buildings

MDPI

Article

Toward Standardised Construction Pipeline Data: Conceptual Minimum Dataset Framework

Elrasheid Elkhidir [1,*], James Olabode Bamidele Rotimi [1], Tirth Patel [2], Taofeeq D. Moshood [1] and Suzanne Wilkinson [3]

[1] School of Built Environment, Massey University, Auckland 0632, New Zealand; j.rotimi@massey.ac.nz (J.O.B.R.); t.moshood@massey.ac.nz (T.D.M.)
[2] Department of Civil and Natural Resources Engineering, University of Canterbury, Christchurch 8041, New Zealand; tirth.patel@canterbury.ac.nz
[3] Faculty of Design & Creative Technologies, Auckland University of Technology, Auckland 1010, New Zealand; suzanne.wilkinson@aut.ac.nz
* Correspondence: e.elkhidir@massey.ac.nz

Abstract

The construction industry is a cornerstone of New Zealand (NZ)'s economic growth, yet strategic infrastructure planning is constrained by fragmented and inconsistent pipeline data. Despite the increasing availability of construction pipeline datasets in NZ, their limited clarity, interoperability, and standardisation impede effective forecasting, policy development, and investment alignment. These challenges are compounded by disparate data structures, inconsistent reporting formats, and semantic discrepancies across sources, undermining cross-agency coordination and long-term infrastructure governance. To address this issue, the study begins by assessing the quality of four prominent pipeline datasets using Wang and Strong's multidimensional data quality framework. This evaluation provides a necessary foundation for identifying the structural and semantic barriers that limit data integration and informed decision-making. The analysis examines four dimensions of data quality: accessibility, intrinsic quality, contextual relevance, and representational clarity. The findings reveal considerable inconsistencies in data fields, classification systems, and levels of detail across the datasets. Building on these insights, this study also develops a conceptual minimum dataset (MDS) framework comprising three core thematic categories: project identification, project characteristics, and project budget and timing. The proposed conceptual MDS includes unified data definitions, standardised reporting formats, and semantic alignment to enhance cross-platform usability and data confidence. This framework applies to the New Zealand context and is designed for replication in other jurisdictions, supporting the global push toward open, high-quality infrastructure data. The study contributes to the construction informatics and infrastructure planning by offering a practical solution to a critical data governance issue and introducing a transferable methodology for developing minimum data standards in the built environment to enable more informed, coordinated, and evidence-based decision-making.

Keywords: construction projects data; pipeline projects data; construction data quality; projects data; construction data standardisation

Academic Editor: Fan Lei

Received: 9 July 2025
Revised: 28 July 2025
Accepted: 5 August 2025
Published: 7 August 2025

Citation: Elkhidir, E.; Rotimi, J.O.B.; Patel, T.; Moshood, T.D.; Wilkinson, S. Toward Standardised Construction Pipeline Data: Conceptual Minimum Dataset Framework. *Buildings* **2025**, *15*, 2797. https://doi.org/10.3390/buildings15152797

1. Introduction

Infrastructure development is widely recognised as a key enabler of economic growth, social well-being, and environmental sustainability. Public and private sectors around

the world invest heavily in infrastructure assets to support national productivity, enhance resilience, and meet long-term service demands. Effective planning for these investments requires access to structured and reliable pipeline data. Such data facilitates evidence-based decision-making by enabling stakeholders to forecast demand, allocate resources efficiently, and coordinate delivery timelines across agencies and sectors. However, despite its strategic importance, construction pipeline data are often fragmented, inconsistently reported, and difficult to integrate across platforms. These limitations constrain the ability of governments and industry to make timely, well-informed planning and investment decisions.

