

## Summary sheet for industrial datasets

Dr Hector Puyosa

Big Data & Innovation, Digital Transformation, February 2025

### I. Introduction

In the fields of machine learning (ML) and data science, datasets serve as the foundation for developing, testing, and validating algorithms. High-quality datasets are essential across industries, enabling advancements in areas such as image recognition, predictive maintenance, and natural language processing. Well-documented datasets streamline research, improve collaboration, and enhance model performance by ensuring transparency and consistency in data handling.

A dataset summary sheet is a crucial tool for documenting key dataset characteristics. It provides an organized and structured overview that helps data teams, researchers, and stakeholders quickly understand and utilize the dataset efficiently. This document outlines the importance of dataset summaries and provides guidance on the essential elements to include in a well-structured dataset summary sheet.

### II. Importance of a Dataset summary

A dataset summary sheet plays a vital role in data management by enhancing organization, transparency, and usability. Key benefits include:

- **Transparency and Collaboration:** Facilitates clear communication and accountability within teams, ensuring that everyone understands the dataset's contents and purpose.
- **Efficiency:** Well-documented data enables teams to work more effectively by reducing the time needed to comprehend data structure and attributes.
- **Data Understanding:** Acts as a guide for users, providing crucial details about the dataset's origin, structure, and transformations.
- **Data Quality:** Helps maintain consistency across models and ensures proper tracking of data freshness, dependencies, and potential errors.
- **Reproducibility:** Supports replicability of experiments by offering a well-defined reference for dataset characteristics and processing steps.

- **Dataset Selection:** Assists researchers and developers in choosing the most suitable dataset for their specific objectives.

### III. Key fields on Dataset summary sheet

A comprehensive dataset summary sheet should include the following elements:

**Dataset Name:** A unique identifier for the dataset.

**Description:** A concise summary of the dataset's content, purpose, and context.

**Data Sources:** Origin of the data (e.g., sensors, production logs, maintenance records).

**File Format:** Specification of data formats (e.g., CSV, JSON, Parquet).

**Data Structure:** Information on file organization and content layout.

**Data Dictionary:** Definitions of variables, abbreviations, missing data codes, and units.

**Metadata:** Tags, categories, or other contextual details to facilitate searchability.

**Time Range:** The period covered by the dataset.

**Size:** The dataset's volume and storage footprint.

**Update Frequency:** How often the dataset is refreshed or updated.

**Data Quality Metrics:** Indicators of data completeness, accuracy, and consistency.

**Preprocessing Requirements:** Any data cleaning or transformation steps required.

**Potential Use Cases:** Possible applications for the dataset.

**Related Datasets:** Links to complementary datasets.

**Responsible Team/Department:** Contact details for dataset owners or maintainers.

**Version History:** A record of changes and updates to the dataset over time.

**Compliance Information:** Adherence to regulatory or industry standards.

**Access and Reuse Conditions:** Guidelines on data usage restrictions or permissions.

**Methodology:** A brief overview of data collection and processing methods.

**Data Lineage:** Information on data transformations and pipeline history.

## IV. Conclusion

A well-structured dataset summary sheet is an invaluable asset for any organization working with data. It enhances transparency, promotes collaboration, and ensures that datasets are used effectively and responsibly. By including key details such as metadata, data dictionary, and quality indicators, organizations can maximize the utility and reliability of their datasets.

Standardizing dataset documentation practices enables better decision-making, improves reproducibility, and supports long-term data governance. Implementing a dataset summary sheet is a best practice that benefits data teams, researchers, and business stakeholders alike.

## References

The following sources provide further insights into dataset documentation best practices. Information was verified as of February 13, 2025:

<https://docs.data.world/en/99081-dataset-best-practices.html>

<https://data.wisc.edu/data-literacy/document/>

[https://datadryad.org/stash/best\\_practices](https://datadryad.org/stash/best_practices)

[https://knowledge.domo.com/Connect/Managing\\_DataSets/Best\\_Practices\\_for\\_Managing\\_DataSets](https://knowledge.domo.com/Connect/Managing_DataSets/Best_Practices_for_Managing_DataSets)