AN ANALYSIS OF INITIAL DATA STAGES IN DATA SCIENCE RESEARCH

Ika Siti Nafisa[1], Aji Setiawan[2], Nadia Nadawina[3]

[1,2,3]*Sekolah Tinggi Keguruan Ilmu Pendidikan Amal Bakti*
E-mail : Ikasitinafisa27@gmail.com[1], Setiawanaji7113@gmail.com[2], nadianadawina273@gmail.com[3]

ABSTRACT

The initial data stage is a crucial component in data science research as it determines the quality of the analytical process and the accuracy of the results obtained. Errors in managing initial data may lead to biased and invalid information. This study aims to analyze the initial data stages in data science research through a literature study approach. The research method employed is library research by reviewing 20 relevent scientific journal articles. The initial data stages examined include data collection, data preprocessing, data validation, and data presentation. The results of the study indicate that although each research has different approaches and contexts, the initial data stages generally share the same principles, namely ensuring data quality, accuracy, and reliability prior to further analysis. The utilization of information technology and data-based systems has been proven to improve effciency and accuracy in managing initial data. This study is expected to serve a conceptual reference for researchers in understanding the importance of initial data stages as the foundation of data science research.

Keywords : *Data Science, Initial Data, Data Preprocesssing, Literature Study* Averse.

INTRODUCTION

The rapid development of information and communication technology has encouraged the increasing use of data in various sectors of lite. Data has become a fundamental component in supporting decision-making processes, policy formulation, and scientific development. In the academic context, data plays a vital role as the primary foundation of scientific research, particularly in the field of data science, which focuses on data processing and analysis to generate valuable information (Turtiantoro, 2021).

Data science is an interdisciplinary field that integrates statistics, computation, and domain knowledge to extract insights from data. The succes of data science implementation largely depends on the quality of the data used. Incomplete, invalid, or poorly structured data may lead to inaccurate analytical results. Therefore, the initial data stages, starting from data collection to data preproccessing, are fundamental aspects that cannot be neglected in data science research (Thamrin et al , 2020).

In both qualitative and quantitative research, initial data management has different characteristic but is based on the same prinsiples of accuracy and reliability. Qualitative research empphasizes understanding context and meaning through descriptive data, while quantitative research focuses, both research findings are scientifically accountable (Turtiantoro, 2021).

The application of information technology in initial data management continues to expand across education, government, and community service sectors. Various studies indicate that the use of web-based information systems, digital media, and data processing applications can enhance efficiency, accuracy, and data security. In the educational field, digital learning media not only support teaching and learning processes but also generate leaning data that can be analyzed to improve education quality (Hidayatullah et al, 2025).

Furthermore, the application of data science through literature studies contributes significantly to enriching both theoretical and practical perspectives. A literuture stundy allows researchers to integrate findings from previous studies to obtain a comprehensive understanding of initial data stages in data science. By reviewing scientific articles, researchers can identify patterns, similarities, and differences in approaches to initial data management (Zunaidi,2024).

Based on the above discussion, this study aims to analzed the initial data stages in data science research through a literature study of 20 scientific journal articles. This analysis is expected to provide a conceptual contribution to understanding the importance of initial data management as the primary foundation of data science research.

METHOD

This study employs a literature study (libarary research) approach to analyze and synthesize the initial data stages in data science research. The literature study method was chosen because it enables researchers to systematically review, compare, and integrate findings from previous studies to obtain a comprehensive understanding of the research topic (Turtiantoro,2021).

The data sources consist of 20 scientific journal articles relevant to data science, data management, data preprocessing, and the application of information technology across various sectors. These articles were obtained from national and international journal based information systems (Hidayatullah et al, 2025; zunaidi, 2024).

Data collection was conducted through several stages. The first stage involved articles indentification, selecting journal articles that aligned with the research focus on initial data stages in data science.The second stage was articles classification, grouping articles based on exammining research objectives, methodologies, and findings related to initial data management (Thamrin et al, 2020).

Data analysis was performed using content analysis techniques. Key information related to data collection, preprocessing, validation, and data presentation was extracted from each articles. The findings were then compared to identify similarities, differences, and general patterns in initial data management practices in data science research (Turtiantoro, 2021).