In many countries, infrastructure pipeline datasets are maintained by multiple institutions using divergent reporting formats, classification systems, and data structures. This heterogeneity impedes data interoperability, reduces transparency, and weakens the consistency of sectoral forecasts. The Global Infrastructure Hub has identified these issues as systemic challenges, characterising pipeline data as siloed, scattered across disconnected platforms, and difficult to compare or consolidate [1]. In response, a number of jurisdictions have introduced structured data frameworks to improve reporting consistency. For example, the Scottish Government's Construction Pipeline Forecast Tool employs a fixed set of fifteen standard data attributes to enable uniform reporting, enhance public–private collaboration, and support market signalling and strategic planning [2]. Such initiatives illustrate how harmonised pipeline data can improve governance, market confidence, and sector coordination.

New Zealand provides a pertinent context in which to explore these challenges. Although several national datasets are publicly accessible and offer forward-looking insights into planned construction activity, they differ considerably in scope, granularity, and structure. Some datasets are limited to public sector works, while others incorporate both public and private initiatives. Most focus on early project discovery to support investment awareness and preliminary planning, rather than providing technical, procurement, or delivery-level detail [3]. Across these datasets, inconsistencies in reporting logic, classification systems, and field definitions undermine the ability to perform cross-dataset comparisons or synthesise comprehensive market insights. These limitations have been acknowledged in [4], which calls for improved data quality and standardisation to support informed, coordinated infrastructure planning. As infrastructure systems become increasingly complex and delivery timeframes more compressed, the absence of a unified data framework in New Zealand presents a critical gap that constrains planning effectiveness and data reliability. Reliable NZ pipeline data is essential for supporting integrated infrastructure delivery and long-term investment planning [5]. For policymakers, such data enables the identification of funding priorities, capacity bottlenecks, and regional delivery needs in the NZ context. For private sector stakeholders, consistent data supports workforce planning, resource allocation, and commercial risk assessment [6]. In contrast, fragmented or incomplete data reduce transparency, erode stakeholder confidence, and introduce misalignment between market activity and policy intent. The need for consistent, interoperable data frameworks is particularly pressing as governments scale up infrastructure investment to meet national development and resilience goals.

Although classification systems such as the RIBA Plan of Work and Statistics New Zealand's construction typologies provide structured guidance for project phases and asset categories [7,8], they do not specify a minimum data structure for reporting across pipeline datasets. This lack of a common standard presents both conceptual and practical barriers. Theoretically, it inhibits the development of a shared understanding of core data requirements for strategic planning across the infrastructure lifecycle. Practically, it forces stakeholders to triangulate from multiple sources, resulting in inefficiencies, du-

plicated effort, and reduced data credibility. The diversity of infrastructure project types, delivery models, and data custodians further reinforces the urgency of a harmonised and standardised reporting approach.

This study responds to this gap by developing a conceptual minimum dataset (MDS) framework for construction pipeline reporting. Using New Zealand as the empirical focus, the research adopts a comparative analysis approach grounded in a multiple case study design [9] and systematic cross-case comparison [10]. Four nationally significant pipeline datasets are assessed using Wang and Strong's multidimensional data quality framework, which evaluates accessibility, intrinsic quality, contextual relevance, and representational clarity. The study places particular emphasis on the contextual dimension, assessing how well existing datasets support planning, coordination, and investment decisions. Based on this analysis, a harmonised MDS framework is proposed, structured around three core thematic categories: project identification, project characteristics, and project budget and timing. The framework introduces standardised definitions, reporting structures, and terminology to enhance data interoperability, usability, and policy relevance. The scope of this research is limited to pipeline-level data that support early-stage project communication and strategic planning, rather than technical documentation or execution-phase metrics. Although developed within the New Zealand context, the proposed framework offers a transferable and practical methodology for improving infrastructure data governance and cross-sector integration in other jurisdictions.

2. Background

Data quality could be defined as data that is "fit for use", and that could satisfy the expectations and requirements of the end users and computerised systems [11]. For construction pipeline data used across multiple stakeholders for planning and investment decisions, this fitness-for-use concept becomes particularly critical as different users, ranging from government planners to private contractors, may have varying quality requirements and specifications for the same project information [12]. Assessing data quality involves nuanced exploration encompassing both subjective and objective methodologies. Subjective approaches draw on the perceptions and experiences of individuals involved in the dataset lifecycle, including collectors, providers, and end-users [13]. In contrast, objective assessments delve into the meticulous evaluation of specific data metrics within the datasets under scrutiny [14,15]. While some researchers contend that data quality hinges on the unique contextual application, as [16] argued, it is crucial to emphasise that subjective data assessments can be further classified as task-dependent or task-independent. Task-dependent assessments are tailored to specific organisational modes or requirements, whereas task-independent evaluations are conducted irrespective of the contextual or task considerations [17].

