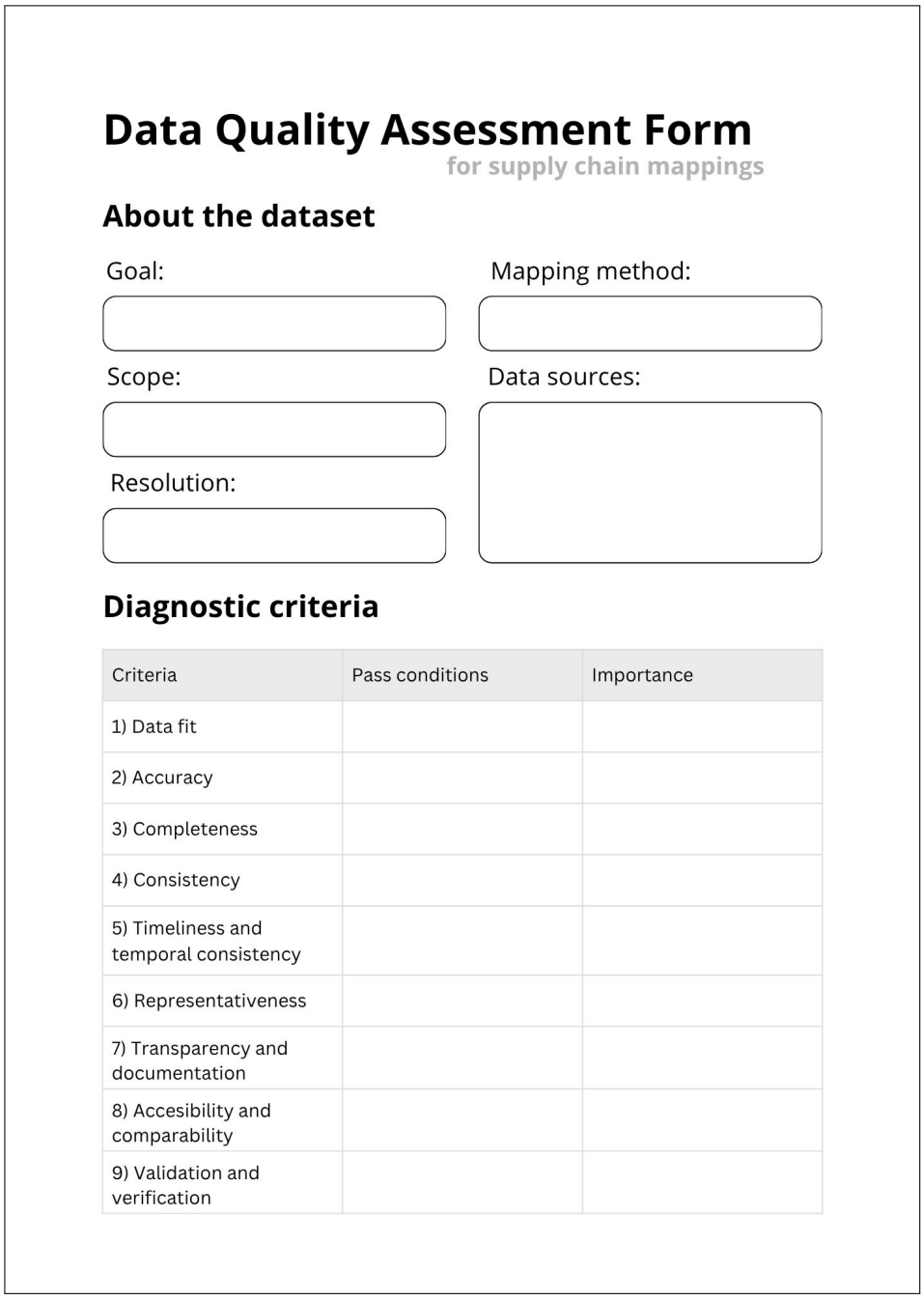
Van Schilt, I. M., Kwakkel, J. H., Mense, J. P., and Verbraeck, A. (2024). Dimensions of data sparseness and their effect on supply chain visibility. *Computers Industrial Engineering*, 191. https://doi.org/10.1016/j.cie.2024.110108.

Wang, R. Y. and Strong, D. M. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems*, 12(4). https://doi.org/10.1080/07421222.1996.11518099.

Appendices

Chapter A. Initial Data Quality Assessment Form

This version of the Data Quality Assessment Form was used in the case study and the expert interviews.



Data Quality Assessment Form

for supply chain mappings

About the dataset

Goal:

Mapping method:

Scope:

Data sources:

Resolution:

Diagnostic criteria

Criteria	Pass conditions	Importance
1) Data fit		
2) Accuracy		
3) Completeness		
4) Consistency		
5) Timeliness and temporal consistency		
6) Representativeness		
7) Transparency and documentation		
8) Accesibility and comparability		
9) Validation and verification		

Assessment

Criteria	Result	Decision
1) Data fit		
2) Accuracy		
3) Completeness		
4) Consistency		
5) Timeliness and temporal consistency		
6) Representativeness		
7) Transparency and documentation		
8) Accesibility and comparability		
9) Validation and verification		

Conclusion

Chapter B. Detailed dataset description - AIPNET

The AI-Generated Production Network (AIPNET) dataset offers a comprehensive, AI-derived map of global production linkages across traded goods. It represents how products are connected through input-output relationships in production processes using a "cutting-edge AI language model", specifically GPT-4o. AIPNET systematically maps these connections for over 5,000 product categories defined in the Harmonized System (HS). The Harmonized System is "a nomenclature of product codes maintained by the World Customs Organization (WCO) and used universally by national trade authorities" (Fetzer et al., 2024).

To construct AIPNET, GPT-4o is prompted to identify, for each product, which other goods serve as inputs or play a role in its production, based on their HS descriptions. A two-step build-prune approach is applied: in the build phase the model generates candidate linkages by asking for vertically related products, and in the prune phase it evaluates each proposed connection to retain only plausible input-output relationships. This methodology ensures that the final network reflects real production dependencies.

AI-generated product descriptions are matched to official HS codes using text embedding techniques that convert textual descriptions into numerical representations that capture their meaning. These representations are compared to find the best matches between AI outputs and the standardized HS codes, producing a structured graph of product linkages. To further validate the network, the model explicitly assesses whether one product can realistically be used to produce another, refining the edge set to improve fidelity.

AIPNET's validity is also supported by comparison with traditional economic data such as Input-Output Tables: the dataset shows strong alignment with real patterns of production and trade, indicating that the AI-generated linkages reflect actual economic relationships.

The resulting network includes more than 5,000 product nodes and nearly one million directed connections at the HS6 level. Each product appears as a node, and each production linkage as a directed edge from input to output. Although the overall network is relatively sparse, clusters of densely linked products emerge in sectors where inputs participate in many downstream production processes.

Products in AIPNET are further categorized by economic function. These classifications distinguish between intermediate goods (used as inputs in production), capital goods (tools and equipment used in production), and final consumption goods (bought directly by end consumers). Intermediate and capital goods typically display high connectivity within the network because they are widely used across multiple production contexts.

A critical indicator included in the dataset is Integrated Global Product Centrality (IGPC). IGPC measures the relative importance of each product in the global production system by combining trade data (such as trade volume and concentration) with network position information. It builds on principles similar to algorithms like PageRank: a product is considered more central if it is widely traded, serves as an essential input to many other goods, and/or is connected to other highly linked products. IGPC thus highlights a product's significance in global economic activity.

The AIPNET data pack comprises several complementary files that together describe the global production network in detail. The Node List contains information for each product, including:

- hs_code_6d: 6-digit Harmonized System code
- short_label: Brief product description

- full_description: Detailed product description
- bec_desc: Broad Economic Category description
- hs_section: HS section classification
- hs_code_2d, hs_code_4d: 2-digit and 4-digit aggregations
- IGPC: Integrated Global Product Centrality score
- GTS: Global Trade Share (%)
- TC: Trade Concentration (%)

The Edge Lists provide the comprehensive set of product-to-product connections identified by the AI. These lists are available at both HS4 and HS6 levels and cover multiple HS vintages (including 2002, 2007, 2012, 2017, and 2022), where each vintage refers to a specific edition of the HS classification updated periodically by the World Customs Organization. Each file enumerates source and target product codes that represent directional production linkages between goods.

Chapter C. Case study: assessment criteria

This appendix contains the detailed data for Step 4 in the case study analysis.

Assessment Data fit
Presence of the wind turbine product category:

1. 6 digit HS code: 850231 (wind-powered generating sets)
2. Check: the HS code is present in the dataset

Coverage of key component categories:

1. Key wind turbine component categories are identified based on standard descriptions of wind turbine production.
2. Key components and their HS codes:
   - Rotor blades - HS 850300
   - Tower (steel structure) - HS 730820
   - Nacelle (housing and integrated systems) - HS 850300
   - Gearbox - HS 848340
   - Generator - HS 850164
   - Yaw and pitch control systems - HS 853710
   - Main bearings - HS 848210
   - Power electronics and converters - HS 850440

   For a visualisation of the key components of a wind turbine, see Figure 2.
3. Check: HS codes are present in the dataset

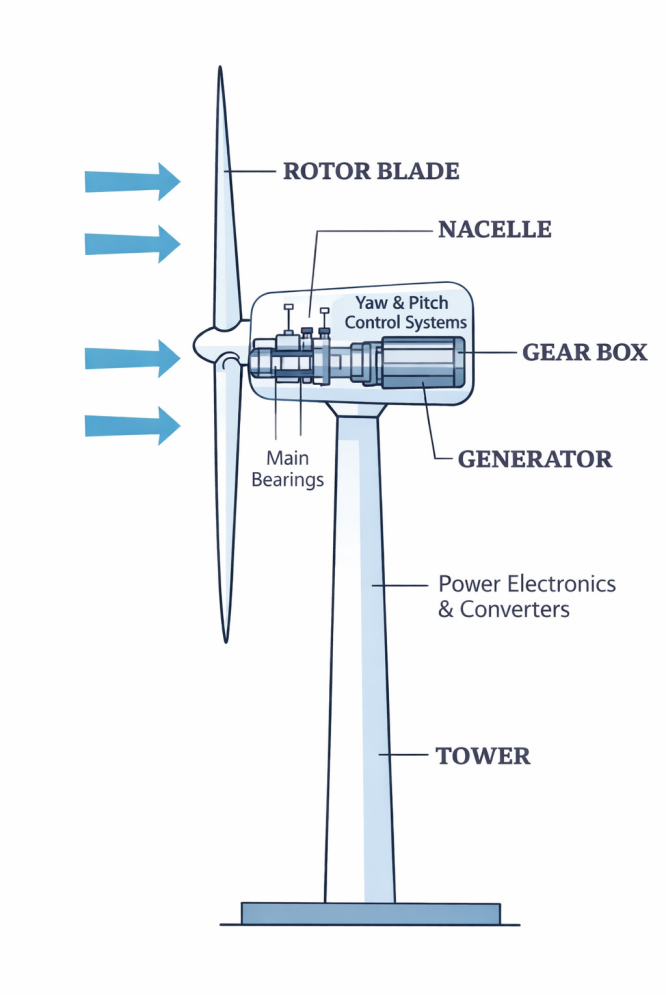


Figure 2: Schematic overview of the main components of a wind turbine. *Source: AI-generated illustration created using ChatGPT (OpenAI)*

Sufficient product-level granularity:

1. Product-level granularity is available at both 6-digit and 4-digit HS levels.
2. The HS classification does not always isolate wind turbine-specific subcomponents, a limitation that is inherent to the HS system rather than to the AIPNET dataset. While this limits component-level separation, the available product-category granularity remains sufficient for identifying turbine-specific upstream inputs and analyzing production and supply chain dependencies.

Global scope:

1. According to the dataset documentation, AIPNET constructs product-to-product linkages using a generative AI model that infers input-output relationships based on HS product descriptions.

2. The resulting network is defined at the product level, rather than being tied to any specific country or region. As a result, the dataset reflects globally applicable production structures, which is what the case study requires.

Assessment Accuracy
Minimum of 95% accuracy:

1. A total of 107 unique direct upstream product categories are identified as inputs to HS 850231 in the Edge List of 2022. These are identified as follows:
   (a) Use Excel's Filter function on the downstream HS code column. Filter the dataset to retain only rows where the downstream product equals HS 850231.
   (b) Select all corresponding values in the upstream HS code column and copy the upstream HS codes into a new worksheet.
   (c) Apply Excel's Remove Duplicates function to the upstream HS code column.
   (d) Exclude the header row from the count.
2. The inputs are verified with the following steps:
   (a) An upstream category is considered plausible if it can reasonably function as:
   - an electrical component or part
   - a mechanical component or transmission element
   - a control, regulation, or power conversion device
   - a raw or semi-processed material commonly used in electrical or mechanical equipment production

   An input relationship is classified as incorrect only if the upstream product category is clearly unrelated to wind turbine production.
   (b) For each upstream HS code, the HS heading (first four digits) is used to determine the broad product category defined by the Harmonized System. Each input is then mapped to a high-level functional group. The relationship is considered plausible if the HS heading corresponds to a product group that is technologically consistent with the production of wind turbines (the list above).
3. Correct relationships count: 107
4. Accuracy = (count correct relationships/total relationships) * 100%
   Accuracy = (107/107) * 100% = 100%

Assessment Completeness
Ideally 100% completeness. If <100%: missing data has minimal impact:

1. A list of key direct input product categories for wind turbine production is compiled at the 6-digit HS level, consistent with the granularity of the AIPNET dataset. Key input categories are defined based on the functional requirements of producing wind turbines and include:

- Electric motors, generators, transformers, converters, and related parts (e.g. HS 850110-850153, 850300, 8504xx)
- Mechanical components and transmission elements such as bearings, gearboxes, shafts, and machinery parts (e.g. HS 848210-848299, 848340-848360, 8479xx)
- Insulated conductors, cables, and electrical insulators (e.g. HS 854411-854470, 854610-854690)
- Metal inputs used in electrical and mechanical equipment manufacturing, including steel, copper, and aluminium in semi-processed form (e.g. HS 72xxxx, 730650, 74xxxx, 76xxxx)
- Control, regulation, and protection equipment (e.g. HS 8535xx-8538xx, 903220)

2. The dataset is checked to verify whether each key input category appears as a direct upstream input to HS 850231:
   (a) Use the 107 unique direct upstream product categories identified as inputs to HS 850231 (in the Edge List of 2022) in the accuracy assessment.
   (b) Check if the HS codes of the key input categories appear as direct upstream input to HS 850231.
3. No key direct input product categories defined in Step 1 are missing from the dataset. All key categories are represented by at least one 6-digit HS code among the direct upstream inputs to HS 850231.
4. Completeness = (missing key direct input product categories / total key direct input product categories) * 100%

   Completeness = 100%

Assessment Consistency
Zero or minimal classification errors:

1. Consistency at the product category level is primarily ensured through the use of the HS classification. The AIPNET dataset adopts HS 2002 at the 6-digit level as its sole product taxonomy. Within a given HS vintage, each 6-digit code corresponds to a unique and internationally standardized product definition maintained by the World Customs Organization.

   A known classification challenge when working with HS data is that HS codes and product definitions change over time, which can lead to inconsistencies in longitudinal analyses. In the present study, however, the analysis is not conducted over time. As a result, changes in HS classifications across revisions do not affect the consistency of product definitions within this use case.

2. To assess whether products are inconsistently classified under multiple product categories, the product descriptions in the HS02 6-digit node list are examined. At the level of short product descriptions, identical or very similar labels (for example, generic descriptions such as “Combed Cotton Yarn”) appear multiple times across different HS codes.

   This repetition does not indicate a classification inconsistency. In the HS nomenclature, short descriptions serve as broad labels, while the full HS descriptions provide the legally

precise product definition, specifying material composition, processing stage, or technical characteristics. In the AIPNET dataset, these distinctions are captured through unique 6-digit HS codes and their corresponding detailed descriptions.

As a result, although short descriptions may repeat, each product category remains uniquely and consistently defined by its HS code and full description. The HS code, rather than the short textual label, functions as the authoritative identifier, preserving classification consistency.

3. To assess internal consistency of the network structure, the HS02 6-digit edge list is compared against the HS02 6-digit node list. A consistent network requires that every upstream and downstream HS code appearing in the links also exists as a defined product node.

   (a) From the HS02 6-digit edge list: all unique HS codes appearing in either column are extracted with the Excel's Remove Duplicates function. This represents the full set of product codes actually used in the network connections.

   (b) The set of HS codes appearing in the edge list is compared to the set of HS codes in the node list. If an HS code appears in the edge list but not in the node list, it is flagged as a missing node. The missing nodes are:

   HS 271011

   HS 271019

   HS 271091

   HS 271099

   HS 293294

   HS 710820

   HS 711890

   Amount of missing nodes: 7 missing nodes

   (c) All edges that contain at least one of these missing HS codes (either upstream or downstream) are counted.

   HS 271011 - 95 edges

   HS 271019 - 65 edges

   HS 271091 - 16 edges

   HS 271099 - 33 edges

   HS 293294 - 29 edges

   HS 710820 - 38 edges

   HS 711890 - 218 edges

   Missing nodes appear in: 494 - 1 (both in upstream and downstream a missing node) = 493 edges

   Full HS02 edge list contains: 429871 edges

   The share of affected links is: 493/429871 = 0,00115 (= 0,115%)

Because the issue affects well below 1 percent of all links and is limited to a small number of codes, it is classified as minor and correctable, not as a systematic classification error. And the discrepancies do not seem to affect the wind turbine case analysis.

In the HS02 6-digit network, the edge list contains 429,871 unique upstream-downstream links and shows no duplicated edges and no self-loops, which supports structural consistency.

Finally, the hs_code_4d column does not contain true 4-digit HS codes, but instead repeats the first three digits of the corresponding 6-digit code (e.g. 270 instead of 2709). While this does not affect the network structure, it may cause confusion if used without adjustment.

Overall, the network structure remains internally coherent, with only limited issues related to referential integrity and minor variable definition inconsistencies that are straightforward to handle through preprocessing.

Assessment Timeliness and temporal consistency

Data not older than 5 years:

According to the AIPNET documentation, the dataset was constructed and released in 2024. Although AIPNET relies on the HS 2002 classification to define product categories, this does not compromise timeliness for the present study. The HS codes serve as a stable product taxonomy, while the production network itself is constructed using contemporary generative AI models and reflects current technological knowledge of input-output relationships. As a result, the use of an older HS vintage does not violate the five-year timeliness requirement.

Assessment Transparency and documentation

Data sources documented:

The data sources are described in the documentation:

- HS 6-digit product codes and official HS descriptions
- Large Language Model (LLM) pre-training corpora
- Official Input-Output and Supply-Use Tables published by national statistical authorities

Mapping methodology described:

The mapping methodology used to construct the AIPNET dataset is described in detail in the accompanying paper. The documentation explains how AI-generated descriptions are mapped to HS product codes and how candidate linkages are filtered to obtain the final production network.

Validation approach documented:

The validation methods used to assess the AIPNET dataset are described in the accompanying paper. The documentation explains how the inferred production network is benchmarked against official Input-Output and Supply-Use Tables and how robustness checks are conducted.

Dataset limitations stated:

The accompanying paper explicitly documents several constraints related to inference, scope, and validation of the AIPNET dataset. It discusses limitations inherent to AI-based inference, including the risk of spurious or implausible relationships and the need for conservative pruning. The documentation also acknowledges trade-offs between precision and recall, as well as limitations arising from validating a highly granular HS-based network against more aggregated industry-level Input-Output and Supply-Use Tables. These statements clarify the scope of applicability of the

dataset and the contexts in which caution is required.

Assessment Accessibility and comparability
Dataset is accessible to researchers:

1. The AIPNET dataset is publicly available for academic and research purposes. Access is granted upon submission of a description of the intended use. While access conditions are framed in terms of research use, the dataset is accessible for policy-oriented analyses that are conducted within a research or analytical context. No technical or licensing barriers prevent such use in practice.

2. Once access is granted, the dataset is immediately available as a downloadable data-pack. The data is provided in standard, machine-readable formats (CSV) and can be processed using commonly available software tools. No specialized infrastructure or proprietary software is required, allowing the dataset to be used in practice for policy analysis.

Product identifiers and dataset structure allow comparison with external datasets:

1. Product categories in AIPNET are defined using internationally standardized HS product codes at the 4-digit and 6-digit level. These identifiers are widely used in policy-relevant datasets, including trade statistics and Input-Output tables.

2. The AIPNET dataset is structured as an explicit product-level network, consisting of a node list of HS coded products and an edge list describing upstream-downstream production relationships. This structure allows the dataset to be aggregated or mapped to external policy-relevant datasets, such as trade statistics or Input-Output tables, using standard mapping procedures. While differences in granularity may require aggregation, the dataset structure enables such alignment in principle.

Assessment Validation and verification
Convincing documentation of triangulation techniques:

1. The AIPNET documentation describes the use of multiple triangulation techniques in the construction of the production network:

    - Repeated model inference: Multiple LLM draws reduce reliance on single outputs and capture stochastic variation.
    - Consensus filtering across iterations: Only relationships consistently appearing across iterations are retained.
    - HS taxonomy-based triangulation: AI-generated descriptions are mapped to official HS codes using embeddings and cosine similarity.
    - Similarity thresholding: Weak or ambiguous text-to-code matches are excluded.
    - LLM-based pruning: Implausible or substitute production linkages are removed through plausibility checks.
    - Conservative ambiguity handling: Uncertain relationships are rejected by default.

- External validation: Aggregated linkages are compared with Input-Output and Supply-Use Tables.
- Network perturbation analysis: Random edge rewiring tests robustness of network structure.
- Cross-country validation: Input-Output data from multiple countries is used to assess generalizability.