To enhance the clarity and accuracy of the analysis, the results of the literature study were synthesized and presented in tabular form. The synthesis table includes information on authors, publication years, research focus, and key findings related to initial data stages. This approach facilitates a clearer undurstanding of each articles contribution to the study (Zunaidi, 2024).

RESULTS AND DISCUSSIONS

Results

Based on the literature review of 20 scientific journal articles, it was found that the initial data stages are fundamental elements in data science research. These stages include data collection, data preprocessing, data validation, and data presentation prior to further analysis. All reviewed articles emphasize that the quality of initial data significantly affects the accuracy of analytical result and research conclusion.

Initial data collection

Initial data collection is the first and most decisive stage in data science research. In the educational sector, data is commonly obtained from digital learning activitles, video-based instruction, and interactions within online learning platforms. Such data represent learning behaviors and serve as the basis for improving educational quality (Hidayatullah et al, 2025).

In community service and social development contexts, data collection is conducted through field observations, interviews, and documentation. This approach enables researchers to obtain contextual data that accurately reflect real community conditions. The use of information technology in data collection has been shown to improve recording efficiency and data accuracy (Zunaidi, 2024).

In governmental and village-level contexts, data is collected through digital administrative systems covering population data, public services, and local potential. Data-based information systems facilitate structured and sustainable data collection processes (Eko, 2020).

Data preprocessing and management

Data preprocessing is a major focus in most of the reviewed studies. This stage ensures that data are clean, consintent, and ready for analysis. Several articles highlight the importance of data cleaning, duplicate removal, and handling missing data to avoid analytical bias (Turtiantoro, 2021).

In information system-based research, preprocessing is often automated through web-based applications, reducing manual input errors and improving efficiency. In qualitative research, data preprocessing is conducted through data reduction and thematic categorization to focus analysis on relevent information (Thamrin et al, 2020, Zunaidi,2024).

Data validation and presentation

Data validation is essential to ensure the credibility of data used in research. In quantitative research, validation involves consistency checks and completeness verification, while qualitative research employs triangulation techniqeus (Turtiantoro, 2021).

Data presentation through tables, charts, and visualizations is widely applied to enhance data interpretation. Effective visualization helps researchers undurstand data characteristics and communicate findings more clearly (Hidayatullah et al, 2025).

Overall, the literature indicates that symtematic and technology-based initial data management is crucial for producing high-quality data in data science research.

No	author(s)	Year	Article Title	Key Findings
1	Turtiantoro	2021	*Implementasi Dana Desa di Desa Lemahireng*	Weak data documentation affects inter-organizational communication
2	Tambunan, B. H., Sihombing, M., & Harahap, R. H.	2020	*Peranan Dana Desa dalam Pengembangan Indeks Desa Membangun*	Data-based indicators influence village development prioritization
3	Thamrin, D. A., Habibi, B., Sari, D. P., & Athief, F. H. N.	2020	*Analisis Strategi Islamic Collective Entrepreneurship (ICE)*	Structured data stages support village economic empowerment
4	Somali, S. G.	2021	*Tata Kelola Pemerintahan Desa*	Governance effectiveness depends on accurate administrative data
5	Eko, S.	2020	*Regulasi Baru, Desa Baru*	Data strengthens village autonomy and governance
6	Septiyanto, D.	2021	Reformasi Birokrasi Desa Panggungharjo	Bureaucratic reform requires transparent data management
7	Puspitasari, C., et al.	2022	*Pola Relasi Pemerintah Desa*	Data-driven BUMDes strengthens village–community relations
8	Ariadi	2021	*Partisipasi Aktor Desa dalam Pembangunan*	Stakeholder data improves development planning
9	Hatuwe, R. S. M., et al.	2021	*Modernisasi terhadap Perubahan Sosial Masyarakat Desa Namlea*	Social change analysis depends on structured social data
10	Astika, A. N., & Subawa, N. S.	2021	*Evaluasi Pembangunan Desa Berdasarkan Indeks Desa Membangun*	Weak documentation limits IDM-based evaluation
11	Kurniawan	2021	*Evaluasi Dampak Dana Desa terhadap Infrastruktur Desa*	Infrastructure outcomes rely on government data quality