The foundational principles shaping hierarchical data quality dimensions, with a focus on end-user perspectives, were established by [18]. This framework remains particularly relevant for pipeline data contexts as it accommodates multiple stakeholder perspectives and varying quality requirements across different organisational boundaries [19]. Within their framework, Ref. [18] delineated four primary categories, namely intrinsic, contextual, representational, and accessible data quality, with 20 distinct data quality dimensions distributed among these identified categories. Data quality dimensions refer to characteristics that, when properly measured, reflect the overall standard or reliability of the data [20]. Figure 1 visually encapsulates the primary data quality categories and dimensions as conceptualised by [18].

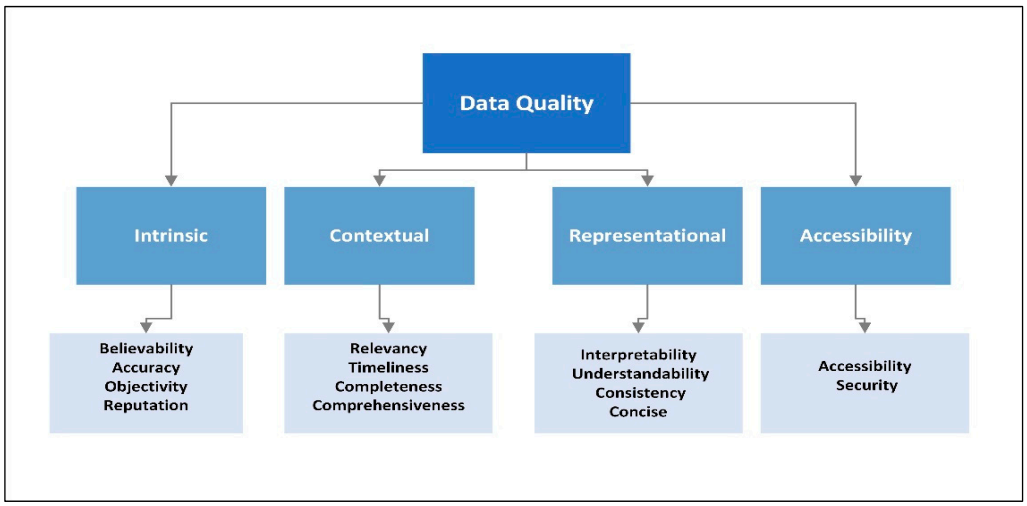

Figure 1. Data Quality Assessment Dimensions. Adapted from Ref. [18].

Over time, researchers have expanded data quality dimensions beyond Wang and Strong's framework. Notable contributions include currency and timeliness [21], relevancy to user requirements [22], and the influence of users' expertise levels [23]. However, these general frameworks do not fully address the specific challenges of construction pipeline data, where temporal evolution, multi-provider integration, and stakeholder diversity create unique quality requirements [24].

The construction industry faces a significant challenge in data selection due to the proliferation of high-level data providers, each operating with distinct reporting practices and pursuing different end goals. This fragmented landscape creates complexity when organisations attempt to identify the most suitable dataset for their operational needs. Some providers focus on real-time project tracking, while others emphasise historical trend analysis or regulatory compliance metrics. The varying methodologies, data formats, and reporting frequencies across these providers make direct comparisons difficult and complicate the decision-making process [5,23]. As the number of data providers increases, there are higher chances of ambiguity around the circumstances and the environment surrounding the actual collection, processing, and communication of the dataset, rendering difficulties in understanding the data characteristics and data quality dimensions such as accuracy, timeliness, and completeness [25]. Recent studies confirm the persistence of challenges in data integration complexities as the number of participating organisations increases [12]. Ref. [26] highlighted several data challenges across different platforms, including different languages and translation schemes, nomenclatural problems, classification differences, and name and number mistakes. Furthermore, different databases may exhibit different levels of breakdown and granularity, adding a higher level of complexity [27].