2. The validation methods used in AIPNET are clearly documented and methodologically transparent, but not without limitations. Internal procedures such as repeated inference, consensus filtering, and LLM-based pruning reduce noise and implausible linkages, but remain dependent on the LLM's underlying knowledge, which cannot be independently audited.

   External validation against Input-Output and Supply-Use Tables provides a credible empirical benchmark, but is necessarily partial. These data are available only at an aggregated, industry-based level, requiring aggregation of AIPNET linkages and limiting direct verification of individual product-level connections.

   Network perturbation analysis further supports credibility by showing that random edge rewiring reduces alignment with official Input-Output tables. This validates the relevance of the overall network structure, but not the correctness of individual product-level linkages.

   Overall, the validation approach is convincing at the network and aggregate level. However, fine-grained product-level linkages cannot always be directly validated, leaving some uncertainty.

Chapter D. Interview protocol

Interview protocol

Begin with a short introduction of yourself and your research. Explain that the purpose of the thesis is to develop a structured framework for assessing the data quality of datasets used for supply chain mapping. Mention that the framework is based on insights from academic literature and that the goal of this interview is to validate the relevance, clarity, and practical feasibility of the Data Quality Assessment Framework, as well as the underlying theoretical concepts on which it is built. Confirm that the interviewee agrees to participate and to be recorded. Inform the interviewee that, after the interview, a short summary will be shared to confirm that their input was understood correctly. Ask the interviewee to briefly introduce themselves, and verify their name, current role, and area of expertise.

1. **Background and dataset types** - 5 minutes
   Purpose: validate whether the dataset characteristics are conceptually sound and reflect real-world practice.

   - To what extent do the dataset characteristics included in the framework (mapping method, scope, resolution, and data sources) reflect the way you typically categorise or distinguish datasets in your work?
     → *Probe:* Are there any characteristics you believe are missing, or any included characteristics that you found unexpected?
   - Could you describe one or two situations in which you decided to reject a dataset because certain characteristics were not acceptable?

2. **Diagnostic criteria** - 10 minutes
   Purpose: assess relevance, clarity, completeness, and practical applicability of the nine diagnostic criteria.

   - To what extent do the nine diagnostic criteria reflect the data quality challenges that you typically observe in practice?
     → *Probe:* Are there any criteria you believe are missing, or any included criteria that you found unexpected?
     → *Probe:* Are there any criteria you believe are broader/narrower than anticipated?
   - Can you provide one or more examples from your own work where a particular criterion became especially important?
     → *Probe:* If you were in a situation where you had to use a particular dataset because no alternative was available, which criteria would you be most willing to compromise on, and why?
   - Has there been, or can you imagine, a situation where a client or end user would dismiss conclusions based on one of the nine diagnostic criteria?

3. **Data Quality Assessment Framework** - 10 minutes
   Purpose: evaluate the logical structure, clarity, usability, and completeness of the Data Quality Assessment Framework.

- Does the framework provide sufficient guidance to support consistent application?
  → *Probe:* Which component of the framework, if any, do you expect users may find challenging to apply in practice?
  → *Probe:* Do you believe the order of the steps is logical and intuitive?
- Do you believe any steps or components are missing that would be necessary for a complete assessment?
- To what extent do you consider this framework practically usable for evaluating real-world supply chain mapping datasets?

4. **Closing question** - 5 minutes
   Is there anything we have not discussed during this interview that you would like to mention?

Chapter E. Interview briefing document

Interview briefing document

Thesis: Data Quality in Supply Chain Mapping

Introduction and context

My name is Emma Glorie and I am a bachelor student Computer Science & Economics at Leiden University. Currently I am conducting my thesis at TNO as a graduation internship. My research focuses on developing a framework for assessing the data quality of datasets used for supply chain mapping. The framework follows the fitness-for-use principle, meaning that data quality is evaluated in relation to the specific analytical purpose for which the dataset is intended.

The purpose of this interview is to gather your expert feedback on the clarity, relevance, and practical feasibility of the framework and its underlying concepts. Your input will help determine whether the framework aligns with real-world practice. Before starting the interview, I will ask for your verbal consent to record the conversation.

This document provides a brief overview of:

- the research context
- the dataset characteristics
- the nine diagnostic criteria
- the Data Quality Assessment Framework
- the structure and questions of the interview

Dataset characteristics

To distinguish between the different types of datasets used for supply chain mapping, we use several key dataset characteristics. These characteristics help clarify how a dataset is constructed and what types of analyses it is suitable for.

Mapping method: How the dataset is created. In practice, many datasets are a combination of the different methods.

- Empirically collected datasets - Based on directly observed information such as company disclosures, surveys, or operational records.
- AI-enhanced datasets - Built on empirical data but expanded or refined using algorithmic inference (e.g. association-rule mining, probabilistic estimation).
- AI-generated datasets - Constructed mainly through model-based reasoning (e.g. large language models identifying product linkages).

Scope:

Scope refers to the geographical and sectoral coverage of the dataset. It concerns whether the data captures linkages within a single country, across regions, or on a global scale, and whether it focuses

on one industry or spans multiple sectors.

Resolution:
Resolution reflects the degree of detail within that scope, such as whether relationships are represented at the company, sector, or even national level.

Data sources:
The origin of the underlying data. This may include primary data (surveys, interviews), secondary data (company reports, customs data, commercial databases), and big data sources (web-scraped supplier lists, logistics records, patent or product registries). Many datasets combine multiple sources.

Diagnostic criteria
The Data Quality Assessment Framework is based on nine diagnostic criteria derived from the academic literature. These criteria help evaluate how suitable a dataset is for a specific analytical or policy purpose. Below is a shortened description of each criterion.

1. *Data fit*
   The extent to which the dataset's structure, coverage, and granularity align with the intended analytical or policy goal.

2. *Accuracy*
   The degree to which recorded relationships reflect real-world supply-chain linkages. Includes classification errors and incorrect matches.

3. *Completeness*
   How fully the dataset captures relevant entities, tiers, and attributes, and whether important suppliers or relationships are missing.

4. *Consistency*
   Uniformity in the use of definitions, classifications, and identifiers within the dataset, and alignment with other data sources.

5. *Timeliness & temporal consistency*
   Whether the dataset is up-to-date and whether time-related information is recorded consistently (e.g. stable reporting periods, documented updates).

6. *Representativeness*
   The extent to which the dataset reflects the full and balanced population relevant for the analysis, without strong sampling biases.

7. *Transparency & documentation*
   Availability of clear documentation on data collection, verification, and processing methods, allowing users to understand how the dataset was constructed.

8. *Accessibility & comparability*
   How easily users can access the data and combine it with other sources, considering format, licensing, and compatibility.

9. *Validation & verification*
   Evidence of internal checks or external comparisons that confirm the reliability of the data or specific relationships.

Data Quality Assessment Framework
The Data Quality Assessment Framework provides a structured and practical approach that can be used to evaluate datasets for supply chain mapping. It outlines how the criteria can be assessed, how pass conditions can be defined and how the overall evaluation can be documented in a consistent manner.

Pass conditions:
For each diagnostic criterion, pass conditions define what constitutes acceptable quality in the given context. These conditions make the evaluation explicit and can be qualitative (e.g. sufficient documentation available) or quantitative (e.g. update frequency or coverage thresholds). Pass conditions are not universal; they are formulated by the user based on the dataset's intended use and characteristics. The framework provides guidance for this process, including indicative importance levels and contextual distinctions that help determine how strict each pass condition should be.

Importance levels:
Each criterion is assigned a verbal importance level ("Required", "Very important", "Important", "Somewhat important", or "Not required"). This reflects how critical the criterion is for the dataset's intended use and ensures alignment with the fitness-for-use perspective.

The Step-by-Step Guide for the Data Quality Assessment Framework:

Step 1: Define the use
Clarify the analytical or policy objective of the study. As the framework is based on the fitness-for-use principle, the research goal helps determine which criteria are most relevant and under what conditions they should be applied.

Step 2: Specify the dataset characteristics
Identify the key characteristics of the dataset. In addition to the research goal, the type of dataset also helps to determine which criteria are most relevant and under what conditions they should be applied. We ask the user to specify the mapping method, scope, resolution and the data sources.

Step 3: Define pass conditions
For each diagnostic criterion, determine what constitutes acceptable quality in the given context. Specify when a criterion is considered sufficient to "pass", and assign the weight each criterion carries in the overall assessment: Required, Very important, Important, Somewhat important, or Not required.

Step 4: Assess the criteria
Evaluate the dataset against each diagnostic criterion. Report the results and assign one of the following outcomes: passed, not passed, or unknown (when information is insufficient).

Step 5: Interpret and conclude
Combine the individual results and their relative importance to derive an overall evaluation. Summarise the findings and provide a conclusion or recommendation on the dataset's suitability for the defined use.

Assessment form:
The framework is operationalised through an assessment form that records the intended use, dataset characteristics, pass conditions, importance levels, assessment outcomes, and the final conclusion. The form illustrates how the framework is applied in practice.

Interview: structure and questions
The interview will take approximately 30 minutes and is structured into three main parts. Each part focuses on a different component to ensure that the discussion remains clear and targeted.

1. *Dataset characteristics*
   We will begin with the dataset characteristics. You will be asked to reflect on whether these characteristics are clear, complete, and representative of how you typically distinguish between datasets in practice. You will also be asked whether you have experienced situations where you decided to reject a dataset because certain characteristics were not acceptable.

2. *Diagnostic criteria*
   We will then review the nine diagnostic criteria. The focus is on whether these criteria reflect your experience with data quality challenges, whether any criteria seem unclear, missing, redundant, or unexpectedly broad or narrow. You will also be asked if you have an example from your own work where specific criteria became particularly important, and which criterion you are most willing to compromise on in a situation where a dataset must be used because no suitable alternative is available. In addition, we will talk about whether you have encountered, or can imagine, situations where a client or end user rejected conclusions based on one of the diagnostic criteria.

3. *The assessment framework*
   Then, we will discuss the framework itself. You will be asked about the clarity and logic of the five-step structure, the feasibility of applying the framework in real-world evaluations, and whether any components seem difficult to use or incomplete.

The interview concludes with an open invitation to share any additional thoughts or suggestions.

Chapter F. Transcripts

F.1 Transcript expert A

Interviewer (Emma): Nou ja, je weet ongeveer wat ik aan het doen was. Dus ik ben aan het kijken, had je het documentje door kunnen lezen? Fijn. Ik ben dus een beetje aan het kijken naar datakwaliteit in de database voor supply chain mapping. Zou jij misschien kort willen zeggen wat je naam is, wat je precies doet, en ja, dat.

Geïnterviewde (A): (00:30) Ja, ik ben [naam]. Ik werk nu als consultant/data-analist voor NMO-projecten, vooral voor twee projecten veel met data. Dat is Materiaaldossiers, waar we veel kijken naar de handelsstromen tussen verschillende landen, en bijvoorbeeld tussen China en de rest van de wereld en Europa en de rest van de wereld, afhankelijk van de dataset die beschikbaar is. En de Landenfiles is gefocust op Canada, en dat is gefocust op Canada en Nederland. Wat kunnen we voor elkaar betekenen in de handel, maar ook: waar heeft Canada ten opzichte van de wereld bijvoorbeeld meer export voor?

Interviewer (Emma): (01:07) Ja, interessant. Dan als eerste: ik neem het even op en dan stuur ik je na het interview een korte samenvatting van wat ik denk dat eruit is gekomen. Dan kan jij het nakijken of je het ermee eens bent. Ik weet niet of jij het documentje voor je hebt?

Geïnterviewde (A): (01:33) Ja, ik had voor de vragen die je wilde bespreken al wat dingen opgeschreven.

Interviewer (Emma): (01:36) Ja, oké, top. Helemaal goed. In mijn onderzoek heb ik natuurlijk wat karakteristieken van datasets onderscheiden. En als eerste wil ik vragen: als je daar naar kijkt, denk je dat dit een beetje overeenkomt met hoe jij het in de praktijk ziet?

Geïnterviewde (A): (01:57) Ja, met de karakteristieken bedoel je die mapping method, scope en resolution?

Interviewer (Emma): (01:59) Klopt, mapping method, scope, resolution en data sources.

Geïnterviewde (A): (02:04) Ja, ik vond het best grappig, want ik denk niet per se... Ik denk gewoon: wat willen we? Wat is de best mogelijke data, en dat gebruik je. Dus zo doen we dit eigenlijk niet, gewoon omdat je ook de luxe niet hebt om het zo te doen. Maar ik las het door en toen dacht ik: ja, dit is eigenlijk wel precies wat je doet. Vooral met die mapping dacht ik: oh ja, wij doen eigenlijk alleen met die empirically collected datasets. Want ja AI, niet zo echt. Nou ja, ik denk wel dat dat steeds meer gaat zijn. Maar bijvoorbeeld ook binnen TNO, minder dan bij andere, minder wetenschappelijke plekken. En verder met scope en resolution: ja, dat is precies hoe we het bekijken.

Interviewer (Emma): (02:50) Oké, dus er is niks waarvan je dacht: dit ontbreekt nog?

Geïnterviewde (A): (02:57) Nee.

Interviewer (Emma): (02:59) Nee is ook een antwoord. Heb je bijvoorbeeld ooit een keer een dataset afgewezen omdat een van deze karakteristieken niet waren zoals je had gewild? Ik weet niet of jij in die positie zit.

Geïnterviewde (A): (03:20) Vooral die projecten die ik nu doe, zijn doorgewerkt op methoden die al eerder ontwikkeld waren, waar ik niet bij was. Dus ik weet niet precies welke afwegingen toen zijn gedaan. Ik weet wel dat het AI-stukje dat ik net al benoemde, daar wordt nu steeds meer mee gedaan, en wij zijn daar best huiverig over, omdat je dan gewoon niet precies weet wat je voor data hebt, hoe je dat moet verifiëren en of je dat mag gebruiken. Dus dat doen we niet. Er komt wel eens wat langs dat je denkt: dat is misschien wel handig, maar dat wijzen we dan af. En bijvoorbeeld met scope en resolution: we gebruiken ook de dataset van de Chinese douane. Dus dat is voor sommige stukken heel interessant, maar dat komt bijvoorbeeld niet overeen met... Als de Chinese douane zegt: dit gaat vanuit China naar Europese landen, dan zegt de Europese data "vanuit China naar Europa" vaak iets wat niet overeenkomt. Dus we kijken per stukje wat je aan het onderzoeken bent: oké, dit klopt niet. Welke neem je dan als waarheid? Of ga je ergens in het midden zitten? Dus je wijst het niet af.

Interviewer (Emma): (04:32) Maar je probeert het wel te valideren met verschillende data.

Geïnterviewde (A): (04:34) Ja. En ik denk dat dat wel een beetje te maken heeft met scope en resolution. We gebruiken een dataset Comext en dat is van wat de EU binnenkomt. Maar China zegt wat naar specifieke landen gaat. Dus daar zit een verschil in scope en resolution. Dus als je dat op elkaar gaat leggen.

Interviewer (Emma): (04:51) Ja, ik snap wat je bedoelt.

Geïnterviewde (A): (04:52) Het is niet per se dat we het afwijzen, maar het is wel dat we...

Interviewer (Emma): (04:55) Dat je andere methodes zoekt om het toch anders te doen.

Geïnterviewde (A): (05:00) Ja, of misschien dat we bij sommige handelscodes zeggen: oké, dit stukje geloven we niet. Dus dit klopt waarschijnlijk niet wat we in de data zien met wat we verwachten. Dus dat stukje wijzen we dan wel af.

Interviewer (Emma): (05:13) In plaats van dat je een hele dataset afwijst. Oké. Dan zou ik het graag willen hebben over de criteria. Eigenlijk weer een beetje dezelfde vraag: wat vind jij ervan? Komen deze negen criteria overeen met de dataproblemen die je in het echt tegenkomt?

Geïnterviewde (A): (05:36) Ja, toen ik erdoorheen keek dacht ik: dit is echt een hele goede lijst. Maar ook wel heel erg soort van wishful thinking, want je had erbij gezet dat je het gaat beoordelen met required of belangrijk, of iets in die zin. Toen dacht ik al: als we dit zouden doen voor de datasets die wij gebruiken. Als je het van tevoren doet met wat je zou willen en het vervolgens gaat gebruiken, dan weet ik niet hoeveel erdoorheen gaat komen. Maar theoretisch vind ik het heel goed. Ik heb ze er ook even bij gepakt. Wil je ze gewoon allemaal afgaan?

Interviewer (Emma): (06:11) Ja, is goed.

Geïnterviewde (A): (06:14) Bijvoorbeeld data fit is superbelangrijk, maar bij gebrek aan beter, wat ik net al zei, daar doe je het mee. Dus dat zou ik eigenlijk altijd een van de belangrijkste vinden, maar daar heb je niet zoveel iets aan. Hetzelfde voor accuracy. Wat ik net zei: als je dan de Chinese hebt en de Comext dataset, ja, dat is allebei niet accuraat, maar je weet niet wat je ermee moet.

Interviewer (Emma): (06:39) Ja, precies.

Geïnterviewde (A): (06:40) Van tevoren zou je misschien zeggen: oké, dit is heel belangrijk. Maar in de praktijk doe je het ermee. Completeness?

Interviewer (Emma): (06:53) Completeness, ja.

Geïnterviewde (A): (06:56) Ik heb nog niet ervaren dat je een complete dataset hebt.

Interviewer (Emma): (07:00) Nee.

Geïnterviewde (A): (07:01) Dus daarvan dacht ik: ik weet niet in welk scenario dat is?

Interviewer (Emma): (07:08) Ik denk dat het meer is, ik zal het even verduidelijken. Ik zie het meer als: Je zou willen dat die iemand is die er al een beetje in zit. Dus dat je bijvoorbeeld inderdaad niet zegt: het moet 100% compleet zijn, maar in ieder geval compleet genoeg.

Geïnterviewde (A): (07:27) Ja, dat je wel de gaten kan vullen. En uiteindelijk denk ik wel: het zou in theorie moeten kunnen om een complete dataset te hebben.

Interviewer (Emma): (07:37) Ja, maar dat is niet zo makkelijk blijkt wel.

Geïnterviewde (A): (07:41) En consistency vind ik iets waar ik heel vaak tegenaan loop. Bijvoorbeeld die HS codes veranderen best veel over de jaren. Sommige veranderen op andere tijdspannen, en dan is zowel de code veranderd over de jaren, en soms nemen ze dan ineens een bredere code of juist een minder brede code. Dus dat is iets waarvan ik denk: waarom doe je dat? Er zullen vast goede redenen achter zitten.

Interviewer (Emma): (08:05) Ja, dit is wel een van de grootste problemen inderdaad die mensen tegenkomen. Ik heb het het hier wel met meer mensen over gesproken, en die zeiden ook dat ze het over tijd veel zien veranderen, en soms zelfs met terugwerkende kracht.

Geïnterviewde (A): (08:21) En waarom doen ze dat?

Interviewer (Emma): (08:24) Ik denk omdat de soorten producten zoveel veranderen. Dan proberen ze er toch een soort consistentie in te zetten, maar daardoor wordt het weer anders. En het is ook verschillende databronnen samenvoegen die net wat anders aanhouden. Maar het is

inderdaad wel een punt waarvan mensen denken: dit is wel belangrijk.

Geïnterviewde (A): (08:46) Ja, en hoe moeilijk is het om het niet te doen?

Interviewer (Emma): (08:50) Het is blijkbaar heel lastig, ja.

Geïnterviewde (A): (08:51) Grappig om te horen dat het wel een echt probleem is.

Interviewer (Emma): (08:53) Ja, zeker.

Geïnterviewde (A): (08:57) Met die timing, die vijfde: je hebt bijna nooit recente data. Dat is lastig, want zeker nu met al die exportrestricties. De Chinese handelsdata is wel heel recent. Dat is eigenlijk de enige.

Interviewer (Emma): (09:09) Hoe recent ongeveer?

Geïnterviewde (A): (09:10) Tot vorige maand.

Interviewer (Emma): (09:11) Oh, dat is best recent.

Geïnterviewde (A): (09:12) Ja, november kan je nu al downloaden. Terwijl alle anderen dat totaal niet zijn. Waardoor je, als je dan dus gaten hebt of wil vergelijken het lastig is. [naam] is nu ook bezig met een Early Warning System, dat ze willen voorspellen wanneer er dingen gaan gebeuren. Dat staat een beetje los hiervan, maar vooral met beleid wil je natuurlijk zoveel mogelijk de meest recente invloeden zien. Dat is wat wij nu af en toe doen. Het is een beetje open end, omdat je de data gewoon niet hebt.

Interviewer (Emma): (09:52) Want vaak is het zeker eens per jaar dat die verslagen uitkomen.

Geïnterviewde (A): (09:57) De BACI-data, volgens mij tot 2022.

Interviewer (Emma): (10:03) Dat is niet heel recent.

Geïnterviewde (A): (10:05) Oké, ik durf daar even niet precies iets over te zeggen.

Interviewer (Emma): (10:07) Maar wel zeker iets langer dan een jaar.

Geïnterviewde (A): (10:08) Ja, maar in ieder geval niet een paar maanden, dus dat is echt onhandig. Representativeness snapte ik niet helemaal wat je ermee bedoelde?

Interviewer (Emma): (10:18) Dat is meer dat je bijvoorbeeld niet alleen data uit Europa en Azië hebt, maar bijvoorbeeld dat de hele wereld wordt gerepresenteerd. Of dat je alleen publieke bedrijven hebt in plaats van private bedrijven, omdat die data makkelijker te verkrijgen is. Daardoor krijg je een onrealistische balans van de wereld.

Geïnterviewde (A): (10:43) Ja, precies.