12	Yulitasari & Tyas, W. P.	2020	*Dana Desa dan Status Desa di Provinsi Jawa Tengah*	Village status changes are not solely funding-dependent
13	Benany, F. E., & Widyastuty, A. S. A.	2020	*Analisis Desa Tertinggal di Kecamatan Maurole*	Data gaps hinder development intervention
14	Zunaidi	2024	*Service Learning Berbasis Teknologi dan Inovasi*	Data supports sustainable community empowerment
15	Hidayatullah, R., et al.	2025	*Efikasi Penggunaan Video Live dalam Pembelajaran Daring*	Learning analytics depend on digital interaction data
16	Chandra & Triwidaryanta	2022	*Peran Lembaga Adat dalam Pemerintahan Desa*	Institutional data imbalance affects governance
17	Gultom	2021	*Perencanaan Pembangunan Desa*	Development planning requires integrated village data
18	Sirabura, S. N., & Yunanto, S. E.	2022	*Demokratisasi melalui Tata Kelola Dana Desa*	Transparency improves through data openness
19	Paais, L. S.	2021	*Keragaman Sosial dan Pembangunan Desa*	Sociocultural data enhances policy accuracy
20	Nasrullah, I.	2021	*Governmentality menurut Foucault*	Power relations are reflected in data governance

CONCLUSSION

Based on the literature study of 20 sclentific journal articles, it can be concluded that the initial data stages play a vital role in data science research. These stages include data collection, preprocessing, validation, and presentation, all of which directly influence the accuracy and relliability of research outcomes.

Despite differences in research contexts and applications, the principles of initial data management remain consistent. Systematic, structured, and technology-based data management significantly enahances data quality and minimizes analytical bias. Therefore, a comprehensive understanding of initial data stage is essential for researchers to produce credible and high-quality data science research.

ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to all researchers and authors whose works were reviewed and analyzed in this study. Their contributions have provided valuable insights and served as an essential foundation for this research. Appreciation is also extended to academic institutions and journal publishers that made these scientific articles accessible. Finally,

the authors acknowledge the support and guidance received from colleagues and reviewers who contributed constructive feedback to improve the quality of this study.

REFERENCES

Astika, A. N., & Subawa, N. S. (2021). Evaluasi pembangunan desa berdasarkan indeks Desa membangun.

Benany F. E., & Widyastuty, A. S. A. (2020) Analisis desa tertinggal di kecamatan maurole.

Chandra, & Triwidaryanta. (2022) peran lembaga adat dalam pemerintahan desa.

Eko, S. (2020). Regulasi baru, yogyakarta:pustaka pelajar.

Gultom. (2021). Perencanaan pembangunan desa

Hatuwe, R. S. M., Tuasalmony, P., Firdaus, R., Riyantika, F., & Wahyuningsih, S. (2025) Efikasi Penggunaan video live dalam pembelajaran daring.

Kurniawan. (2021). Evaluasi dampak dana desa terhadap pembangunan infrastruktur desa di indonesia.

Nasrullah, I. (2021). Governmentality menurut foucault

Paais, L. S. (2021). Keragaman agama, etnis, bahasa, dan pembangunan desa.

Puspitasari, C., surnarjono, perdana p., & Mardhika, J.G. (20220. Pola relasi pemerintahan Desa.

Septiyanto, D. (2021). Reformasi birokrasi Desa panggungharjo

Sirabura, S. N., & Yunanto, S. E. (2022). Demokratisasi melalui tata kelola dana desa.

Somali, S. G. (2021). Tata kelola pemerintahan desa.

Tambunan, B. H., Sihombing, M., & Harahap, R. H. (2020). Peranan dana desa dalam pengembangan indeks Desa Membangun

Thamrin, D. A., Habibi, B., Sari, D. P., & Athief, F. H. N. (2020). Analisis strategi Islamic collective Entrepreneurship (ICE).

Turtiantoro. (2021). Implementasi dana desa di Desa Lemahireng.

Yulitasari, & Tyas, W. P. (2020). Dana desa dan status desa di provinsi Jawa Tengah.

Zunaidi. (2024). Service learning berbasis teknologi dan inovasi.