The same data may be collected across several platforms, but with different user specificity, dimensions, and quality attributes. Ref. [25] quoted the example that a manager may seek to measure cost as a monetary value, while traders perceive it in the timeliness of the data in terms of the opportunity or competitive advantage provided by the data. Similarly, Ref. [28] contended that different actors view and interpret the same dataset in different semantics, although all may be correct. The authors addressed the challenges with simple data strings, such as ambiguities that could arise in assigning identifiers and specific names to entities without confusion or overlaps, aggregation, and relationships.

Moreover, the existence of a common vocabulary and metadata, along with a shared understanding of the data typologies and structure, is crucial for integrating and using multi-sourced data [26,29,30]. While government data standards have advanced significantly in other domains, construction pipeline data lacks comparable standardisation

frameworks [19]. In designing databases, Ref. [25] suggested consolidating the different views and attributes that are related to a single unified global view to eliminate redundancy and inconsistency and ultimately achieve a higher data quality for the dataset. Ref. [26] considered that the standardisation of data across multiple platforms was necessary due to the complexity and high variability of the data in the biodiversity field and the data entry and usage processes involving several people or entities. Those two features are also found within the realm of project data. Ref. [26] further argued that enhancing information exchange between databases relies on a standardised data structure, such as the type and number of fields and columns. To achieve standardisation across databases, Ref. [26] suggested developing and agreeing on the semantics and syntax for each data field/column, using appropriate and well-defined classifications/standards for each field/column, and acknowledging exceptions. The goal would be to reach a consensus on a defined set of classes and properties that adequately reflect and represent the required data and its associated level of information domains and fields [27].

The construction industry has witnessed several initiatives aimed at standardising pipeline data, though these efforts have produced fragmented results across different jurisdictions and sectors. The UK's Infrastructure and Projects Authority pipeline database and Australia's Infrastructure Australia project repository exemplify this fragmentation, as each has developed distinct reporting frameworks tailored to their national requirements rather than facilitating cross-provider data integration. Building Information Modelling (BIM) standards have proven effective for managing detailed project execution data; these frameworks do not address the high-level project discovery information that stakeholders require for strategic planning and market intelligence purposes [24]. Ref. [31] highlighted that construction data integration faces persistent challenges, including data heterogeneity, source credibility differences, system compatibility problems, and varying interpretation approaches. Government data inconsistency presents a significant obstacle to effective pipeline project management, even within single jurisdictions where coordination should theoretically be straightforward. Different agencies often maintain incompatible data structures, with some using asset-based classifications while others organise information by geographic regions or regulatory categories. Reporting timeframes vary dramatically, with agencies updating their datasets on monthly, quarterly, or annual cycles that rarely align with one another. These disparate classification methods create additional complexity, as one agency might categorise a project as "transmission infrastructure" while another labels the same asset as "distribution network equipment." The resulting data fragmentation forces project planners and investors to manually reconcile conflicting information sources, leading to delayed decision-making and potential gaps in comprehensive risk assessment. This internal inconsistency undermines the efficiency that unified jurisdictional oversight should provide and creates unnecessary barriers to strategic planning and capital allocation decisions [12]. This situation reveals a clear gap between existing technical standards and the specific requirements for pipeline information sharing across organisational boundaries.

3. Methodology

This study adopts a structured six-stage methodological framework, adapted from [32]'s conceptual framework development process, as shown in Figure 2. The approach integrates multiple case study methodologies [9] with systematic cross-case comparative analysis techniques [10], tailored to the evaluation of construction pipeline datasets in New Zealand. The objective is to systematically assess the structural, semantic, and functional quality of pipeline data sources and to formulate a harmonised minimum dataset (MDS) to support standardisation and interoperability.