Interviewer (Emma): (10:46) Dit wat je hebt gelezen hebt zijn ook wat kortere uitleggen natuurlijk.

Geïnterviewde (A): (10:48) Ja. Oké, ik weet niet helemaal hoe ik die dan moet interpreteren als ik kijk naar de datasets die ik gebruik. Want als je het hebt over handelsstromen, dan kijk je specifiek naar landen. Dus als je de Chinese dataset hebt, dan is dat natuurlijk China.

Interviewer (Emma): (11:06) Ja, ik zou dan meer denken: welke data wordt publiek vrijgegeven, bijvoorbeeld volgens de wet. En sommige data van private bedrijven hoeft helemaal niet publiek te zijn, en daardoor kan je een disbalans krijgen.

Geïnterviewde (A): (11:26) Ja, dat is trouwens wel heel relevant. Wij vragen af en toe informatie op van het CBS en dat geven ze vaak wel, maar ook niet aan iedereen, maar omdat we samenwerken krijgen we dat soms wel. Maar wij weten als Nederland best weinig over wat er allemaal binnenkomt, omdat dat qua privacy niet wordt vrijgegeven. Maar [namen] waren vorige week in Korea en daar zijn ze veel bezig met hetzelfde als wij. En zij hebben veel meer data, omdat door andere wetgeving de Koreaanse CBS veel meer verzamelt en dat ook allemaal deelt. Daardoor kunnen ze veel meer, omdat zij veel representatievere data hebben. En dat is in Europa en Nederland, door privacy en zo, wel een beetje minder.

Interviewer (Emma): (12:14) Ja, precies.

Geïnterviewde (A): (12:16) Transparantie is sowieso een heel groot probleem. Want de Chinese douane kan exportrestricties omzeilen. Die douane zit daar gewoon. Het is allemaal een beetje natte vinger werk van hoe noemen we het? Dus dat is met handelsdata een heel groot probleem. Accessibility en comparability is ook een groot probleem. Omdat je het niet altijd goed kan vergelijken, omdat het op cruciale vlakken anders is. Je hebt ook gewoon heel veel niet openbare dingen. Je hebt bijvoorbeeld ook het data science team, zij zijn bezig met een dashboard maken met FactSet en andere dingen. Ik denk dat zij hier wel meer over kunnen vertellen. Ik denk dat die sowieso top zijn. Die laatste is lastig: we checken in ieder geval hoe we de dataset verwerken. Die kunnen we wel goed valideren, maar de dataset zelf...

Interviewer (Emma): (13:32) Ja, je hebt gewoon niet heel veel vergelijkings materiaal.

Geïnterviewde (A): (13:37) Ja. Vaak is het gewoon bekijken of het ongeveer overeenkomt.

Interviewer (Emma): (13:45) Dat zijn wel nuttige dingen die je zegt. Als je naar dit lijstje kijkt: zijn er nog dingen waarvan je denkt dat het ontbreekt?

Geïnterviewde (A): (13:56) Nee, want je kijkt echt naar de dataset en niet naar wat je met die dataset doet. Maar vaak is dat in de praktijk handig. Nee, ik denk dat dit wel een vrij compleet plaatje is.

Interviewer (Emma): (14:15) Oké. Je noemde het net al een beetje, maar als we het wat concreter moeten maken. Heb je een specifiek voorbeeld van een criterium dat een keer heel belangrijk was? Je noemde bijvoorbeeld al consistentie met die codes. Of zeg je, er springt er eentje echt bovenuit die je vaak ziet, is dat dan die consistentie?

Geïnterviewde (A): (14:44) Waar ik nu veel tegenaan loop is dat de codes, bijvoorbeeld dat je een handelscode hebt die verschillende materialen bevatten. Dus dat is denk ik transparantie of consistency?

Interviewer (Emma): (15:02) Ja, een beetje van beide misschien.

Geïnterviewde (A): (15:09) Ja, want het is best vaak dat je er eigenlijk niet echt iets over kan zeggen. Dan zie je dat een bronland Brazilië is. Je hebt zo'n handelscode met gallium, niobium en nog een paar. En bronland Brazilië is het grootste. Dat is blijkbaar algemeen bekend dat niobium vooral uit Brazilië komt. Dus dan zegt die hele handelscode niks over al die andere materialen. Waardoor je eigenlijk niks over bijvoorbeeld gallium kan zeggen. En dat is ook met die handelscodes die af en toe veranderen. Dan zou ik zeggen: verander het dan naar eentje voor gallium en eentje voor niobium. Hoe je het verdeelt, zeg maar.

Interviewer (Emma): (15:52) Gewoon specifieker.

Geïnterviewde (A): (15:54) Ja en verander het dan niet ook nog de hele tijd. Want soms pakken ze er poeder bij en waste en scrap, en vervolgens veranderen ze het weer terug naar die losse. Waardoor je die hele vergelijking over de jaren niet goed kan maken.

Interviewer (Emma): (16:06) Dan heb je er eigenlijk bijna niks meer aan, omdat je niet op een specifiek product uitkomt.

Geïnterviewde (A): (16:12) Ja. Ik weet niet onder welk kopje je dat zou scharen, want het is een combinatie van.

Interviewer (Emma): (16:24) Ik zou het meest zeggen: consistency. Maar als er heel veel transparantie zou zijn in hoe het is veranderd, dan zou het ook beter zijn. Het is ook een beetje hoe je het verwoordt.

Geïnterviewde (A): (16:37) En ik denk dat accuracy ook echt een ding is. Want soms, als er uit twee verschillende iets heel anders komt. Dan ga ik aan alles twijfelen daarna.

Interviewer (Emma): (16:50) Logisch.

Geïnterviewde (A): (16:51) Dus het komt niet per se heel vaak naar voren, maar vaak genoeg dat je het in je het de hele tijd in je achterhoofd houdt. En data fit trouwens ook. Bijvoorbeeld granularity, dat is dan met dat ze misschien te veel op een hoop gooien. Wat ik net zei.

Interviewer (Emma): (17:12) Ja, precies. Het is een beetje hoe je het verwoord, want dan kan je

ook onder andere dingen plaatsen. Oké, dankjewel. Dan nog een laatste vraag. Heb je wel eens gehad dat een klant of opdrachtgever zei: de conclusies die jullie uit deze dataset hebben gehaald, vind ik niet goed, of ik verwerp dat? / Op basis van een van die criteria, dat ze zeggen: ik geloof niet wat jullie eruit hebben gehaald, omdat er iets niet klopt aan de dataset?

Geïnterviewde (A): (17:57) Ik denk het niet. Want wij zijn best voorzichtig, dus we doen ook vaak disclaimers. Toevallig gisteren nog: dat die chinese en Comext dataset niet overeen kwamen. En het was een code die zowel germanium als zirconium had en je dus ook niks over germanium kan zeggen. Eerst hadden we het weggelaten, omdat je het liefst die dubbelen niet in wil. Maar toen was het wel onder de exportrestricties dus het is wel volledig om mee te nemen. Dus heb ik er heel veel disclaimers omheen gezet: we kunnen eigenlijk geen conclusies trekken. Die en die bron hebben besloten het niet mee te nemen, maar die bron doet het wel. Voor de volledigheid nemen we het wel mee, maar neem het met een korreltje zout. Dus dan is het heel duidelijk dat we iets geven, maar we zeggen deze punten expliciet bij. Dan mogen zij zelf de conclusie trekken wat ze ermee doen.

Interviewer (Emma): (19:05) Je zegt het eigenlijk uit jezelf al.

Geïnterviewde (A): (19:07) Ja, dat doen wij eigenlijk met alles. Dat is goed, maar daardoor kunnen ze het ook niet zo goed verwerpen. Want als je dat niet doet, dan hadden ze prima kunnen zeggen: dit klopt niet, dit slaat nergens op. Maar omdat je het zo inpakt en ook veel refereert naar andere bronnen. Dus we zeggen bijna nooit echt conclusies als we dat niet ergens anders deels kunnen bevestigen. Dus dan nee, maar dat komt wel echt daardoor.

Interviewer (Emma): (19:35) Ja, precies. Dan wil ik graag door naar het volgende onderdeel. Dat gaat over het hele framework als geheel en het formuleren van die pass conditions, en de praktische toepassing van deze theorie. Als eerste: als je naar dit geheel kijkt, denk je dat dit houvast biedt om het echt in de praktijk toe te passen?

Geïnterviewde (A): (20:11) Ja en nee. Ja, omdat ik vind dat het er heel goed uitziet. Ik heb het idee dat het vrij compleet is. En als je er iets mee wil, is dit een duidelijke manier om het stapsgewijs te doen. Als je kijkt naar de praktijk waar wij nu mee bezig zijn: je hebt niet echt de luxe dat je kan zeggen ja of nee tegen een dataset. Dus ik weet niet of wij dit zo zouden gebruiken. We zouden het misschien eerder achteraf doen, om te zeggen: oké, misschien is de volgende keer deze dataset beter voor dit doeleinde en die voor dat. In plaats van dat wij aan het begin, met een onderzoeksvraag, dit doorlopen en dan besluiten: deze dataset wel of niet. Want als je op basis van deze criteria besluit dat het niet goed is, dan heb je niks.

Interviewer (Emma): (21:01) Dus meer dat je zegt: er zijn gewoon niet veel opties. Je doet vaak: wat is het beste van wat ik heb?

Geïnterviewde (A): (21:13) Ja, en dan probeer je de punten waar ze niet goed op scoren te compenseren door te vergelijken met iets anders, of te valideren. Maar dan probeer je het af te vangen, in plaats van het afschuiven op een ja of nee voor de dataset. Ik kan me wel voorstellen dat als je meer datasets hebt, of een onderzoeksvraag waarbij je meer keuze hebt, dat het dan wel goed is. En op zich de toepassing om het achteraf toe te passen, om te kijken wat hebben we gedaan en

hoe wetenschappelijk onderbouwd is het, vooral daarvoor kan het voor ons vooral goed voor zijn. En om te onthouden voor de volgende keer, voor iemand anders: dit is hiervoor goed en daarvoor goed.

Interviewer (Emma): (22:01) Ja. Dus misschien bijna meer als een disclaimer achteraf: dit had beter gekund, dit ging minder goed. Logisch. Als we specifiek naar de stappen kijken: zijn er onderdelen waarvan je zegt dat je die lastig vindt in de praktijk? Bijvoorbeeld het opschrijven van pass conditions, importance levels bepalen, karakteristieken ontdekken, of criteria beoordelen?

Geïnterviewde (A): (23:03) Welke dan het moeilijkste is op zo'n individuele manier?

Interviewer (Emma): (23:04) Ja, waar je het snelst tegenaan loopt, of wat niet logisch is, of misschien iets in de volgorde.

Geïnterviewde (A): (23:12) Ik vond de eerste heel goed. Het is überhaupt goed om daarmee te beginnen, dat moet je gewoon doen. Die characteristics doe je denk ik automatisch.

Interviewer (Emma): (23:31) Ja, onbewust doe je dat sowieso.

Geïnterviewde (A): (23:33) Ja, want anders weet je niet of je de vraag die je bij stap 1 hebt opgesteld wel kan beantwoorden. Dus die is sowieso goed. Die pass conditions zijn denk ik het lastigst. Want als ik het tegen onze eigen datasets aan leg, dan denk ik: oh, dan scoort die niet zo goed. Maar je gaat hem toch gebruiken, want je hebt geen alternatief. Dus dan ga je het een beetje aanpassen, dat je denkt: zo slecht is die ook niet, zodat die beter scoort ofzo.

Interviewer (Emma): (24:01) Ja, dat je een soort bevooroordeeld die pass conditions gaat invullen.

Geïnterviewde (A): (24:05) Omdat er in ons geval eigenlijk niet uit kan komen dat we het niet kunnen gebruiken. Dus ik snap dat het heel goed kan zijn, maar in de praktijk zie ik die een beetje niet zo serieus, of niet zo eerlijk toegepast. Daarnaast denk ik ook dat lastig is: Veel wat wij doen binnen TNO verandert over de tijd. Dus hoe de eerste materiaaldossiers zijn gedaan, is nu alweer anders. Dus ik kan me voorstellen dat de dataset in die eerste varianten beter of slechter scoorde, maar met voortschrijdend inzicht het nu ook weer anders wordt. En dat heb je in onderzoek sowieso wel, waardoor het ook lastig is om het een harde pass te geven.

Interviewer (Emma): (24:58) Ja dus om vanaf het begin te zeggen: dit is ons doel en dat blijft het voor de rest van de jaren.

Geïnterviewde (A): (25:05) Ja. Dus ik heb het idee dat je daar flexibel mee omgaat. Assess the criteria, wat was dat ook alweer?

Interviewer (Emma): (25:15) Dat je de dataset toetst op die criteria. Ja dus eerst zeg je: het moet voldoen aan dit. En stap 4 is daar naar kijken.

Geïnterviewde (A): (25:28) Ja, ik denk dat dat in de praktijk door elkaar gaat lopen. Je wil dat

die moet passen, dus dan ga je hem anders behandelen. Dus ik denk dat stap 3 en 4 de meeste uitdaging heeft.

Interviewer (Emma): (25:44) Dat had ik ook wel verwacht. Het is inderdaad misschien meer een hulpmiddel dan een harde eis: hier komt iets slechts uit, dus dat gebruiken we niet.

Geïnterviewde (A): (25:59) Ja, en ik denk dus wel dat het voor je methodologie, en dat is bij het TNO onderzoek heel belangrijk, dat dit wel echt een goede toevoeging kan zijn daarvoor. Dus meer om te verantwoorden wat je aan het doen bent. Daar zie ik wel echt het nut van in.

Interviewer (Emma): (26:09) Ja. Als ik het samenvat: dus vooral omdat je gewoon heel weinig opties hebt van goede datasets. En je kan aan het begin niet volledig je doel bepalen van een dataset. Je weet misschien wat je het eerste jaar ermee gaat doen, maar vaak koop je zo'n dataset met de intentie voor meerdere jaren, en dan wordt het doel breder dan aan het begin.

Geïnterviewde (A): (26:48) Ja. En zeker omdat we binnen het NMO nog best wel aan het opstarten zijn. Er komen steeds meer datasets bij, omdat het dan is: dat kunnen we voor dat project gebruiken. En de vraag verandert ook. Nu met de paar materialen die we doen, gaat het alleen maar over Chinese exportrestricties. Dus hebben we de Chinese douane nodig. Maar bij de eerste materialen was dat nog niet zo, want toen waren er nog geen exportrestricties dus dan ging je daar nog niet naar kijken. Dus in het begin hadden we misschien niet gezegd dat we de Chinese douane gaan gebruiken, en later wel. Dus je doel verandert, maar ook je onderzoeksrichting. Hoe meer tijd je hebt, hoe meer je erbij pakt.

Interviewer (Emma): (27:34) Heel logisch. Ik heb hier zeker wat aan. Zou je zeggen dat er nog een specifieke stap mist?

Geïnterviewde (A): (27:57) Ik was benieuwd naar de assessment form. Hoe ziet dat eruit? Ik ben heel benieuwd.

Interviewer (Emma): (28:03) Ja, het is eigenlijk heel simpel. Het is meer een visualisatie zodat het duidelijker is. Je zou eerst kort omschrijven: hoe ziet je dataset eruit, waarvoor willen we hem gebruiken? En daaronder heb je al die criteria, en dan zou je opschrijven: die criterium moeten voldoen aan, bijvoorbeeld 80% accuracy halen we waarschijnlijk nooit. En dan is het bijvoorbeeld very important voor ons doel. En daaronder heb je weer alle criteria, en dan schrijf je het resultaat van de assessment op. En als je dat vergelijkt met elkaar, dan zou ik zeggen: het krijgt een pass of niet. En daaruit trek je een conclusie. Het is gemaakt om flexibel te zijn en juist rekening te houden met dat het niet altijd even goed is. Dat hoorde ik van veel mensen: het is behelpen met wat we hebben. Ik probeer het wel specifiek te maken zonder flexibiliteit te missen. In de tekst staan ook wat specifiekere richtlijnen voor hoe je het bepaalt, maar het blijft erg flexibel. Dit is ook waarom ik in gesprek wilde gaan met experts. Eigenlijk ook: wat zijn dan misschien mijn disclaimers die ik erbij moet geven?

Geïnterviewde (A): (29:50) Die forms ziet er wel goed uit, want als je het leest dan denk je: dit zijn allemaal goede punten, maar hoe krijg je dat in overzicht? Dus dit is wel heel handig.

Interviewer (Emma): (30:00) Ja. En bij die importance heb ik ook geprobeerd het te visualiseren, dat de criteria in sommige situaties belangrijker of minder belangrijk zijn. Dus ik probeer dat zoveel mogelijk te doen.

Geïnterviewde (A): (30:19) Ja, wat goed. Dit ziet er goed uit.

Interviewer (Emma): (30:22) Dankjewel. Ik wilde alleen wat ik naar jullie opstuurde niet te lang te maken, dus vandaar dat dit er niet instond.

Geïnterviewde (A): (30:30) Ja, dat moet ik wel zeggen. Ik was aan het doorlezen en dacht: aan het eind van de dag had ik zoiets van: dit zijn echt allemaal woorden.

Interviewer (Emma): (30:36) Ja, ik probeerde het zo kort mogelijk te houden.

Geïnterviewde (A): (30:42) Ja, maar leuk om te zien.

Interviewer (Emma): (30:43) We zijn er al redelijk doorheen. Mijn laatste vraag: heb je nog iets in het algemeen dat je wil toevoegen, tips?

Geïnterviewde (A): (31:00) Ik had op het einde opgeschreven: uiteindelijk is het gewoon alleen maar wat het best mogelijke is, en daarmee doe je het. Dat is misschien niet iets waar jij in je onderzoek veel aan hebt, maar dat is wel hoe ik het zie.

Interviewer (Emma): (31:11) Dat kan ik zeker meenemen. Dat je het meer in dat licht moet zien, in plaats van dat je zegt: deze dataset moet perfect zijn of echt voldoen aan iets. Meer de vergelijking: deze is heel slecht en deze is goed genoeg voor wat we willen bereiken.

Geïnterviewde (A): (31:34) En ook als je meerdere datasets bij elkaar pakt, dan heb je misschien minder gaten als je ze over elkaar legt. Maar dat is lastig, want je kan in die assessment niet meerdere datasets naast elkaar leggen of beoordelen. Maar in de praktijk zouden we dat wel doen. Dan zouden we drie datasets allemaal invullen, en dan hoop je dat de ene misschien heel slecht scoort op een punt, maar de andere juist heel goed op dat punt. Dat als je ze allebei gebruikt, heb je gemiddeld een goede score hebt.

Interviewer (Emma): (32:07) Ja, precies. Dus als je ze naast elkaar beoordeelt, kijk je of ze de zwakke punten van elkaar kunnen rechttrekken.

Geïnterviewde (A): (32:18) Ja. Dit is heel erg op één dataset. Maar in de praktijk doe je er meer dan één. En dat kan misschien nog meer toevoegen. Want dan moet je ook zoeken naar een dataset die op die paar punten goed scoort. Misschien is dat wel het meeste waar we het voor zouden gebruiken: bewust zijn van de gaten, en kunnen we die met een andere dataset opvangen?

Interviewer (Emma): (32:40) Ja, wat zijn de sterke en zwakke punten?

Geïnterviewde (A): (32:42) Ja. Dus niet per se: we gebruiken de dataset wel of niet. Meer: we moeten er nog een dataset bij hebben om het aan te vullen. Ik denk dat dat oprecht is waarvoor we dit in de praktijk echt zouden gebruiken.

Interviewer (Emma): (32:54) Ja, oké. Dat is geen slecht idee. Ik zeg wel vaak: je moet valideren met andere datasets. Maar je kan inderdaad ook heel specifiek twee of drie datasets naast elkaar leggen en kijken wat de sterke en zwakke punten zijn.

Geïnterviewde (A): (33:12) Ja, dat is eigenlijk ook wat we doen, maar dan zonder een framework, maar gewoon in de praktijk. Misschien moet je dat in je conclusie zetten of in ieder geval in je verhaal. Want daarvoor zou ik het echt nog gebruiken, binnen wat we doen.

Interviewer (Emma): (33:31) Oké. Heb jij nog iets?

Geïnterviewde (A): (33:35) Nee, dit was het. Maar wel leuk om te zien wat je allemaal hebt gemaakt.

Interviewer (Emma): (33:39) Dankjewel. Ik vond het heel nuttig. Als je er zelf aan zit, blijf ik de hele tijd denken: zou je dit wel gebruiken? Het is lastig om het echt in één strakke beoordeling te gieten. Maar waar je echt tegenaan loopt, en waar jij het echt voor zou gebruiken in je werk, dat is toch anders dan dat het er theoretisch mooi uitziet op papier.

Geïnterviewde (A): (34:10) Ja, ik ben ook benieuwd: hoeveel mensen ga je spreken?

Interviewer (Emma): (34:15) Vijf in totaal.

Geïnterviewde (A): (34:16) Ik denk dat mensen echt wel verschillende dingen zeggen. Ook interessant wat anderen zeggen.

Interviewer (Emma): (34:21) Ja, ik ben ook benieuwd.

F.2 Transcript expert B

Interviewer (Emma): (0:09) Nou, volgens mij doet hij het. Iedereen werd even gemute.

Geïnterviewde (B): (0:16) Ja, ik zat in het document al. Sorry, ik ben er weer.

Interviewer (Emma): (0:19) Is het gelukt om het document te openen?

Geïnterviewde (B): (0:32) Ja, ik heb hem voor me.

Interviewer (Emma): (0:33) Oh ja, helemaal goed. Als eerste zou ik je willen vragen om je kort voor te stellen: wat je precies doet binnen TNO. Ik weet het zelf al een beetje, maar toch.

Geïnterviewde (B): (0:48) Ja, ik ben [naam]. Ik ben teamlead van het Data Innovation Team dat zich bezighoudt met AI en data science bij de Geologische Dienst. We zitten binnen de Energy and Material Transition Unit van TNO. We werken veel samen met alle andere TNO-units. Een van de projecten waar we heel actief in zijn, is het Nederlandse Materialen Observatorium. Dat is een project dat geleid wordt door de Geologische Dienst en samen met TNO Vector uitgevoerd wordt. De samenwerking met TNO Vector gaat ook specifiek over supply chains. Ons team ontwikkelt de analyses, waarbij we verschillende databronnen bij elkaar brengen. Dus we maken de data-infrastructuur om alle data bij elkaar te brengen, en we maken daar AI-oplossingen bovenop om supply chains te analyseren.

Interviewer (Emma): (1:37) Ja, mooi. Dat is precies waar ik ook veel van heb gehoord. Voordat we beginnen: ik stuur na het interview een korte samenvatting van wat we hebben besproken, zodat je kan checken of ik het goed heb geïnterpreteerd. Als eerste wil ik het graag hebben over de dataset karakteristieken. Die staan onder het eerste kopje. Daarover zou ik willen vragen: als je naar die karakteristieken kijkt, dus mapping method, scope, resolution en data sources, komt dat overeen met hoe jij dat in je eigen werk ziet? Misschien maak je dat onderscheid niet zo direct, maar als je ernaar kijkt: ontbreekt er iets, of zijn er dingen die onverwacht overkomen?

Geïnterviewde (B): (2:37) Dus die dataset characteristics: kan je ze nog een keer noemen?

Interviewer (Emma): (2:42) Ja, vier: als eerste mapping method, dus empirically collected of AI, etc. en dan scope, resolution en data sources. Het staat als het goed is allemaal onder het eerste kopje.

Geïnterviewde (B): (2:58) Ja, ik zie het. Data sources. Wil je dat ik ze allemaal doorneem, hoe ik denk dat het eruit ziet voor elk?

Interviewer (Emma): (3:10) Ja bijvoorbeeld, of vooral of het compleet is en of het is hoe je het in de praktijk ziet.

Geïnterviewde (B): (3:21) Ik zal beginnen met de databronnen. We kopen datasets in, bijvoorbeeld van FactSet of S&P Global. Een paar andere: daar zijn we nog mee in gesprek, verschillende bedrijven die data kunnen aanbieden. Dat zijn serviceproviders die bijvoorbeeld data van het internet afschrapen en verzamelen. Ze proberen het zo veel mogelijk bij de bron te houden, maar ze aggregeren het ook op een bepaalde manier zodat het makkelijk te consumeren is. Wil je dit specifiek in supply chain-context?

Interviewer (Emma): (3:58) Ja, het liefst wel.

Geïnterviewde (B): (3:59) Voor supply chain: er is eigenlijk niet één databron die alles heeft om dit inzicht te maken. Dus je moet sowieso verschillende bronnen bij elkaar halen en zelf de combinatie maken om tot dat inzicht te komen van hoe die supply chain er echt uitziet. Sommige databronnen zijn gescrapet van het internet. Of sommige kopen zelfs in bij shippers of transportbedrijven. Dus daar hebben ze dan een afspraak mee: tegen betaling verzamelen ze data en dat verkopen ze dan weer door als onderdeel van hun service. We zijn zelf ook aan het kijken naar dat soort mogelijkheden: dat we zelf direct bij dat soort partijen ook data zelf inkopen, en dan meer specifiek

voor wat wij nodig hebben voor de analyse.
Dus de data is niet compleet, dat is zeker. Want wat gescrapet wordt van het internet betekent dat je alleen publieke data in die datasets hebt. Want in Europa en de EU wordt dat soort data niet publiek gemaakt. Dus hoe dan ook: wat we ook kopen van de markt, het is niet compleet. We moeten het hoe dan ook zelf verrijken met eigen databronnen. Dus we moeten binnen Europa, binnen de EU, afspraken gaan maken met douanes, shippers en andere partijen.
Om hun informatie aan ons te leveren voor de gezamenlijke vergaring van die inzichten over de supply chains voor Europa. Dat is een beetje de databronnen: we weten dat het niet compleet is en dat we zelf achter de data aan moeten gaan. Daarnaast gebruiken we ook technieken om de data te verrijken, bijvoorbeeld met AI, waarbij we interpreteren welke kritieke materialen er in producten zitten.

Interviewer (Emma): (5:32) Dus jullie gebruiken AI? Ik sprak net iemand, en die zei: wij zijn daar huiverig voor. Misschien is dat omdat jullie het zelf toepassen, maar niet per se datasets gebruiken waar AI al in zit. Begrijp ik dat goed?

Geïnterviewde (B): (5:39) Absoluut. Wij passen het zelf toe.

Interviewer (Emma): (5:55) Ja, precies.

Geïnterviewde (B): (5:56) Onze data innovaties binnen NMO zijn ook het toepassen van AI toe om data te verrijken en nieuwe inzichten te maken, maar ook om automatisering te doen.

Interviewer (Emma): (6:10) Ja, precies.

Geïnterviewde (B): (6:11) En soms, bijvoorbeeld omdat die databronnen gewoon niet compleet zijn, hebben we de context nodig van: als we in kaart willen brengen hoe kritieke materialen supply chains lopen, van bron tot bij onze bedrijven, dan moeten we eigenlijk weten in welke producten ze zitten. Die informatie zit niet in de databronnen die we tot nu toe hebben. Dus om toch inzicht te krijgen, doen we een AI-gestuurd onderzoek waarbij we interpreteren met AI: deze producten bevatten typisch deze set kritieke materialen. En dan kunnen we via die interpretatie door de supply chain tracken hoe een kritiek materiaal via een product terechtkomt in Nederland. Dus we gebruiken AI wel voor analyse en research, en integreren dat in de analyses.
Daarnaast doen we ook AI-analyses: met deze verschillende bronnen, als dit en dit samen gebeurt, dan is dat misschien een signaal dat we moeten oppikken.
Ik denk dat ik daarmee de scope and resolution ook wel heb meegenomen. En de method ook.

Interviewer (Emma): (7:18) Ja.

Geïnterviewde (B): (7:20) Dus AI-generated datasets zitten er zeker ook in. We verrijken echt datasets met AI-generated content.

Interviewer (Emma): (7:32) Dus als je kijkt naar dat onderscheid dat daar gemaakt wordt met de verschillende karakteristieken, dat geeft dan wel een compleet beeld?

Geïnterviewde (B): (7:38) Ja.

En empirically collected is ook nog iets. We hebben ook researchgroepen binnen NMO, dus echt onderzoekers die zelf data produceren. Dat komt ook weer terecht in wat wij analyseren. Dus dat is echt empirisch verkregen data door onszelf.

Interviewer (Emma): (7:58) Ja, oké. Dus jullie kopen niet alleen datasets in, maar verzamelen ook zelf data.

Geïnterviewde (B): (8:10) Ja.

Interviewer (Emma): (8:11) Oké. Dan iets naar beneden in het document. Vanuit problemen die ik in de literatuur en uit gesprekken met mensen heb gevonden heb ik negen criteria ontwikkeld waarmee je die dataproblemen kan samenvatten. Als je daar naar kijkt: omvat dit de problemen die jij in de praktijk ziet, of ontbreekt er iets, of is er iets onverwachts?

Geïnterviewde (B): (9:16) Ja, dit klopt wel. Ik zou niet nog een tiende kunnen bedenken.

Interviewer (Emma): (9:24) Ja, dat kan. Of bijvoorbeeld iets wat heel specifiek is of wat breder zou kunnen.

Geïnterviewde (B): (9:52) Misschien moet ik er nog wat langer over nadenken of ik nog iets kan bedenken.
Of vanuit een AI-toepassing perspectief, daar is het vooral belangrijk dat de data die je erin stopt accuraat en compleet is. Alleen weten we niet: wat is "compleet"? We kunnen alleen proberen zoveel mogelijk data te verzamelen en op basis daarvan zo goed mogelijk inzicht te maken. Maar het zal nooit compleet zijn. We zullen nooit alle data hebben die er is. Dus dat is altijd moeilijk af te wegen: is het compleet genoeg voor het beantwoorden van deze vraag? En daarnaast: transparantie en het kunnen verifiëren van de correctheid.
Dat is ook altijd heel belangrijk. Dus validation en verification. Ook als wij zelf iets maken, moeten we ook altijd kunnen uitleggen hoe dat inzicht tot stand is gekomen. Via welke transformaties, via welke samenvoegingen en cross-references van databronnen. We kunnen niet zomaar iets in een AI stoppen, dat dat wat uitspugt en zeggen: kijk, dit is het. Je moet het kunnen uitleggen.

Interviewer (Emma): (10:59) Ja, dus vooral in relatie tot hoe jullie AI toepassen, zijn die drie criteria extra belangrijk?

Geïnterviewde (B): (11:06) Ja. Accuracy, completeness, transparantie en validation. Die vier.

Interviewer (Emma): (11:15) Ja, precies.

Geïnterviewde (B): (11:17) En voor de rest, bijvoorbeeld representativeness, of timeliness and temporal consistency. Opzich is dat ook wel belangrijk, maar met al die dingen zijn met hoe wij AI oplossingen ontwikkelen, is het vooral: je brengt zoveel mogelijk data bij elkaar, en dan schonen we het op als onderdeel van de taak. We doen wel transformaties, die het harmoniseren, zorgen dat alles in dezelfde tijdlijn zit of dezelfde tijdstappen. Het helpt wel als alle bronnen dat al hebben, maar dat is eigenlijk nooit het geval. Dus we pakken wat er is. Wat beschikbaar is, dat gebruiken we.

Interviewer (Emma): (11:59) Ja, precies. Stel dat je een dataset moet gebruiken omdat er geen alternatief is: welk criterium zou je dan het minst belangrijk vinden, waar je het eerst concessies op zou doen? Misschien representativeness?

Geïnterviewde (B): (12:35) Waarschijnlijk data fit. Want we transformeren de data zodat het gewoon past in hoe we het allemaal aggregeren. Dus het maakt minder uit hoe het is geformat, want we herformatteren het toch.

Interviewer (Emma): (12:53) Ja, precies. Dus vooral dat de data klopt, en dat er genoeg is om er iets mee te doen en te valideren.

Geïnterviewde (B): (13:01) Ja, valideren is belangrijk.

Interviewer (Emma): (13:08) Ik weet niet of jij in deze situatie zit, maar: hebben jullie wel eens gehad dat jullie een conclusie uit data trekken en dat een klant of opdrachtgever die verwerpt omdat hij zegt: de data waarop jullie dit hebben gebaseerd niet klopt, of vanwege een van deze problemen?

Geïnterviewde (B): (13:35) Dat zou kunnen, maar binnen NMO hebben we daar domeinspecialisten op. Wij helpen hen met deze inzichten te vergaren. En als dan hun klanten zeggen: dit klopt niet hierom en hierom, dan zouden wij dat via domeinspecialisten als feedback terugkrijgen. Dan moeten wij kunnen uitleggen: dit is toch echt wat eruit komt om deze en deze redenen. Als zij bedenkingen hebben bij een bron die wij gebruiken, dan moeten we onderzoeken of valideren of die bron correct is, waarom die misschien iets mist. En dan kunnen we daar misschien iets aan doen. Ik heb daar niet direct mee te maken, maar via domeinspecialisten komt dat wel bij ons terug. Alleen: ik heb dat tot nu toe nog niet gehoord.

Interviewer (Emma): (14:17) Oké. Dan wil ik het graag hebben over het framework. Ik heb een vijfstappenplan gemaakt, en daarbij een formulier (dat staat niet in het documentje) om deze negen criteria te testen op je dataset. Vind je deze vijf stappen voldoende houvast om het in de praktijk toe te passen? Of zijn er onderdelen waarvan je denkt: dit wordt heel lastig?

Geïnterviewde (B): (15:19) Laat me het even opnieuw lezen.

Interviewer (Emma): (15:21) Ja, geen haast.

Geïnterviewde (B): (16:01) Ja, dit is wel toepasbaar voor alle databronnen denk ik. Ik zit alleen te denken of het binnen het raamwerk dat wij in AI gebruiken, nog verder uit te breiden is. Wij redeneren bijvoorbeeld niet zozeer vanuit databronnen, maar vanuit use cases. “Define the use” zit daar dichtbij. Dus voor alle innovaties, inzichten en analyses die we als taak hebben om te ontwikkelen. Elke research en innovation die we doen, breken we op in use cases, dus stukjes technologie die we moeten ontwikkelen. En daarin beschrijven we: dit is het probleem, dit is de oplossing die we voorzien, en deze typen technologieen zijn waarschijnlijk nodig om ze op te lossen. Als je dat weet, dan kan je plannen hoe verschillende dingen die te ontwikkelen zijn binnen het

programma samenhangen, en dan kan je efficiënt prioriteren: wat doe ik eerst, wat daarna, wat ga ik hergebruiken. En data types en databronnen zijn daar één van de criteria binnen het beschrijven van een use case. Dus dat zit als onderdeel van de use case orchestration, zoals ze dat noemen. Maar we zijn tot nu toe niet op deze resolutie door gegaan om de databron echt te analyseren.

Interviewer (Emma): (17:25) Ja, precies.

Geïnterviewde (B): (17:29) We hebben de use case beschreven en we weten waarvoor we het nodig hebben. Dan kijken we naar verschillende databronnen die bepaalde inhoud hebben waarvan we denken: dat hebben we nodig. Bijvoorbeeld: deze databron heeft verschillende producten die verkocht worden van de een naar de ander. Als we weten dat het erin zit, gaan we kijken: hoe verbinden we dit met de rest, en brengt dat het inzicht dat we nodig hebben? Dus we kijken niet naar condities bij die criteria. Dit is conform de kwaliteit die we nodig hebben ofzo.

Interviewer (Emma): (18:01) Dus meer: het eindproduct moet voldoen aan dit, maar de bronnen die je daarvoor gebruikt beoordeel je zelf niet echt systematisch op zulke criteria? Maar je kijkt vooral: klopt het?

Geïnterviewde (B): (18:18) Ja. En ik denk dat we wel dit zouden moeten doen, deze vijf stappen. Maar misschien zijn niet alle negen criteria even relevant. Data fit zou ik bijvoorbeeld bijna nooit echt naar kijken, want we vinden altijd wel een manier om het te laten passen en te verbinden met de rest. Maar het is wel belangrijk dat we zouden moeten checken van: waar komt die data van S&P Global vandaan? Hoe compleet is dat? Eigenlijk met elke andere vendors of service providers zouden we dat moeten evalueren. Dat hebben we inderdaad nog niet eerder gedaan, en dat is wel goed.

Interviewer (Emma): (18:39) Ja, want de insteek is ook dat je zelf per criterium de importance zelf kan bepalen, en dat je ook kan zeggen: dit is voor ons niet relevant. Omdat ik het relevant wilde maken voor allerlei verschillende soorten datasets.

Geïnterviewde (B): (19:14) Jazeker, als een bron AI gegenereerde data heeft, dan zouden we dat meteen flaggen als questionable om te gebruiken in onze analyse.
Dus we kijken daar wel naar, op een high level, als specialisten in data-analyse en data science. Maar we hebben geen systeem dat we dit op dezelfde manier zouden evalueren voor alle bronnen. Eigenlijk zouden we dat wel moeten hebben. Ik denk dat dit heel nuttig is.

Interviewer (Emma): (19:43) Ja, dus ik ben vooral ook benieuwd naar in deze interviews van: Dus zouden jullie dit ook echt in de praktijk gebruiken? En hoe dan precies? In een ander gesprekje gaf zij aan dat het misschien bijna achteraf gebruikt zou worden om te kijken naar hoe goed was onze data.

Geïnterviewde (B): (20:16) Nee, ik zou dit echt van tevoren doen. Wel echt aan het begin. We hebben met al die use cases nu al in kaart gebracht: waar willen we naartoe ontwikkelen? Een soort intelligentie platform dat misschien ooit als service aangeboden kan worden. Maar daar komen allemaal verschillende databronnen bij elkaar. Dus om te weten hoe data bij elkaar komt en hoe het elkaar complementeert, moeten we weten: wat is inhoudelijk de kwaliteit van al die aanbieders van

die bronnen en de data zelf. En misschien bij het selecteren van welke bron wel en welke bron niet. Dan moeten we een systematische assessment kunnen doen: waarbij je vergelijkt, deze wel en deze niet, en om welke redenen.

Interviewer (Emma): (20:57) Ja, precies.

Geïnterviewde (B): (20:59) Anders ben je met verschillende criteria aan het vergelijken en besluiten welke bron wel in het uiteindelijke intelligentievoorziening komt en welke niet.

Interviewer (Emma): (21:10) Dus je ziet dit wel echt als iets dat in de praktijk gebruikt zou worden.

Geïnterviewde (B): (21:16) Ja en echt aan het begin van het hele proces.

Interviewer (Emma): (21:18) Dat was ook de intentie, dat het vanaf het begin gebruikt wordt.

Geïnterviewde (B): (21:24) En het kan ook aan het begin, want bij NMO hebben we van tevoren in kaart gebracht welke toepassingen van AI en innovatie en überhaupt van MNO er ontwikkeld moeten worden, en welke eerst en welke daarna ontwikkeld moet worden. Dus je weet voor welke use case welke bron wanneer er in moet komen. Nog voordat we zijn begonnen met de ontwikkeling.

Interviewer (Emma): (21:46) Dus je zegt eigenlijk dat het doel goed vaststaat van tevoren?

Geïnterviewde (B): (21:50) Ja, het doel staat goed vast.

Interviewer (Emma): (21:51) Ja. Ik had net een ander gesprek waaruit kwam dat het niet altijd even duidelijk was wat het doel was en het doel soms veranderde over de tijd. Jij zegt: het doel staat goed vast? Want dat is waar ik nu over nadacht, als het doel heel erg verandert: hoe handig is het dan om heel erg te focussen op deze dataset die gaan we precies gebruiken voor dit doel.

Geïnterviewde (B): (22:12) Ja, het doel kan altijd veranderen. En we hebben dan ook een systeem dat we herevalueren: is de roadmap van ontwikkeling nog accuraat, of moeten we dingen veranderen? Dus ook de databronnen die we voor ogen hadden voor die roadmap. Maar eigenlijk heb je altijd, voordat je begint met de innovatie, heb je een soort doel. Je weet waarvoor je het doet.
En het is prima als dat verandert: terwijl je door de ontwikkeling heen gaat leer je nieuwe dingen en dan kan je doel ook veranderen. En dat is prima. Dan blijkt misschien: deze bron is toch niet geschikt, we moeten een andere hebben. Hoe beslis je dan waarom je een andere bron neemt? Dan moet je toch echt een goede assessment gemaakt hebben van: het doel is veranderd, en daarom is deze bron op basis van deze criteria niet meer geschikt. Dus we moeten een andere bron hebben die wel voldoet aan de nieuwe criteria.
Dat betekent dat je bronnen die je al hebt en nog niet hebt, op dezelfde manier moet evalueren. En we hebben niet eerder een systeem eerder gebruikt zoals dit om dat te doen.

Interviewer (Emma): (23:30) Ja. En als je specifiek kijkt naar stap 3, waar je per criterium de

pass conditions opschrijft. Het staat hier natuurlijk wat korter opgeschreven. Maar denk je dat dat in de praktijk te doen is?

Geïnterviewde (B): (24:01) Ja, ik denk het wel. Toen ik door die negen criteria heen ging, begon ik al te bedenken: voor die bron zou ik dit cijfer geven, en voor die bron dat. Dus ja, ik denk het wel.
Niet elk criterium is dus even relevant. Data fit zou ik misschien geen cijfer geven, tenzij het later belangrijk wordt als de infrastructuur heel erg verandert. Per project kan het ook verschillen. Maar binnen NMO zou ik zeggen: vijf of zes van deze criteria zou ik altijd in een evaluatie van een databron willen doen, op dezelfde manier en met hetzelfde formulier. En dat zou ik doen voor elke keer dat we een gesprek hebben met een nieuwe provider.

Interviewer (Emma): (24:52) Oké fijn, dat is inderdaad waar ik achter wil komen: zouden jullie dat zo gebruiken. Heb jij nog tips of iets om toe te voegen als je hier zo naar kijkt?

Geïnterviewde (B): (25:29) Heb je ook iets van een workflow? Dat je op een makkelijke manier het proces kan doorlopen voor elke nieuwe evaluatie van een bron, en dat de outputs op dezelfde gestructureerde manier gebruikt kunnen worden. Dus elke keer als je een nieuwe bron zou evalueren, dat je dezelfde tool gebruikt met dezelfde workflow, en dat de output er altijd hetzelfde uitziet, zodat je makkelijk alle bronnen daarna kan vergelijken.

Interviewer (Emma): (25:58) Ja, ik heb ook een forms gemaakt die je op dezelfde manier kan invullen. En ik heb geprobeerd die importance te ondersteunen met een tabel: wanneer is iets heel belangrijk wordt en wanneer minder belangrijk. Dat staat niet allemaal in dat document, maar het is er wel. Ik heb het wel zo gehouden dat er flexibiliteit is, dus er is wel variatie, je moet bijvoorbeeld wel zelf de conclusie trekken. Ik kan even mijn scherm delen.

Geïnterviewde (B): (26:53) Ja, ik zie je scherm.

Interviewer (Emma): (26:58) Hier zie je de stappen. En dan de form: al deze vakjes zouden in het echt groter zijn, dit is meer ter illustratie. Maar op deze manier probeerde ik in ieder geval het invullen consistenter te maken: dat je hier de dataset omschrijft, en dan per criterium de pass conditions en importance opschrijft en dan het resultaat.

Geïnterviewde (B): (27:30) Ik bedenk nog iets. NMO is een goed voorbeeld hiervan. Dat verschillende programmalijnen vaak dezelfde data of informatie nodig hebben uit vaak dezelfde databronnen. Elk programmalijn is een soort eigen afdeling die op zoek gaat naar databronnen. Soms vinden ze dezelfde bron en gaan ze onafhankelijk dezelfde engagement aan. De kans kan zijn dat je dan misschien twee keer dezelfde licentie koopt vanuit twee verschillende programma's. Dus je wil dat ook in kaart brengen. De "define the use" stap moet ook zijn: wat is de toepassing van de data, dus binnen welk domein is het toe te passen, en wie wordt de eigenaar ervan?
Dat kan een project zijn, een persoon of een programmalijn, maar iemand moet eigenaar worden van die databron. Ook al is die is ingekocht of als service is aangenomen van externe. Net zo goed als dat een onderzoeksafdeling zelf data genereert: zij zijn de eigenaar van die data. Als een extern iemand dat inkoopt, dan moet iemand daar verantwoordelijk voor zijn. Dat echt in stap 1 definiëren

zou echt heel goed zijn.

Interviewer (Emma): (28:41) Ja. Ik zit even mee te denken hoe dit vorm zou krijgen in hoe dit eruit ziet.

Geïnterviewde (B): (29:08) Uiteindelijk gaat het erom dat iemand die data evalueert. Iemand is de eigenaar, iemand heeft de evaluatie gedaan. Dus andere mensen intern moeten weten: wie heeft deze assessment gedaan, en waarom, en vanuit welk perspectief. Iemand kan het vanuit supply chain-perspectief doen, maar misschien heeft iemand vanuit geologisch onderzoek die bron ook nodig. Degene die die eerste assessment gedaan heeft, heeft nooit vanuit een geologisch perspectief ernaar gekeken. Dus die assessment is dan misschien niet helemaal representatief voor een ander domein. En dan moet die assessment opnieuw, of in ieder geval aangevuld.

Interviewer (Emma): (29:44) Ja, dat is een slimme toevoeging. Nu staat er inderdaad niet wie dit heeft gedaan, dus je kan het niet achterhalen.

Geïnterviewde (B): (30:04) Precies. En als iemand die assessment gedaan heeft. Dan zijn ze eigenlijk een soort expert van die databron ook. Dus dan kunnen mensen ze ook over die assessment vragen. En dan misschien samen de assessment verrijken met inzichten vanuit een ander domein perspectief. Dan hoeft die assessment niet helemaal opnieuw, maar kan hij verbeterd of uitgebreid worden. Dat kun je alleen maar weten als je van tevoren weet wie het eerst heeft gedaan en vanuit welk perspectief. Dus er moet een eigenaar zijn van de bron, of in ieder geval van de assessment.

Interviewer (Emma): (30:34) Ja, eens. Bedankt daarvoor. Dat was denk ik wat ik wilde laten zien. Ik heb ook bijvoorbeeld meer tekst waarin ik dingen meer uitleg, en importance iets meer uitgewerkt met richtlijnen, ook voor mensen met minder inhoudelijke kennis.

Geïnterviewde (B): (31:05) Ja, super. Ik denk heel nuttig.
En echt goed te gebruiken binnen de organisatie als je dus in "define the use" daarin al koppelt naar daar waar de databron bedoeld is te gebruiken. Dus dat je daarin kan zeggen: voor dit project of deze use case is deze assessment gedaan. Dat is denk ik wat je afvang wilt vangen. Want dezelfde bron kan in een andere use case, vanuit een andere context, een andere kwaliteit hebben.

Interviewer (Emma): (31:41) Ja, precies. Ik heb inderdaad minder gefocust op welke persoon de assessment heeft gedaan, maar ik probeer het wel te zien in waarvoor de data wordt gebruikt.
Ik denk dat ik veel wijzer ben geworden.

Geïnterviewde (B): (32:05) Dat is fijn, blij dat ik kon helpen.

Interviewer (Emma): (32:07) Ja, zeker. Het is heel behulpzaam om deze gesprekken te voeren en om te horen wat mensen in de praktijk meemaken en hoe dat samenhangt. Dat voegt veel toe.

Geïnterviewde (B): (32:28) Top. Oké, klinkt goed. Ik ben benieuwd: als het een tool is die je zo kan gebruiken, dan zou ik hem meteen op alle bronnen loslaten. Laat maar weten wanneer hij te gebruiken is.

Interviewer (Emma): (32:38) Is goed. Op 22 januari moet ik klaar zijn, dus rond die tijd kan ik het delen.

Geïnterviewde (B): (32:44) Top, geweldig.

F.3 Transcript expert C

Interviewer (Emma): Okay, nice. I think it is recording, but I think it will be saved on your laptop.

Interviewee (C): (0:26) Yes. I can share it with you via SharePoint or email. We can try that.

Interviewer (Emma): (0:35) Yes, if that would be possible, that would be really nice. I am sorry, I did not have this problem before.

Interviewee (C): (0:44) That is fine. I hope it works.

Interviewer (Emma): (0:45) Okay. Thank you for meeting with me today. Did you have a chance to read what I sent you before the vacation?

Interviewee (C): (0:58) Yes, I had a look.

Interviewer (Emma): (1:00) Okay, thank you. That is really nice. We have met before, but I am currently working on developing a framework to assess the data quality of datasets used for supply chain mapping.
I want to see if the framework works in practice and whether the characteristics I identified in the literature also hold up. That is why I am talking to experts, to understand how this works in practice.
Could you start by giving a short introduction of who you are and what your role is at TNO?

Interviewee (C): (1:58) Yes. My name is [name]. I started at TNO last year in December, so it has been more than one year now. I started in the data science team and now that is the data innovation team.
Within the team, I generally work on NMO projects, especially supply chain analysis. For this analysis, I mainly use FactSet data, sometimes recycling data and SNP data. I also mostly work on visualizing this data, such as showing relationships between companies in different countries.
In addition, I work on enriching the data and generating more insights from it. Besides this, I also support the teamlead on the business, unlicensed part, such as describing use cases and collecting use cases for the roadmap. I think that is it.

Interviewer (Emma): (3:01) Yes, that is really interesting. I saw some of the things you are working on, and it looks really good.
I want to start with my broad research on what the datasets look like. I derived several characteristics from the literature. My first question is whether these characteristics reflect what you see in

your day-to-day work, or whether you think something is missing or unexpected.

Interviewee (C): (3:42) Yes, actually, I think it is really nice. I mean, especially for FactSet. But are you asking specifically about FactSet, or about datasets in general, such as SNP data, etc.?

Interviewer (Emma): (3:51) In general. In the document, on the first page, I describe mapping methods, scope and resolution, and data sources. That is how I defined how you characterize different types of datasets in general

Interviewee (C): (4:12) Okay. First of all, I am not that much of an expert on the datasets themselves, as I mainly focus on visualization. But I did look at the data.
I think I worked with three datasets for this use case, especially for supply chain analysis. For FactSet, it was very well documented. The tables were explained very well, although there were almost 200 tables. Company names were standardized, and there were relationship keywords describing relationships between companies. It also included information about the companies, such as sectors, industries, and company profiles.
For the SNP data, it was not as standardized. Some company names did not match for example the FactSet company names. SNP data is shipment data, so it focuses more on physical transportation, while FactSet focuses more on trade relationships so more the supply chains data part.
I also used a recycling dataset from the recycling team. That dataset was useful to understand what was going on in the FactSet data. Because I created a match between the recycling data and the FactSet data, and I was able to find company names from the recycling database in FactSet. This gave insights into which companies are involved with which materials or recycling activities.

Interviewer (Emma): (5:51) Yes, that makes sense. So I hear you talking about some of the criteria I did in part two.
I am trying to rephrase my question. I tried to differentiate between different types of datasets by assigning characteristics to them, such as AI-generated datasets or empirically collected datasets. Do you think this way of describing dataset types works in practice? For example if you look at FactSet does this help to determine the type.

Interviewee (C): (6:51) Yes, I think it is not AI generated data, but it is empirical data.

Interviewer (Emma): (6:55) Yes, but I mean more generally. Do you think this way of differentiating dataset types works, or do you think there are more aspects to consider?

Interviewee (C): (7:18) Do you mean whether it works for our use cases, depending on how the data is generated?

Interviewer (Emma): (7:27) Yes, so in my framework, I differentiate between dataset types using a few characteristics. And by "if it works" I meant: Do you think those characteristics work in practice?

Interviewee (C): (8:01) Yes, I think they work very well. Empirically collected data usually refers to data collected from public sources. This applies to all of the datasets: recycling data, SNP data, and FactSet data. All of them are collected via public sources.

FactSet is the one using the most AI techniques to classify data, such as sectors and industries. This is very useful when analyzing the data. However, this does not apply to SNP data. It was not standardized the data, including company names and addresses.

Interviewer (Emma): (8:48) Okay, thank you. We already touched on consistency in names etc.. But to reflect these data quality challenges, I identified nine diagnostic criteria in the literature. Do you have the document open?

Interviewee (C): (9:16) Yes, I have it open.

Interviewer (Emma): (9:26) Okay good, then I will refer to the document. Do you think these nine criteria reflect the problems you see in your work? Or are some not important, or is something missing?
For example, you already said consistency is a really important one.

Interviewee (C): (9:50) Yes, consistency is very important. Timelines are also really important because we need to know whether relationships are currently ongoing. If there is no end date for the relationships, you can assume the relationship still exists.
Completeness is also important to gain more insights into materials of supply chains, and shipment data. And to know what companies do. Accuracy is also very important for the relationships.
What do you mean by representativeness?

Interviewer (Emma): (10:38) These are the shortened descriptions. For example, if only data from Europe and the US are represented in the dataset, because those data are easier to obtain. Or of public companies because the data is publicly available, then certain regions or company types may be overrepresented.

Interviewee (C): (11:12) Yes, I think that is also very important. But it was challenging for me when analyzing the data. For SNP data, it was only port data. For FactSet, it was only public company data. We do not always know exactly what private companies do or which companies ship which materials or products.

Interviewer (Emma): (11:52) Do you think it is clear from the dataset documentation where these biases or problems of representativeness are, or is that often unclear?

Interviewee (C): (12:10) Yes sometimes it is vague, when you need to figure out yourself when there is no data about a company for example. The documentation states where the data comes from, but it does not clearly state what is missing or underrepresented. It is a bit vague.

Interviewer (Emma): (12:52) Yes, it usually shows what is included, not what is missing or underrepresented. That could be a big issue?

Interviewee (C): (13:03) Yes, I agree.

Interviewer (Emma): (13:09) Is there any criterion that you think is less important in your work,

or one you would compromise on first?

Interviewee (C): (13:20) What do you mean?

Interviewer (Emma): (13:29) For example, you can say I would compromise on consistency because we can fix that easily.

Interviewee (C): (13:47) Let me take a look. I think it is good that you defined these nine criteria. It is really well defined. Accuracy and validation are very important, especially to create dependency chains and for visualization. Documentation is also important.
Representativeness would be the one I would compromise on more easily, compared to accuracy or timeliness that are important.

Interviewer (Emma): (14:58) So representativeness is less critical because most of the time you mainly want to see connections between companies? Or how would you motivate that representativeness is less important?

Interviewee (C): (15:28) Yes. I mean when we come to the accuracy part: if the data is correct in my opinion it does not matter as much where the data is coming from. It is important if it is up to date, the timeliness.

Interviewer (Emma): (16:03) Thank you. Those are my questions about the criteria. Now I want to talk about the framework itself I developed from these criteria etc. Did you understand how it works?
So you first define the use, then you specify which data set you're using. And then you'll define for each criteria. OK, these are the past conditions. The criteria has to have this at least to pass. And then you'll also assign an importance on how important the criterion is for your framework and then. And you'll assess it and conclude and I have a form where you can visualize it a bit more on how you would write it down. I can also show you if you want.
But do you think the framework will work in practice? Do you think it's logical? You could actually use it?

Interviewee (C): (17:33) Yes, I think it would be really helpful. We are still looking for other datasets, such as [name database], and this framework could help assess whether those datasets work for us.

Interviewer (Emma): (17:51) Do you think any steps are particularly difficult or unclear? A part you would get stuck on or is vague?

Interviewee (C): (18:19) The last step, interpreting and concluding, would be the most difficult. Datasets are very large, and it takes time to analyze them and draw conclusions.

Interviewer (Emma): (18:39) So the assessment itself, like defining the conditions a dataset needs to pass is manageable, but making a final interpretation and decision is difficult? Because maybe you have a decision but what would you do then? Do you mean that?

Interviewee (C): (19:33) Yes, especially when decisions such as renewing licenses are involved. Maybe that would be difficult with this framework to make a decision on the datasets.

Interviewer (Emma): (19:55) Do you think that is because it is hard to determine if a dataset is good enough or do you think it is because there are maybe no alternatives?

Interviewee (C): (20:21) This framework would be really nice and helpful to make a decision. But we also need to consider other use cases beyond one specific analysis. That is why I said the last step would be difficult.

Interviewer (Emma): (20:53) What other factors would you want to include?

Interviewee (C): (21:01) We often use datasets for multiple use cases, such as visualization and analysis of the supply chain. The framework helps for one use case, but we also need to consider whether the dataset is useful for others. That's why I think the last step would be difficult.

Interviewer (Emma): (22:53) So more to have a broad view. So maybe for this use case the dataset is good, but if you want to purchase a license then you also want to know if it is good for the other use cases (look at the whole view).

Interviewee (C): (21:50) Yes. For example, I think it is really good to visualize relationships between companies, especially for FactSet data. But when it comes to the material or product side, it was not that useful, because it did not include any information on that. It was only based on keywords.

Interviewer (Emma): (22:17) Yes. Okay, I think I really understand what you are saying. It helps for a specific use case and a specific dataset, but for the real questions in your work, whether to purchase a dataset or not. You need to take a broader view of the purposes.

Interviewee (C): (22:47) Yes.

Interviewer (Emma): (22:53) That makes sense. But if we look specifically at one use case and one dataset, which of all the steps in the assessment do you think is the hardest? Or do you think something is missing?

Interviewee (C): (23:18) Do you mean the hardest to implement?

Interviewer (Emma): (23:22) Yes, when you are actually doing the assessment for a dataset and a use case, where do you think you would get stuck first?

Interviewee (C): (23:35) I think it is defining the pass conditions. It can be difficult, and it also requires making decisions. So defining the conditions to pass would probably be the hardest part.

Interviewer (Emma): (23:39) Yes, that makes sense. In my thesis, I try to explain more about

how to define pass conditions, but it is still difficult to give clear guidelines because it really depends on the use case. It definitely requires judgment from the assessor using the framework.

Interviewee (C): (24:14) Yes. You have the criteria and characteristics, and you can define them according to the dataset. You can also assess the criteria, but defining the pass conditions is the challenging part.

Interviewer (Emma): (24:38) Do you think something is missing that could help define the pass conditions? Or is it just difficult in general?

Interviewee (C): (25:01) I think a thresholds could help. That would be useful.

Interviewer (Emma): (25:17) In the text, I describe some scenarios. For example, if you have certain types of datasets, you often see certain types of pass conditions. I also describe how the importance of criteria changes depending on the situation. But it was hard to define very concrete guidelines, because there is no single case.

Interviewee (C): (25:48) Yes. For example, it is written here as required, very important, important, somewhat important, but it can be difficult to define that. Something may be very important for me, but not for another use case. I think if it is well explained it can be easier to define this.

Interviewer (Emma): (26:13) Yes, definitely. The framework is really developed to assess one use case at a time.

Interviewee (C): (26:25) Yes.

Interviewer (Emma): (26:30) I did not describe how to define pass conditions when you want to assess multiple use cases at the same time.

Interviewee (C): (26:40) Yes, it might be important for me but not important for anyone else.

Interviewer (Emma): (26:44) Yes, I understand. This is really helpful to see how different people would actually use the framework. I have already received slightly different answers from others. Let me see. We still have a few minutes, so I can quickly show you the forms I made?

Interviewee (C): (27:37) Do you need me to fill it in?

Interviewer (Emma): (27:39) No, I will just show it. I am on my other laptop, so I need a moment to open it.

Interviewee (C): (27:44) Okay, that is fine. By the way, did you have a chance to look at the FactSet data and the Databricks?

Interviewer (Emma): (27:54) In the end, I did not do the FactSet case. Instead, I used AIPNET. I do not know if you have heard of it.

Interviewee (C): (27:54) No.

Interviewer (Emma): (28:10) It is an AI-generated dataset developed by researchers in Maastricht. I think it is also very interesting, but of course very different.

Interviewee (C): (28:17) Nice.
I am sorry that I could only help with FactSet and not with another supply chain case.

Interviewer (Emma): (28:38) Oh no, that is completely fine.
Do you see my screen?

Interviewee (C): (28:53) Yes, I can see it.

Interviewer (Emma): (28:55) I showed this to others as well, and they found it made the framework clearer because it is visualized. The first two steps focus on the dataset, where you define the criteria and the goal. Then, for each criterion, you define pass conditions and importance. After that, you fill in the results and the decision for each criterion, and based on that, you conclude.

Interviewee (C): (29:41) Which decision exactly? Is it whether it passed or did not pass?

Interviewer (Emma): (29:45) Yes. It shows whether the result is good enough to pass the pass conditions. Then you compare those decisions with the importance to conclude whether the dataset is good enough for your use case. This is very small here, but that is the idea.

Interviewee (C): (30:20) Can we also use calculations, like multiplying by an importance rate or something?

Interviewer (Emma): (30:26) I chose not to incorporate numbers and instead use wording. But you can definitely assign numbers to importance. For example, not required could be zero, and very important could be five.

Interviewee (C): (30:41) Right.

Interviewer (Emma): (30:46) I chose not to do that because everything is very context-dependent. It is hard to assign numbers. That is why I used wording instead, but you could definitely quantify it and assign numbers to both importance and decisions.

Interviewee (C): (31:12) Yes.

Interviewer (Emma): (31:23) And also give the decision a number. For example, you could say four out of five or five out of five. You can definitely do that. But I chose a wording-based interpretation.

Interviewee (C): (31:35) Yes.

Interviewer (Emma): (31:36) If you really want to quantify it, it is definitely possible.

Interviewee (C): (31:41) I think it is really nice.

Interviewer (Emma): (31:43) Okay. Those were my questions. Is there anything else you would like to add?

Interviewee (C): (31:58) Thank you for your questions. If you need anything else, you can always email me.

Interviewer (Emma): (32:10) Thank you so much. It was really nice speaking with you and very helpful.

Interviewee (C): (32:13) Thank you too. I hope it was useful. Sorry, go ahead.

Interviewer (Emma): (32:34) Especially the point about decision-making, such as whether to buy a license. You are not only looking at whether you already have the dataset, but also at broader considerations.

Interviewee (C): (32:55) Yes.

Interviewer (Emma): (32:56) Okay. Thank you so much.

F.4 Transcript expert D

Interviewer (Emma): (0:14) Om te beginnen zal ik mezelf even voorstellen. Ik ben Emma en ik loop nu stage onder Elmer. Hij heeft mij naar jou doorverwezen. Ik kijk naar de datakwaliteit van datasets die worden gebruikt voor supply chain mapping.
Daarvoor heb ik een framework ontwikkeld, en met deze interviews wil ik kijken hoe mensen die daadwerkelijk met deze datasets werken hiertegen aankijken.
Ik wil graag begrijpen hoe dit in de praktijk werkt en of de theorie die ik uit de literatuur heb gehaald ook overeenkomt met hoe jullie dit in de praktijk ervaren.
Zou je jezelf eerst even willen voorstellen? Wat doe je bij TNO?

Geïnterviewde (D): (1:06) Ja. Ik ben [naam]. Ik heb economie gestudeerd in Tilburg. Daarna heb ik een tijd bij Economische Zaken gewerkt, maar daar was ik vooral bezig met afstemmen en had ik te weinig ruimte om inhoudelijk na te denken. Daarom ben ik overgestapt naar TNO, en dat bevalt eigenlijk heel goed.

Interviewer (Emma): (1:25) Hoe lang werk je nu bij TNO?

Geïnterviewde (D): (1:27) Sinds de zomer van 2023, dus nu ongeveer twee tot tweeënhalf jaar. Ik ben hier begonnen met meer kwalitatief werk, zoals interviews en literatuuronderzoek, maar ben steeds kwantitatiever gaan werken. Dat vind ik namelijk interessanter.
Wat ik interessant vind aan data is dat je minder uiteenlopende antwoorden krijgt. Bij interviews, als je acht verschillende experts dezelfde vraag stelt, krijg je vaak acht verschillende antwoorden. Bij data heb je dat minder, en dat maakt het voor mij wat duidelijker en concreter.
Ik houd me vooral bezig met innovatiebeleid en R&D-beleid. Voor het NMO heb ik samen met [namen] gekeken naar hoe economen aankijken tegen kritieke materialen en leveringsonderbrekingen. Daarvoor heb ik samen met [naam] een analyse gedaan naar hoe investeerders aankijken tegen leveringsonderbrekingen, op basis van aandelendata. Daarbij heb ik FactSet-data gebruikt via [naam], met wie je volgens mij ook al hebt gesproken.

Interviewer (Emma): (2:55) Ja, [naam] ken ik wel.

Geïnterviewde (D): (2:59) Ja, hij weet het meeste van de supply chain data.
Ik ben dus eigenlijk econoom en werk veel met innovatiebeleid, maar ook steeds meer met data-gedreven analyses die daarbij aansluiten.

Interviewer (Emma): (3:16) Ja. Leuk, dat is een interessante ontwikkeling. Ik heb ook het idee dat veel mensen daar een beetje in rollen, misschien hier bij TNO, doordat het een relatief nieuw vakgebied is.

Geïnterviewde (D): (3:34) Ja, binnen TNO kun je inderdaad veel verschillende kanten op. Dat is heel leuk. Als je iets interessant vindt, kun je daar gewoon bij aanhaken. Dan maakt het niet zoveel uit wat voor achtergrond je hebt.

Interviewer (Emma): (3:46) Ja, precies. Had je de kans gehad om het documentje dat ik had doorgestuurd door te kijken? Zo niet, is dat ook helemaal goed.

Geïnterviewde (D): (3:57) Ik had het nog niet gezien.

Interviewer (Emma): (4:01) Ik denk dat ik het nog voor de vakantie heb opgestuurd. Ik kan me voorstellen dat het ergens onder is blijven liggen.

Geïnterviewde (D): (4:06) Ja, ik zie het. Het staat inderdaad nog ongeopend.

Interviewer (Emma): (4:13) Als je het er even bij wilt pakken, is dat handig. Dan kan ik er makkelijker naar verwijzen.

Geïnterviewde (D): (4:14) Ja. Ik heb het nu voor me.

Interviewer (Emma): (4:25) Dat is helemaal goed hoor. Het is ook een drukke periode zo voor de vakantie.

Geïnterviewde (D): (4:28) Ja, inderdaad. Het eerste halfjaar is meestal vrij rustig, maar het

tweede halfjaar is richting het einde altijd erg druk.

Interviewer (Emma): (4:37) Ja, dat begrijp ik. Het eerste onderdeel gaat over kenmerken die ik uit de literatuur heb gehaald, waarmee je een dataset kunt typeren.
Mijn eerste vraag is of deze typering, dus de verschillende kenmerken zoals mapping method, scope, resolution en data sources, reflecteert hoe je in je werk datasets onderscheidt. Misschien niet heel bewust, maar wel impliciet. Neem gerust even de tijd om het te bekijken.

Geïnterviewde (D): (5:20) Ja. Dus dat gaat dan over hoe je verschillende datasets typeert, zeg maar.

Interviewer (Emma): (5:30) Ja, welke verschillende datasets je hebt. Dat heb ik geprobeerd te typeren met deze kenmerken.

Geïnterviewde (D): (5:36) Oké.

Interviewer (Emma): (5:47) Of er bijvoorbeeld nog iets ontbreekt, of iets dat je mist.

Geïnterviewde (D): (6:11) Ik denk niet echt. Je zou misschien tijdshorizon kunnen toevoegen, maar dat zou ook onder resolutie kunnen vallen.

Interviewer (Emma): (6:22) Ja, bijvoorbeeld qua scope, maar dan meer in de zin van welke tijd dekt de data. Is het recente data of meer historisch?

Geïnterviewde (D): (6:29) Ja. De meest waardevolle datasets zijn denk ik die heel ver teruggaan, maar die zijn ook het moeilijkst te vinden, zeker bij supply chains. Dat is pas in de afgelopen paar jaar echt relevant geworden.
Ik heb bijvoorbeeld gekeken naar China in 2010, toen zij de export van bepaalde kritieke materialen naar Japan stopzetten. Dat wilde ik analyseren met de FactSet Supplier Relations dataset, maar die data ging maar heel beperkt terug tot 2010.
Daarom hebben we uiteindelijk de hele dataset van 2010 tot 2024 gebruikt en de aanname gedaan dat als bedrijven nu met kritieke materialen werken, ze dat toen ook deden. Dat blijft wel een aandachtspunt.

Interviewer (Emma): (7:31) Ja, tijd is inderdaad ook iets dat een dataset typeert. Wordt de data bijgewerkt, bevat het recente informatie, of is het meer een historisch overzicht?

Geïnterviewde (D): (7:35) Ja. De meeste datasets zijn denk ik vooral recent. Het tijdsinterval, dus of data dagelijks of maandelijks beschikbaar is, kan ook onder resolutie vallen. Dat geldt ook voor hoe ver je terug kunt gaan en hoe real time de data is.
Bij het NMO zijn ze bijvoorbeeld bezig met een early warning system. Daarvoor wil je zo actueel mogelijke data. Bij supply chains is dat lastig, maar wel heel belangrijk. Ik weet niet of [naam] dit al heeft verteld, maar ik heb van [naam] gehoord dat DHL een dataset heeft opgebouwd met alle pakketstromen van land naar land.
Als er ergens een verstopping is, weten zij dat als eerste. Daardoor kunnen ze heel real time werken.

Wereldwijde gebeurtenissen die impact hebben op logistiek kunnen zij heel goed in kaart brengen. Dat geeft die dataset veel waarde en maakt die heel anders dan andere datasets.

Interviewer (Emma): (9:05) Dat is wel heel interessant. Daar had ik nog nooit van gehoord.
Als je kijkt naar deze kenmerken, kun je dan een situatie bedenken waarin je een dataset zou afwijzen op basis van één van deze kenmerken?

Geïnterviewde (D): (9:42) Afwijzen?

Interviewer (Emma): (9:45) Ja, bijvoorbeeld in een specifieke situatie.

Geïnterviewde (D): (9:50) Ik denk dat de grootste filtering zit in het niveau waarop de data beschikbaar is: bedrijfsniveau, sectorniveau of landniveau.
Als je echt op bedrijfsniveau wilt kijken, vallen heel veel datasets af. Tegelijkertijd is dat juist een heel interessant niveau. Het kost veel moeite om bedrijfsspecifieke inzichten te krijgen. Resolutie is daarin dus erg belangrijk.

Interviewer (Emma): (10:35) Ja, dat klinkt logisch.
Dan het volgende onderdeel. Dat gaat over negen diagnostische criteria die uit de literatuur komen en die proberen data-kwaliteitsproblemen samen te vatten. Die criteria kun je gebruiken om datasets te beoordelen.
Mijn vraag is of deze criteria de problemen reflecteren die jij in je werk tegenkomt, en of er iets ontbreekt of juist onverwacht is. Neem gerust even de tijd om ze te lezen.

Geïnterviewde (D): (11:22) Even kijken.
Dit zijn vooral kwaliteitskenmerken, denk ik.

Interviewer (Emma): (12:27) Ja, klopt. In het framework toets je de dataset uiteindelijk aan deze criteria. Ik ben benieuwd of je dingen herkent, of dat je bepaalde criteria juist heel belangrijk vindt of vaak tegenkomt.

Geïnterviewde (D): (12:39) Ja. Het lijkt mij een goed en uitgebreid overzicht. Als je dit naast de datasets legt, kan het een goede indicatie geven van hoe de kwaliteit is en hoe bruikbaar een dataset is. Ik denk wel dat een paar dingen lastig zijn om te bereiken of te meten.

Interviewer (Emma): (13:30) Is het misschien lastig om te meten?

Geïnterviewde (D): (13:30) Ja, met name completeness. Dat weet je natuurlijk nooit helemaal zeker. Hetzelfde geldt voor representativeness. Dat kun je op landniveau of sectorniveau bekijken, maar bij supply chains heb je geen uitputtend overzicht.
Dat maakt het lastig om completeness en representativeness echt goed te meten. Je zou wel samples kunnen nemen en waarschijnlijk een methode kunnen ontwikkelen om daar iets over te zeggen, maar op het eerste gezicht lijken dit de lastigste criteria om te meten.

Interviewer (Emma): (14:10) Als je even los van de meetbaarheid kijkt, welke criteria vind je

inhoudelijk het belangrijkst in je werk? Zijn er criteria waarvan je zegt dat je daar concessies op kunt doen, terwijl andere echt goed moeten zijn?

Geïnterviewde (D): (14:47) Ja, dat vind ik lastig.
Ik denk dat je in principe om veel dingen heen kunt werken als je toegang hebt tot andere datasets. Bij supply chains heb je vaak niet de luxe om heel kieskeurig te zijn op heel specifieke onderdelen. Als een dataset niet heel accuraat is... Ik vind het lastig.
Als ik de Supply Relations dataset van FactSet als voorbeeld neem, dan bestaat die uit keywords die relaties tussen partijen beschrijven. Die keywords geven aan dat er een relatie is, maar niet om hoeveel goederen het gaat of hoe belangrijk die relatie is. Dat staat er allemaal niet in.
Dus het is niet heel accuraat, niet compleet en niet heel consistent. En ook niet heel timeliness, de data is gebaseerd op documenten die bedrijven publiek beschikbaar maken. Het is ook niet echt representatief. Het bouwt voort op wat bedrijven zelf publiceren. Het is niet heel transparant, voor zover ik weet. Ik heb daar ooit naar gekeken, maar ik kan me voorstellen dat het lastig is om daar volledig transparant over te zijn.

Interviewer (Emma): (16:34) Wat ik heb gelezen, is dat ze niet heel specifiek beschrijven wat ze precies hebben gedaan.

Geïnterviewde (D): (16:40) Nee, klopt. Validatie en verificatie zijn ook lastig.
Als je de dataset op alle criteria beoordeelt, zou ik hem niet heel hoog scoren. Toch is hij best handig, omdat er weinig betere alternatieven zijn. In die zin valt er overal wel iets op af te dingen. Het liefst heb je alles perfect, maar het is niet erg als je niks hebt, zolang je maar goed beseft wat je wel en niet kunt zeggen met de data die je gebruikt.

Interviewer (Emma): (17:11) Dus je zegt niet dat één criterium minder belangrijk is, maar dat het vooral gaat om weten wat de beperkingen zijn?

Geïnterviewde (D): (17:27) Ja. Ik denk dat het helder communiceren over de kwaliteitseisen en over wat een dataset vooral niet bevat belangrijker is dan het strikt voldoen aan alle criteria.
Stel dat wij de Supply Relations dataset gebruiken en een andere partij heeft precies een dataset met alle HS codes tussen bedrijven over de afgelopen twintig jaar, dan heb je ineens niets meer aan die Supply Relations dataset. Maar omdat er niets beters is, blijft hij bruikbaar.

Interviewer (Emma): (18:05) Ja, je doet er dan maar mee wat er is.

Geïnterviewde (D): (18:08) Ja, eigenlijk wel. Dat maakt het ook een uitdaging en verklaart misschien waarom die mappingmethodes bij die data characteristics. Toen ik dat las, dacht ik eerst dat AI-gegenereerde datasets niet zo nuttig zijn. Maar als je websites scrapt en daar een large language model overheen laat lopen om patronen te herkennen, dan kom je eigenlijk op hetzelfde niveau uit als de Supply Relations dataset.

Interviewer (Emma): (18:43) Ja. Ik weet niet of [naam] dit ook aan jou heeft verteld, maar ik gebruik nu AIPNET als case. Dat is gebaseerd op HS codes en de bijbehorende omschrijvingen. Daarmee genereren ze product-tot-productrelaties en logische verbanden tussen producten.

Ze hebben wel gevalideerd op hoe accuraat het is, maar tegelijkertijd denk ik dat het vooral in combinatie met andere datasets nuttig kan zijn. Omdat er zoveel data ontbreekt, kun je op deze manier toch een vollediger beeld krijgen. Dat is misschien wel leuk om naar te kijken dan als het je interesseert.

Geïnterviewde (D): (19:31) Ja zeker. Hoe heet het?

Interviewer (Emma): (19:36) AIPNET, het AI-Generated Production Network. Het is ontwikkeld door onderzoekers uit Maastricht, onder andere door Mau en Karsten dacht ik.

Geïnterviewde (D): (19:46) Oh ja, ik zie het hier. Grappig, dat ze dat op die manier doen.

Interviewer (Emma): (20:05) Dat is best interessant. Maar hoe accuraat het is, ja dat is de vraag. Het is in ieder geval wel echt een nieuwe methode. Misschien werkt het juist goed in combinatie met andere datasets.

Geïnterviewde (D): (20:24) Ja, het ziet er wel interessant uit. Een interessante methodologie. En het is zo te zien gewoon te downloaden?

Interviewer (Emma): (20:35) Ja. Je moet kort een beschrijving geven en dan kun je het downloaden. Het is een Excel-bestand met eerst alle HS codes en bijbehorende omschrijvingen. Daarna zijn er berekeningen gedaan en heb je product-tot-productrelaties, over meerdere jaren.

Geïnterviewde (D): (20:59) Oké, dat is grappig.

Interviewer (Emma): (20:59) Het is ook handig om er een case mee te doen, omdat het een Excel-bestand is en daardoor voor mij ook een wat makkelijkere en toegankelijkere keuze.

Geïnterviewde (D): (21:09) Ja, dat is interessant.

Interviewer (Emma): (21:15) Oké, maar dan weer even terug.
Vanuit deze criteria heb ik geprobeerd een logisch framework te ontwikkelen waarmee je stap voor stap een dataset kunt beoordelen. Je ziet daar stap één tot en met vijf. Ik heb daar ook een formulier bij, maar dat heb ik hier niet toegevoegd.
Als je naar die stappen kijkt, denk je dan dat dit iets is wat je in de praktijk zou kunnen gebruiken? Of zie je punten waarvan je denkt dat die lastig zijn? Je zei net al misschien dat beoordelen van de criteria?

Geïnterviewde (D): (22:16) Dat is stap 2 dat beoordelen van die criteria?

Interviewer (Emma): (22:19) Stap twee is het definiëren van de karakteristieken, dus het type dataset. Stap drie is het formuleren van pass conditions en het toekennen van een importance level. En stap vier is het beoordelen of de dataset voldoet aan de pass conditions die je hebt opgesteld.

Geïnterviewde (D): (22:45) Ja. Ik zou er eigenlijk nog een stap aan toevoegen. Een stap waarin

je kijkt hoe je je methodologie kunt aanpassen om om te gaan met criteria waaraan de dataset niet voldoet, maar die je wel belangrijk vindt.
Waarschijnlijk is er geen dataset die alle negen criteria volledig afvinkt. Gegeven dat je altijd criteria hebt waar een dataset niet aan voldoet, en die misschien wel belangrijk zijn voor je studie, zou je moeten kijken of je je methodologie kunt aanpassen om daar toch mee om te gaan.
Kun je bijvoorbeeld een andere dataset toevoegen, of een aanname maken die ook in andere studies wordt gebruikt? Op die manier ervoor zorgen dat die criteria die niet worden gehaald misschien toch gedeeltelijk ondervangen. Dat kan je misschien niet allemaal van tevoren bedenken, maar meer een iteratief proces.

Interviewer (Emma): (24:11) Dus dat je eigenlijk al tijdens het beoordelen van de datakwaliteit kijkt of problemen op te lossen zijn binnen de methodologie? Dus dat je kijkt of problemen echt onoplosbaar zijn, en de dataset daarom afvalt, of dat het problemen zijn die je binnen je methodologie kunt oplossen. In dat geval kan de kwaliteit in zekere zin alsnog worden verbeterd.

Geïnterviewde (D): (24:57) Ja. Ik zou vooral kijken naar de belangrijkste redenen waarom je een dataset niet zou kunnen gebruiken. En dan kijken of je die op de een of andere manier kunt omzeilen, bijvoorbeeld door een andere dataset toe te voegen.
Aangezien bijna alle supply chain data best wel matig is.

Interviewer (Emma): (25:22) Ja, precies.

Geïnterviewde (D): (25:23) Neem zo'n AI-gegenereerde mapping van productrelaties. Die zal nooit honderd procent accuraat zijn en er zullen consistentieproblemen in zitten. Maar kun je dan nog een extra validatiestap toevoegen, bijvoorbeeld met een expertgroep of door een andere dataset te gebruiken?
Je zou bijvoorbeeld de FactSet-dataset kunnen gebruiken om te kijken of bedrijven binnen bepaalde HS codes daadwerkelijk handelen met bedrijven in andere HS codes die volgens het model gelinkt zijn. Zie je dat ook terug in andere data? Zoiets.

Interviewer (Emma): (26:17) Ja, precies. Dus echt al kijken naar mogelijke oplossingen, ook in vergelijking met andere datasets.

Geïnterviewde (D): (26:25) Ja, dat zou ik inderdaad toevoegen. Ik denk dat het een belangrijke stap is, omdat het bij elke dataset wel een beetje zoeken en bijschaven is om het werkbaar te krijgen.

Interviewer (Emma): (26:40) Ja, dat had ik ook al met iemand anders besproken. Bijna meer kijken naar wat de sterke en zwakke punten zijn en meerdere assessments naast elkaar leggen. Dan kun je zien hoe datasets elkaar kunnen aanvullen, of hoe je er op een andere manier mee omgaat. Dat vind ik zeker een goede toevoeging.
Als je kijkt naar de stappen die er nu staan, is er dan een stap waarvan je denkt dat die lastig is om uit te voeren? Bijvoorbeeld het formuleren van pass conditions, of een andere stap waar je tegenaan zou lopen bij toepassing van het framework?

Geïnterviewde (D): (27:44) Ja, ik denk dat het wel werkbaar is. Als ik die lijst met criteria zie, kan ik vrij makkelijk een indicatie geven van wanneer iets goed genoeg is en hoe belangrijk een criterium is.
Ik denk dat het prima en relatief snel toepasbaar zou moeten zijn.

Interviewer (Emma): (28:16) Ja, oké. Dus jij zou zeggen dat dit framework in het algemeen praktisch bruikbaar is?

Geïnterviewde (D): (28:24) Ja, dat denk ik wel.

Interviewer (Emma): (28:25) Met misschien nog wat toevoegingen.

Geïnterviewde (D): (28:27) Ja. Ik denk dat het prettig is om een soort meetlat te hebben waar je datasets langs kunt leggen, en die je consistent kunt gebruiken over meerdere datasets.
We krijgen vaak demo's van datasets van heel veel verschillende dataproviders. Iedereen heeft dan een goed salesverhaal. Het helpt om een aantal criteria te hebben, met het doel van je studie in het achterhoofd, waarmee je overzichtelijk kunt afvinken of een dataset voldoet.
Stel dat ik een demo krijg met een salespersoon en [naam] krijgt een demo van een andere provider, dan kun het goed naast elkaar leggen. Dat lijkt me heel handig.

Interviewer (Emma): (29:25) Oké. Ik denk dat dat mijn vragen zijn. Is er vanuit jou nog iets dat opkomt als je alles zo leest, of opmerkingen die je nog wilt meegeven?

Geïnterviewde (D): (29:44) Ik vraag me af wat precies je onderzoeksvraag is binnen je thesis.

Interviewer (Emma): (29:54) Mijn hoofdonderzoeksvraag is: "How can the data quality of datasets for supply chain mapping be systematically assessed?"
Daarnaast heb ik deelvragen: "What types of datasets are used for supply chain mapping, and what are their main characteristics?", "Which criteria can be applied to assess the data quality of supply chain mapping datasets?" en "For each criterion, how can the pass conditions and importance for the evaluation of the different datasets be defined?"

Geïnterviewde (D): (30:37) Oké. En wat is jouw beeld van supply chain data?

Interviewer (Emma): (30:47) Wat ik veel hoor en ook veel heb gelezen, is dat we eigenlijk moeten werken met slechte en beperkte data. Misschien niet per se slechte data, maar data met veel tekortkomingen.
Er zijn veel fouten en beperkingen. Bijvoorbeeld verandert het HS codesysteem voortdurend, worden codes samengevoegd of weer opgesplitst. Er valt dus zeker nog veel te verbeteren.
Ik heb zelf nog niet heel veel met deze datasets gewerkt, dus ik baseer me vooral op literatuur en gesprekken met anderen.

Geïnterviewde (D): (32:04) Nee, maar je hebt wel veel gelezen over die datasets en met veel mensen gesproken die ermee werken. Wie heb je allemaal geïnterviewd?

Interviewer (Emma): (32:16) [namen] heb gesproken en ga ik nog spreken.

Geïnterviewde (D): (32:26) Oké, nice.

F.5 Transcript expert E

Interviewer (Emma): Als eerste, ik ben Emma, maar dat weet je al. Ik heb gekeken naar datakwaliteit voor supply chain mapping en daar een framework voor ontwikkeld. Dit interview is bedoeld om te kijken of de onderliggende literatuur, zoals de characteristics en criteria, een overeenkomt met de praktijk. En om te beoordelen of het framework volgens jullie toepasbaar is in de praktijk.
Zou je jezelf kort willen voorstellen voor het interview, wat je functie is en wat je doet?

Geïnterviewde (E): (01:01) Mijn naam is [naam]. Ik werk nu ongeveer anderhalf jaar bij TNO. Ik ben vooral betrokken bij het Nederlands Materialen Observatorium. Mijn functie is een combinatie van junior consultant en data analyst. Ik houd me veel bezig met supply chain mapping en met het analyseren van trends in handels- en productiedata, vooral met focus op kritieke grondstoffen. Daarnaast werk ik ook aan projecten die meer vanuit een specifieke technologie worden benaderd, zoals technologieafhankelijkheden.

Interviewer (Emma): (01:48) Dank je wel. Als eerste wilde ik het hebben over het volgende. Vanuit de literatuur heb ik datasets op verschillende manieren gekarakteriseerd. Ik vroeg me af of die indeling, zoals mapping methods, scope, resolution en data sources, een beetje overeenkomt met hoe jij dat (misschien onbewust) in de praktijk doet.

Geïnterviewde (E): (02:16) Kun je het op het scherm laten zien?

Interviewer (Emma): (02:19) Ja, zeker.
Dit is het document. Bij Mapping Method onderscheid ik Empirically Collected Datasets, AI Enhanced Datasets en AI Generated Datasets. Dit is dus eigenlijk een manier om verschillende typen datasets te onderscheiden. Daarbij gebruik ik vier characteristics: Mapping Method, Scope, Resolution en Data Sources. Ik weet niet of je het al gelezen had?

Geïnterviewde (E): (03:03) Nee, ik weet ze niet helemaal uit mijn hoofd.

Interviewer (Emma): (03:05) Nee, dat snap ik. Neem je tijd.

Geïnterviewde (E): (03:12) En de vraag was?

Interviewer (Emma): (03:35) De vraag is of je vindt dat hier iets ontbreekt of juist overbodig is, of dat deze typering eigenlijk overeenkomt met hoe het in de praktijk werkt. Of mis je dus misschien nog een characteristic?

Geïnterviewde (E): (04:05) Ik denk dat dit op zich wel de belangrijkste zijn. Wat wel opvalt, is dat je mapping methods en data sources apart zet. Maar als je AI-generated mapping gebruikt, is dat ook een soort data source, toch?

Interviewer (Emma): (04:30) Ja. Dan zou het de underlying data zijn waarop je dat model hebt losgelaten.

Geïnterviewde (E): (04:40) Ja.

Interviewer (Emma): (04:41) Dan heb ik bijvoorbeeld nog een vraag. Heb je weleens een dataset afgewezen op basis van één van deze characteristics? Bijvoorbeeld dat je sowieso niet met AI-generated datasets werkt?

Geïnterviewde (E): (05:10) Over het algemeen vind ik het vooral belangrijk dat, als een dataset via AI is gegenereerd, duidelijk is op welke onderliggende bron die is gebaseerd. Anders is het lastig in te schatten in hoeverre de uitkomsten daadwerkelijk kloppen. Maar ik heb soms zulke soort datasets wel gebruikt, bijvoorbeeld het AIPNET (AI generated production network). Maar ik gebruik het dan vaak meer als een soort voorselectie. Om het daarna via andere bronnen of aanvullende research te controleren om verder uit te zoeken hoe het eruitziet.

Interviewer (Emma): (06:22) Je gebruikt misschien nooit maar één dataset. Maar je zou het meer als een soort beginpunt gebruiken dan de dataset waar je echt alles op baseert.

Geïnterviewde (E): (06:31) Ja, ik zou het niet als absolute waarheid gebruiken inderdaad. Maar het kan ons wel al veel schelen als je via een AI-methode al een voorselectie kunt maken. Bijvoorbeeld bij het koppelen van producten die worden gebruikt om andere producten te produceren. Daar zitten soms links tussen die niet heel waarschijnlijk zijn, maar die kun je er altijd nog later uit filteren. Het is in ieder geval beter dan elke keer helemaal opnieuw beginnen met een dataset van bijvoorbeeld 6000 producten.

Interviewer (Emma): (07:17) Logisch. Dan ...

Geïnterviewde (E): (07:19) Mag ik nog één ding aanvullen over de scope. Heb je daarin ook iets van de tijdsdimensie meegenomen?

Interviewer (Emma): (07:33) Niet expliciet. Ik heb het hier eerder met [naam] over gehad dat dat misschien nog ontbreekt, zoals hoe vaak de data wordt geüpdatet, hoe ver het teruggaat of hoe recent de data is. Je zou het misschien deels in de scope kunnen interpreteren, maar het is niet expliciet meegenomen. Want daarmee zou je eigenlijk wel een dataset typeren?

Geïnterviewde (E): (08:07) Dat neem ik zelf wel mee. Om te kijken of bepaalde supply chain relaties nog steeds actief zijn. Of als je naar een historische periode kijkt, bijvoorbeeld om de impact van een historische exportrestrictie te analyseren, dan wil je zoveel mogelijk reconstrueren hoe de supply chain er toen uitzag. Ook omdat supply chains continu veranderen, is dat relevant. Dat zou je wat opzich gewoon binnen de scope kunnen meenemen, net zoals geografische coverage dat je

ook een temporal coverage heeft.

Interviewer (Emma): (09:04) Ja precies. Want bedoel je dan vooral welke jaren het dekt, of ook hoe vaak het wordt geüpdatet en hoe recent de laatste update is, zoals tijdsinterval en tijdshistorie?

Geïnterviewde (E): (09:26) Beide zijn inderdaad relevant, maar dat zijn wel twee verschillende dingen die je inderdaad moet onderscheiden. Wat ik ook heb gemerkt, is dat je soms kunt achterhalen dat datasets een bepaalde bias hebben, bijvoorbeeld omdat ze gebaseerd zijn op heel recente bronnen.
Bijvoorbeeld jaarverslagen. Die worden niet continu geüpdatet, waardoor het beeld dat je hebt van de supply chain vaak achterloopt op de meest recente situatie.

Interviewer (Emma): (10:12) Ja, dat snap ik. Ik denk dat dat een waardevolle toevoeging is om te overwegen.
Ik heb geprobeerd om vanuit de literatuur alle datakwaliteitsproblemen te achterhalen en daar negen criteria uit te halen om datakwaliteit te toetsen. Mijn vraag is of je vindt dat deze negen criteria een volledig beeld geven van de problemen die jij in je werk ziet, of van de aspecten die je belangrijk vindt in een dataset. Of dat er misschien dingen ontbreken, of juist onverwachts zijn.

Geïnterviewde (E): (11:05) Dus het gaat erom deze criteria te gebruiken om te beoordelen of je een dataset wel of niet zou gebruiken?

Interviewer (Emma): (11:15) Ja, met deze criteria in het framework toets je de dataset. Maar hoe precies we dat doen, daar kom ik zo op terug.

Geïnterviewde (E): (11:25) Ik zal er even naar kijken. Oké, ik denk dat dit inderdaad wel de belangrijkste criteria zijn. Kun je misschien nog iets meer toelichten wat hier het verschil is tussen completeness en representativeness?

Interviewer (Emma): (12:24) Completeness betekent bijvoorbeeld dat alle landen zijn opgenomen, of dat elk continent vertegenwoordigd is. Representativeness gaat meer over de verdeling binnen die data. Bijvoorbeeld dat er veel meer data beschikbaar is uit Europa, of dat de data vooral afkomstig is van publieke bedrijven in plaats van private bedrijven. Completeness is dus de vraag of alles erin zit, ja of nee, terwijl representativeness gaat over de verhouding. Dat hangt af van de vraag die je stelt en of dat relevant is. Dit zijn ook de korte omschrijvingen in het document.

Geïnterviewde (E): (13:19) Wil je dat ik kijk in hoeverre het aansluit bij de datasets die ik gebruik?

Interviewer (Emma): (13:25) Ja, je kunt aangeven welke criteria je het belangrijkst vindt, of ze één voor één langsgaan. Soms is een criterium ook helemaal niet relevant voor een specifiek vraagstuk.

Geïnterviewde (E): (13:46) Een aantal dingen hiervan herken ik in ieder geval wel. Zou het misschien handig zijn om het te vergelijken aan de hand van de database FactSet die ik gebruik?

Interviewer (Emma): (14:07) Ja, dat is goed.

Geïnterviewde (E): (14:08) Een aantal dingen herken ik zeker. Waar ik bijvoorbeeld bij de FactSet-database ben aangelopen, is de consistency. De supply chain relaties worden bijvoorbeeld omschreven met een soort key words, maar dat is niet op een gestandaardiseerde manier gedaan. Dat raakt aan de vraag of er consistente definities worden gebruikt, en dat is daar niet echt het geval. Dat maakt het veel moeilijker om die database op een gestructureerde manier te analyseren. Voor elke use case moet je de database op een andere manier bevragen. Dat maakt het een lastiger proces dat steeds use-case specifiek is. Daarnaast heb ik gemerkt dat criteria zoals completeness en representativeness ook problematisch zijn. Veel databases hebben een bias richting westerse landen, Engelstalige publicaties en beursgenoteerde bedrijven. Kleine MKB-bedrijven vind je daar niet snel in terug.

Interviewer (Emma): (15:59) Dus je ziet ook echt veel meer Engelse publicaties. Hoe denk je dat dat komt?

Geïnterviewde (E): (16:09) Dat observeer ik niet direct, maar het heeft te maken met de onderliggende data. FactSet is volgens mij grotendeels gebaseerd op Engelstalige bronnen. Dat zie je terug in de database doordat de coverage beter is voor Amerikaanse bedrijven dan voor bijvoorbeeld Chinese bedrijven.
Het is lastig te zeggen in hoeverre dat alleen door taal komt, of ook door verschillen in transparantie in hoe landen en bedrijven rapporteren.

Interviewer (Emma): (16:54) Dat is interessant, daar heb ik nog niet echt zo over nagedacht.

Geïnterviewde (E): (17:00) Dat is in ieder geval wel iets wat ik hier al heb gemerkt. En ook als het gaat over type bedrijven: bijvoorbeeld in Nederland hebben we vooral MKB-bedrijven en weinig grote industriebedrijven. Die zijn moeilijk terug te vinden in dit soort data. Omdat ze ook minder hoeven te rapporteren en komen dus ook minder snel voor in die database. Dat is een probleem, want dit zijn juist de bedrijven waar we voor analyses in Nederland geïnteresseerd in zijn.

Interviewer (Emma): (17:56) Want stel dat je als eerste een concessie zou moeten doen. Je moet met een dataset werken, welk criterium zou je dan als eerste loslaten? Welke vind je het minst belangrijk?

Geïnterviewde (E): (18:38) Goede vraag. Dat hangt af van je insteek. Neem consistency als voorbeeld. Als een dataset niet volledig gestandaardiseerd is, maar wel een heel goed beeld geeft voor een specifieke use case of industrie, dan zou ik die alsnog gebruiken. Dat geldt vooral als je één specifieke use case hebt. Maar als je een databron zoekt voor veel verschillende analyses, wordt consistency juist heel belangrijk, omdat je niet voor elke use case een andere aanpak wilt gebruiken. Ik denk dus dat vooral ook veel te maken heeft met data fit. Je wilt voor de use case waar je naar kijkt een zo goed mogelijk beeld krijgen. Dat is het belangrijkste.

Interviewer (Emma): (20:08) Dat geeft een goed beeld. Op basis van deze criteria heb ik geprobeerd een framework te ontwikkelen om de dataset te toetsen binnen een specifieke use case. Eerst bepaal je waarvoor je de dataset wilt gebruiken. Daarna kijk je naar het type dataset en

dus de characteristics. Vervolgens stel je per criterium de pass conditions op, waaraan de dataset moet voldoen. Dan beoordeel je alle criteria en trek je een conclusie. Ik heb ook een visuele form ontwikkeld waarin dit duidelijker wordt, ik laat hem even zien. Je moet eerst de use case vastleggen en daarna de pass conditions en het importance level.
Je kan het zo even precies lezen. Maar denk je dat deze frameworkstappen haalbaar zijn in de praktijk? Of waar zie je de meeste moeilijkheden? Oh en je geeft dus per criterium ook een importance level aan. Zo kan consistency minder belangrijk zijn, terwijl goede documentatie juist een hogere importance krijgt.

Geïnterviewde (E): (22:06) Ja, dat lijkt mij inderdaad juist goed, dat je kunt variëren in hoe belangrijk je criteria vindt. Dat hangt sterk af van de use case. Op het eerste gezicht vind ik het logische criteria. In de praktijk is het wel lastig om sommige dingen te testen.
Sowieso is een algemene opmerking: Dit kun je over het algemeen best goed doen voor publieke databronnen, maar veel supply chain databases zijn betaald. Dan heb je maar beperkte mogelijkheden om van buitenaf te beoordelen.

Interviewer (Emma): (23:07) Ja dus eigenlijk voordat je hem überhaupt koopt.

Geïnterviewde (E): (23:09) Ja, om die dingen te kunnen controleren. Om een voorbeeld te geven: we hebben weleens gekeken naar databronnen die op basis van shipment records laten zien welke productcodes, zoals HS-producten, tussen welke bedrijven worden verhandeld. S&P Panjiva is daar bijvoorbeeld een aanbieder van. Dan kun je vaak een demo-sessie doen met die organisatie, en zij kunnen wat inzicht geven in de algemene coverage van regio's of handelstromen. Maar de vraag is dan hoe goed dat aansluit bij je specifieke use case, bijvoorbeeld als je naar de handel in een specifieke kritieke grondstof wilt kijken. Daar kun je dan niet zo goed op voorhand achterhalen hoe goed die coverage is.
Sowieso voor veel van die betaalde databronnen geldt inderdaad dus dat je ze eerst moet aanschaffen voordat je het goed kunt controleren hoe goed het is.

Interviewer (Emma): (24:38) Dus jij denkt dat het wel te doen is om die pass conditions op te stellen, maar dat vooral het beoordelen lastig is. Bijvoorbeeld of de accuracy klopt, omdat dat moeilijk te meten is, zeker als je nog geen toegang hebt tot de dataset.

Geïnterviewde (E): (24:57) Ja, ik zie dat als het belangrijkste obstakel: volledige toegang hebben tot de data om die criteria überhaupt te kunnen testen.
Misschien is het ook goed om te noemen dat een eerste stap kan zijn om, voordat je een bepaalde dataset aanschaft, te kijken wat je erover kunt vinden, bijvoorbeeld in de academische literatuur. Maar daar is best weinig over gepubliceerd, zoals je waarschijnlijk hebt gemerkt. Researchers moeten dan ook toegang krijgen tot dat betaalde platform, en er moeten afspraken worden gemaakt over wat wel en niet gepubliceerd mag worden. Daardoor is het ook via de academische literatuur lastig om erachter te komen hoe goed ze scoren.

Interviewer (Emma): (25:52) Oké, maar dat zou je dus wel als een mogelijke toevoeging zien: dat je in ieder geval kijkt of er iets over gezegd is. Bedoel je dat zo?

Geïnterviewde (E): (26:06) Ja, daar zou ik zeker naar kijken. Voor bijvoorbeeld FactSet en voor een aantal andere databronnen is er wel zoiets te vinden.

Interviewer (Emma): (26:18) Ja, voor de grote vaak wel. Maar dan bedoel je dus echt literatuur niet van de dataprovider zelf, maar academische literatuur waarin bijvoorbeeld een eigen assessment is gedaan of waarin er iets over is gezegd.

Geïnterviewde (E): (26:34) Ja, dat is een goed punt. Een onafhankelijke assessment door een partij die niet direct gerelateerd is aan de data-aanbieder zelf, daar zou ik wel vertrouwen in hebben. Dat zou ook helpen in de beoordeling.

Interviewer (Emma): (26:57) Dat doet misschien ook wel al wat werk van jou.

Geïnterviewde (E): (27:00) Ja, maar ik zou er alsnog niet blind op afgaan. Uiteindelijk hangt hoe goed een databron is ook af van jouw specifieke use case. Dat is niet per se hetzelfde als wat een externe partij heeft getest. Maar ik zou het wel meenemen.

Interviewer (Emma): (27:26) Zijn er nog andere dingen waarvan je zegt: als je dat toevoegt, wordt de toepassing van dit framework makkelijker? Of dit is echt nog iets wat erin mist?

Geïnterviewde (E): (27:50) Wat ik hier niet direct zie staan, maar dat relateert aan de pass conditions, is of je een benchmark hebt waarmee je vergelijkt? Dat is vaak makkelijker dan een harde pass condition formuleren. Dan vergelijk je gewoon met wat het alternatief is.

Interviewer (Emma): (28:20) Hoe bedoel je dat precies? Wat is die benchmark dan? Bedoel je hoe belangrijk het is?

Geïnterviewde (E): (28:30) Nee, ik bedoel ten opzichte van anderen, bijvoorbeeld andere data-aanbieders. Wat is het beste alternatief? Als je aanbieders vergelijkt op deze criteria, dan heb je op basis van die andere aanbieders misschien een soort benchmark, een vergelijkingsbasis van wat het beste is.

Interviewer (Emma): (29:06) Dus dat je meer kijkt naar wat gemiddeld is tussen andere datasets, en dat je daarop baseert, in plaats van dat je een heel onrealistische pass condition zet. Bedoel je dat?

Geïnterviewde (E): (29:17) Ja, op zich wel. Maar ik denk ook wel dat voor sommige criteria het ook best lastig is om vooraf een harde pass condition te formuleren. Dan is het misschien makkelijker om het als vergelijking te definiëren: we willen dat de coverage, bijvoorbeeld de geografische coverage, minimaal zo goed is als die van een andere aanbieder die wij nu gebruiken.
Stel dat we nu FactSet gebruiken en daaruit blijkt een bepaalde geografische coverage. Als je een andere databron hebt met net iets betere coverage, maar die is drie keer zo duur, dan kies je die misschien alsnog niet. Dan heb je in ieder geval een vergelijkingsbasis. Dat kan helpen bij het formuleren van de pass conditions.
Kun jij een voorbeeld geven van wat een pass condition zou kunnen zijn?

Interviewer (Emma): (30:35) Ik kan wel even mijn eerste uitwerkingen van mijn case laten zien. Ik heb uiteindelijk AIPNET gebruikt als case. In mijn scriptie staat het sowieso allemaal veel uitgebreider uitgelegd dan in dit documentje. Maar in deze tabel staan voorbeelden, in de tekst heb ik het wat meer uitgeschreven. Ik zou het bijvoorbeeld zo formuleren. Het hangt heel erg af van welke use case je hebt. Je kunt het kwantificeren door te zeggen: minimaal zoveel coverage, of iets anders. Maar je kunt ook zeggen: het moet een global scope hebben.

Geïnterviewde (E): (31:22) Ja, ik denk dus dat veel van deze dingen makkelijker te formuleren zijn als je ze kunt vergelijken met een bepaalde benchmark. Bijvoorbeeld bij completeness of representativeness. Voor landen is dat goed te formuleren: er zijn zoveel landen in de wereld en we willen minimaal een bepaald percentage. Maar als het gaat om bijvoorbeeld het aantal supply chain relaties dat in een database terugkomt, is er geen definitief getal waarmee je kunt vergelijken. Je weet niet hoeveel er in totaal zijn, en dat maakt het lastiger.

Interviewer (Emma): (32:27) Ja, je weet dan niet wat 100% completeness zou zijn. Dus dat je het eigenlijk vergelijkt ten opzichte van iets.

Geïnterviewde (E): (32:43) Ja, het percentage completeness of accuracy ten opzichte van wat. Dat wat kun je soms gewoon zo definiëren. Maar voor bepaalde criteria is het dan waarschijnlijk makkelijker om ze ten opzichte van een benchmark dataset te definiëren.

Interviewer (Emma): (33:04) Ja, ik snap wat je bedoelt.
Mijn vraag was vooral hoe je dit in de praktijk zou toepassen. Je zegt dus dat die pass conditions op zich wel te formuleren zijn, maar dat het voor sommige criteria lastiger is. Een benchmark zou daarbij kunnen helpen.

Geïnterviewde (E): (33:39) Ja.

Interviewer (Emma): (33:39) En denk je dat het lukt om die importance te benoemen? Bijvoorbeeld dat je voor een bepaalde use case zegt dat bij AI-generated datasets het belangrijk is dat je kunt zien hoe goed ze gevalideerd zijn en of de documentatie op orde is. Denk je dat dat goed op te schrijven is als jij zo'n assessment zou doen?

Geïnterviewde (E): (34:15) Ja, ik denk het wel. Het is soms best wel moeilijk om te quantificeren misschien.

Interviewer (Emma): (34:24) Daarom heb ik ook woorden gebruikt in plaats van cijfers. Het is lastig om zulke dingen echt in cijfers uit te drukken. Door woorden te gebruiken heb ik die nuance proberen te behouden.

Geïnterviewde (E): (34:45) Ja, dat lijkt me goed. Zoals je zegt, maak je bij een AI-generated dataset ook andere afwegingen dan bij een empirically collected dataset.

Interviewer (Emma): (35:11) Dat waren denk ik mijn vragen. Dus over het algemeen denk je dus dat dit relevant kan zijn, maar vooral stap drie en vier kunnen lastig zijn.

Dat waren mijn vragen. Heb je nog iets algemeens toe te voegen, of iets waarvan je denkt dat daar nog een knelpunt zit?

Geïnterviewde (E): (35:50) Zo één, twee, drie niet. Misschien bedenk ik later nog wat.

Interviewer (Emma): (35:54) Dat is dat helemaal goed. Ik wil altijd aan het einde nog even de ruimte geven om iets toe te voegen. Dankjewel voor je feedback. Dit waren mijn vragen.
Die assessment zal altijd wel lastig blijven.

Geïnterviewde (E): (36:21) Ja, bijvoorbeeld van dat [database], soms kost zo'n databron tien keer zoveel als wat we nu gebruiken. Dan vraag ik me af of het ook echt tien keer zo goed is.

Interviewer (Emma): (36:40) Dat vraag ik me ook af. Ik ben benieuwd wat ze in zo'n demo precies laten zien, want op de website staat vaak niet veel. Moet je dan maar aannemen wat ze vertellen.

Geïnterviewde (E): (36:58) Ja, precies. Wat ik in zo'n geval probeer te doen, is tijdens een demo zoveel mogelijk te kijken vanuit een concrete use case die we ook in de praktijk gebruiken. Vaak demonstreren ze het platform aan de hand van een specifiek product of een bepaald type bedrijf, maar dat is meestal niet representatief voor onze use case. Daarom probeer ik altijd te zorgen dat ze ook kijken naar de soort producten of bedrijven waar wij mee bezig zijn. Anders kiezen ze gewoon het beste.

Interviewer (Emma): (37:36) Ja, dan kiezen ze gewoon de beste. Dankjewel voor je tijd.

Chapter G. Final Data Quality Assessment Framework

Data Quality Assessment Framework

This document presents the Final Data Quality Assessment Framework, grounded in the fitness-for-use perspective (Wang and Strong, 1996). Data quality is assessed in relation to the research context.

Dataset characteristics

To distinguish between the different types of datasets used for supply chain mapping, we use several key dataset characteristics. These characteristics help clarify how a dataset is constructed and what types of analyses it is suitable for.

Mapping method: How the dataset is created. In practice, many datasets are a combination of the different methods.

- Empirically collected datasets

  Based on directly observed information such as company disclosures, surveys, or operational records.

- AI-enhanced datasets

  Built on empirical data but expanded or refined using algorithmic inference (e.g. association-rule mining, probabilistic estimation).

- AI-generated datasets

  Constructed mainly through model-based reasoning (e.g. large language models identifying product linkages).

Scope:
Scope refers to the geographical and sectoral coverage of the dataset. It concerns whether the data captures linkages within a single country, across regions, or on a global scale, and whether it focuses on one industry or spans multiple sectors

Resolution:
Resolution reflects the degree of detail within that scope, such as whether relationships are represented at the company, sector, or even national level.

Data sources:
The origin of the underlying data. This may include primary data (surveys, interviews), secondary data (company reports, customs data, commercial databases), and big data sources (web-scraped supplier lists, logistics records, patent or product registries). Many datasets combine multiple sources.

Diagnostic criteria

The diagnostic criteria used in the Data Quality Assessment Framework are:

1. Data fit

Data fit refers to the extent to which the dataset's structure, coverage, and granularity are aligned with the intended analytical or policy goal. This criterion operationalises the broader fitness-for-use principle by testing whether the dataset is appropriate for its specific purpose. In supply chain research, Culot et al. (2022) note that one must assess the fit between the data, the supply chain constructs, and the research questions to avoid drawing invalid conclusions.

Example: suppose policymakers want to analyze multi-tier supplier risks for critical products. A dataset that only contains immediate supplier links would be unfit for this purpose, because it lacks the depth (tiers) needed to map cascading risks.

2. Accuracy

Accuracy is the degree to which the recorded information truthfully represents real-world supply-chain relationships. It focuses on the correctness of existing data entries, whether supplier links, firm identifiers, and classifications accurately describe actual relationships. Culot et al. (2022) highlight "errors caused by the databases' classifications and assumptions" as a key issue, referring to data accuracy problems.

Example: a supply chain mapping dataset for a major automaker lists companies as suppliers that, in reality, never supplied that automaker. Such inaccuracies could obviously lead the automaker to the wrong conclusions.

3. Completeness

Completeness measures the extent to which the dataset contains all relevant entities, tiers, and attributes. Missing records, unreported supplier links, or omitted small firms can distort analyses and bias results (Van Schilt et al., 2024).

Example: consider a global electronics manufacturer trying to map its supply chain. An incomplete dataset might only list that manufacturer's top 5 suppliers (perhaps derived from a few public disclosures), omitting smaller or second-tier suppliers. As a result, any analysis of supply risk or diversity would be skewed.

4. Consistency

Consistency refers to the uniform application of definitions, classifications, and identifiers within the dataset and the alignment of those with other sources (Van Schilt et al., 2024). Inconsistent terminology, duplicate entries, or varying aggregation rules reduce comparability.

Example: if a company appears under slightly different names or codes in different parts of the dataset, it might be treated as two different suppliers incorrectly. In supply chain mapping, consistency issues often arise with firm identifiers (e.g. one record uses a legal name, another uses a common name for the same entity) or with classification of industries and locations. Such inconsistencies make it hard to merge or compare datasets and can lead to double counting or omissions.

5. Timeliness and temporal consistency

Timeliness ensures data is up-to-date and reflects the current state of supply chain relationships (Lee and Yoon, 2025). In fast-moving supply chains, outdated data (even by months) can mislead

decisions, particularly in real-time operational use.
When the focus is on analyzing historical data, temporal consistency is important. This means the data is recorded in a time-consistent manner: reporting periods are aligned, changes over time are tracked without unexplained gaps, and historical data is not retroactively altered without documentation. Culot et al. (2022) observe, for instance, that many commercial supply chain databases had methodological adjustments over the years in sources, languages, industry coverage, etc., which complicates longitudinal analysis.
Example of timeliness: in a low-quality dataset, a supplier that has already been replaced still appears in the data.
Example of temporal consistency: imagine tracking a company's supplier count over a 5-year period using a supply chain database. In a poor-quality scenario, the database changed its inclusion criteria in year 3 (say, it started including smaller suppliers or added new industries). If this change is not transparent, an analyst might see a spike in year 3's supplier count and conclude the company dramatically expanded its supply base.

6. Representativeness
Representativeness measures the extent to which the dataset captures the full and balanced population of firms, sectors, regions, and supply-chain tiers relevant to the analysis. From a scientific standpoint, representativeness is related to sampling bias: if the data sample is not representative of the population, conclusions may not generalize. Recent studies (such as Dong et al. (2024)) highlight issues like selection bias in supply chain data and emphasize aligning data choices with research questions to improve generalizability.
Example: in a mapping of the global textile supply chain, a non-representative (biased) dataset might only include suppliers that are in developed countries or only those suppliers that voluntarily report to a certain platform. Such a sample would not be representative of the real situation.

7. Transparency and documentation
Transparency refers to how openly the data source discloses its collection methods, definitions, and any cleaning or verification procedures. Documentation is the detailed information about how data was gathered, what each field means, and how often it is updated. Transparency and documentation are crucial for evaluating whether a dataset can be trusted, replicated, and correctly interpreted (Culot et al., 2022).
Example: a particular supply chain mapping platform provides a public dataset of supplier relationships but gives no information on how it collects these links (are they from surveys, from purchase orders, from web scraping news?). It also does not explain its company matching process or how often data is refreshed. This lack of transparency/documentation is risky: if an error or odd pattern appears, users cannot tell if it is a real supply chain insight or a data error.

8. Accessibility and comparability
This criterion addresses whether the data is easily obtainable and usable, and whether it can be readily compared or integrated with other data sources. Accessibility refers to the extent to which data can be accessed by those who need it, considering technical format, cost, and access rights. Data that cannot be accessed or practically used has limited value regardless of its intrinsic accuracy or completeness (Wang and Strong, 1996). Expensive subscriptions or delivery in highly specialized software environments may therefore limit both the dataset's impact and the level of external

scrutiny it receives. Comparability concerns the alignment of definitions, classifications, and formats across datasets, enabling meaningful comparison or integration. A lack of comparability complicates data integration and increases the risk of mismatches or biased results. Limited comparability is a recurring challenge in business and supply chain research and poses a barrier to replication and cumulative knowledge building (Liu, 2020).
Example of accessibility: a dataset of supplier relationships that is only available through a paid subscription costing thousands of dollars and delivered in a very specialized software format would be a quality concern, as ideally data should be as accessible as possible (without compromising necessary confidentiality).
Example of comparability: suppose one supply chain dataset identifies companies by their full legal names, while another dataset uses standardized business identifiers. If you try to combine them, you face a big reconciliation task, potentially matching names one by one and risking mismatches. This is an example of poor comparability.

9. Validation and verification
Validation and verification refer to the processes that confirm the dataset's accuracy and reliability, both through internal checks and external comparisons. Essentially, this criterion asks: has the data been cross-checked and quality-assured from multiple angles? Scientific arguments for this criterion come from the principle of triangulation: by comparing data with independent sources, one can significantly increase confidence in its correctness (Bans-Akutey and Tiimub, 2021).
Example: imagine a supply chain mapping dataset that lists Supplier X as providing components to Manufacturer Y. If this dataset is unverified, it is possible this link came from a single press article or a one-time survey response. In a high-quality scenario, the data provider or user would verify this relationship. For instance, they might check Manufacturer Y's annual report or Supplier X's customer list (if available), or see if trade data shows commerce between the two. If no independent confirmation is found, the link could be flagged as low confidence or removed.

Pass conditions and importance

To apply the framework in practice, each diagnostic criterion must be assessed against clear pass conditions and assigned a relative importance based on the dataset's intended use and characteristics. Pass conditions specify what qualifies as acceptable quality. After assessing each criterion we assign one of the following decisions:

- Passed: sufficient quality
- Not passed: insufficient quality
- Unknown quality: when information is missing

Importance reflects the weight each criterion carries in the overall assessment. Together, these elements operationalise the fitness-for-use principle by translating conceptual quality criteria into measurable evaluation standards.

Formulating and documenting pass conditions and importance
Pass conditions describe, in clear terms, what constitutes acceptable quality for each criterion within the specific analytical or policy context. They can be described either quantitatively or qualitatively.

Quantitative descriptions may include, for example, minimum coverage levels, update frequencies, or accuracy rates that indicate whether the dataset meets the expected standard.
Qualitative descriptions may refer to aspects such as the availability of documentation on data collection methods, the clarity of definitions and classifications, or the presence of validation procedures.
While quantitative specifications are generally preferred because they allow for more objective comparison and replication, qualitative descriptions remain appropriate when numerical information is unavailable, for simple yes/no conditions, or when evaluative judgement is required.
This flexible approach aims not to impose universal criteria, but to make explicit how "acceptable quality" is defined in the specific context.

The relative importance of criteria is written down using the following five-level verbal scale:

- Required
- Very important
- Important
- Somewhat important
- Not required

This format ensures clarity while avoiding false precision and enables consistent interpretation across users. The scale can be translated into numeric weights (5-1) if a numerical comparative analysis is desired. However, this study does not recommend such conversion, as it may oversimplify the qualitative nuances that the verbal scale is intended to preserve.
Each dataset characteristic influences the definition of pass conditions and the relative importance of criteria in its own way. Because the analytical contexts in which datasets are used vary greatly, it is not possible to identify a single criterion per characteristic that is universally most important. However, it is possible to provide guidance on how certain criteria become more or less critical depending on specific contexts, and to explain the underlying reasons for these differences. The following subsection outlines how the criteria importance changes in different contexts.

Criteria importance
Before presenting a schematic overview, this subsection first discusses the relative importance of each criterion. The narrative explanation below clarifies how the relevance of each criterion varies across different contexts. After this textual discussion, a summarising table is provided that captures the conceptual extremes of all nine criteria in a concise and systematic format.

1. Data fit
Data fit is most important when the dataset is used for targeted analytical or policy purposes, as the structure and level of detail must align with the research goal. It becomes less critical when data is explored in a more general or descriptive way, where precision in structure is not essential.

2. Accuracy
Accuracy carries the most weight for empirical datasets, which rely on factual information to represent supply-chain relationships. In AI-enhanced or AI-generated datasets, its importance depends on how strongly the data is grounded in observed reality, making it slightly less decisive for overall quality.

3. Completeness
Completeness is particularly important when datasets aim to provide a full overview of supply-chain networks or when quantitative measures depend on coverage. It is less crucial when the dataset serves exploratory or illustrative purposes, where partial information can still be informative.

4. Consistency
Consistency is key for ensuring that variables, classifications, and identifiers are applied uniformly within the dataset. It is especially relevant when information is collected from multiple sources or compiled over several stages. Minor inconsistencies are less problematic when the dataset is used for descriptive or one-time analysis.

5. Timeliness and temporal consistency
Timeliness is most important when analyzing volatile, rapidly changing phenomena, where it is critical that the dataset is up to date. Temporal consistency is critical for longitudinal analyses, as stable definitions and measurement methods are required to ensure comparability over time.

6. Representativeness
Representativeness is most relevant when findings are intended to reflect broader sectors, countries, or populations. For narrowly defined or case-specific analyses, it is of lower importance, as detailed accuracy within the selected boundary is the main concern.

7. Transparency and documentation
Transparency is crucial to understanding how data were obtained and processed. It is especially important in AI-enhanced or AI-generated datasets, where users must be able to interpret how relationships were inferred and validated.

8. Accessibility and comparability
Accessibility becomes more critical when the project constraints are very tight, e.g. when budgets, expertise, the available tools, time, etc. are limited. Comparability is most relevant when datasets are used across studies, sectors, or organisations. It is less critical for internal or single-purpose datasets, where external replication is not required.

9. Validation and verification
Validation gains importance when datasets include inferred or automatically collected information, as it confirms that the results are reliable. For empirical data, validation is often more straightforward but remains important to ensure credibility.

To provide a concise overview of how each criterion conceptually ranges between two extremes, Table 7 summarises the findings above.

Criterion	**Extreme 1: When the criterion becomes highly important**	**Extreme 2: When the criterion becomes less important**
Data fit	Targeted analytical or policy use	Explorative or descriptive use
Accuracy	Empirically collected, fact-based datasets	AI-inferred datasets
Completeness	Full network coverage required	Exploratory or illustrative purposes
Consistency	Multi-source or longitudinal analysis	One-time or descriptive analysis
Timeliness & Temporal consistency	Timeliness: volatile, dynamic data Temporal consistency: longer time period, more or varying data sources	Timeliness: stable data Temporal consistency: shorter time period, fewer or consistent data sources
Representativeness	Broad scope (generalisable)	Narrow scope (case-specific)
Transparency & Documentation	To interpret modelling assumptions (e.g. AI inferred datasets)	Internal or exploratory use
Accessibility & Comparability	Accessibility: limited budget, expertise, tools available Comparability: cross-study use or data integration	Accessibility: ample availability of budget, expertise, tools Comparability: internal or single-purpose use
Validation & Verification	AI-inferred datasets	Empirically collected datasets

Table 7: Conceptual extremes of the nine diagnostic criteria

In summary, the context determines how data quality should be interpreted and prioritised. The mapping method shifts the focus from factual accuracy to methodological transparency, while scope and resolution balance representativeness against precision. The data sources influence what level of validation and documentation is feasible. Together, these characteristics guide how the framework should be applied in practice and ensure that data quality is always assessed in relation to the dataset's intended use.

Step-by-Step Guide

To support its practical application, the framework includes a clear step-by-step guide.

Step 1: Define the use

Clarify the analytical or policy objective of the study.

As the framework is based on the fitness-for-use principle, the research goal helps determine which criteria are most relevant and under what conditions they should be applied.

Step 2: Specify the dataset characteristics

Identify the key characteristics of the dataset.

In addition to the research goal, the type of dataset also helps to determine which criteria are most relevant and under what conditions they should be applied. We ask the user to specify the mapping

method, scope, resolution and the data sources.

Step 3: Define pass conditions and assign importance
For each diagnostic criterion, determine what constitutes acceptable quality in the given context. Specify when a criterion is considered sufficient to "pass". And assign the weight each criterion carries in the overall assessment: Required, Very important, Important, Somewhat important, or Not required.
→ *Where pass conditions involve comparisons with other datasets (e.g. a dataset is required to perform at least as well as a reference dataset), references to these benchmark datasets are recorded in the Data Quality Assessment Form, including dataset names, version numbers, and data providers.*
→ *Where academic reviews of the evaluated dataset are available, references or links to these studies may be recorded in the corresponding field of the Data Quality Assessment Form.*

Step 4: Assess the criteria
Evaluate the dataset against each diagnostic criterion. Report the results and assign one of the following outcomes: passed, not passed, or unknown (when information is insufficient).

Step 5: Interpret and conclude
Combine the individual results and their relative importance to derive an overall evaluation. Summarise the findings and provide a conclusion or recommendation on the dataset's suitability for the defined use.

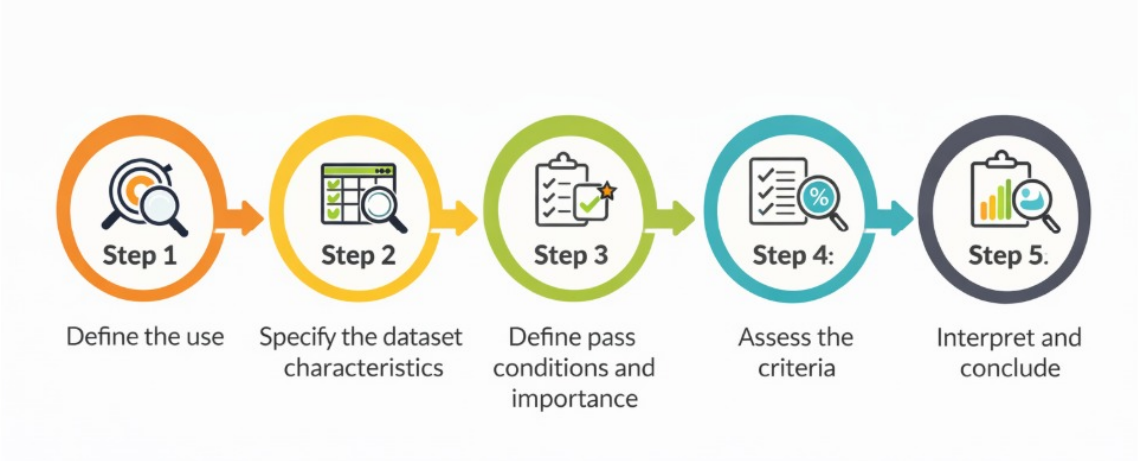
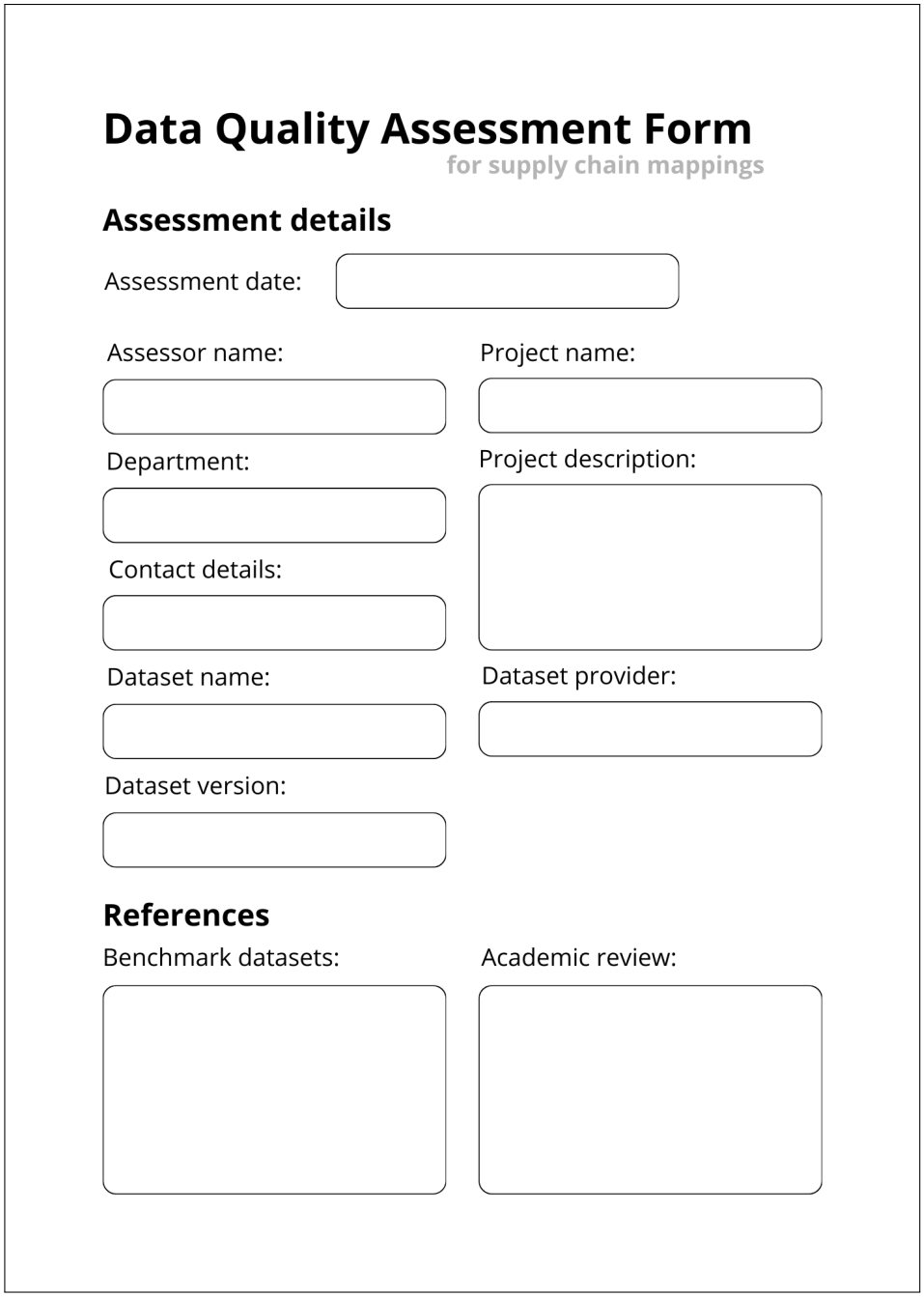


Figure 3: Overview of the Step-by-Step Guide of the Data Quality Assessment Framework

To facilitate a structured format to document these steps, an evaluation form is developed and an example (for illustration purposes only) is shown in Appendix H.

Chapter H. Final Data Quality Assessment Form



Data Quality Assessment Form

for supply chain mappings

Assessment details

Assessment date:

Assessor name:

Project name:

Department:

Project description:

Contact details:

Dataset name:

Dataset provider:

Dataset version:

References

Benchmark datasets:

Academic review:

About the dataset

Goal:

Mapping method:

Scope:

Data sources:

Resolution:

Diagnostic criteria

Criteria	Pass conditions	Importance
1) Data fit		
2) Accuracy		
3) Completeness		
4) Consistency		
5) Timeliness and temporal consistency		
6) Representativeness		
7) Transparency and documentation		
8) Accesibility and comparability		
9) Validation and verification		

Assessment

Criteria	Result	Decision
1) Data fit		
2) Accuracy		
3) Completeness		
4) Consistency		
5) Timeliness and temporal consistency		
6) Representativeness		
7) Transparency and documentation		
8) Accesibility and comparability		
9) Validation and verification		

Conclusion