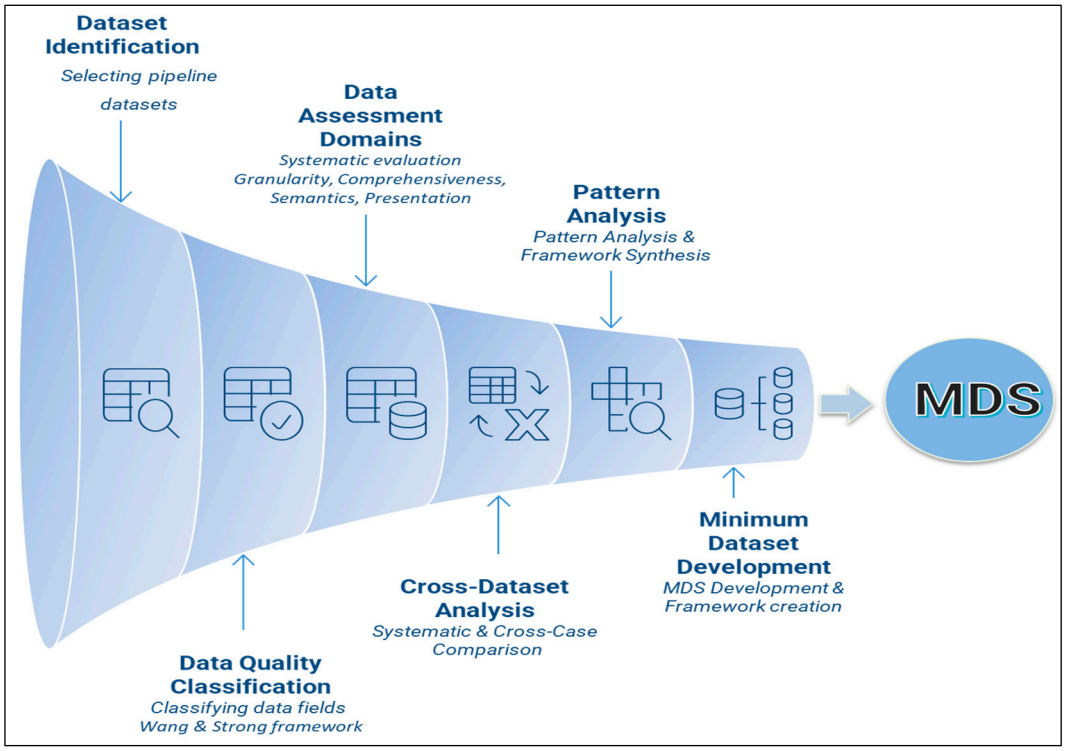

Figure 2. Methodological approach for the development of a standardised construction pipeline minimum dataset.

3.1. Stage 1: Dataset Identification and Selection

The first stage involved identifying and selecting key construction pipeline datasets that are widely referenced across the New Zealand infrastructure planning landscape. The selection criteria were based on three factors: the frequency of use by industry and government stakeholders, the ability to represent different client segments, and the comprehensiveness of coverage across various types of infrastructure projects nationwide. Based on these criteria, four datasets were selected for comparative analysis. The first dataset, published by a central government agency, consolidates pipeline information from both public and private sector contributors. Its principal objective is to support national-level planning and policymaking. In addition, it offers industry stakeholders access to forward-looking insights and demand projections to inform their strategic planning. The second dataset, produced by a well-established private New Zealand firm, compiles information on different types of projects, including infrastructure, residential, and commercial projects. Owing to its wide industry reach, it serves a diverse user base that includes local and central government agencies, private developers, and construction firms. The third dataset is released quarterly by a government agency tasked with delivering capital works, maintenance, and redevelopment projects. It is primarily intended to communicate upcoming project activity to the construction sector, thereby supporting alignment and delivery preparedness. Finally, the fourth dataset is derived from the Long-Term Plans (LTPs) of local authorities across New Zealand. These plans outline infrastructure investment priorities over a ten-year horizon and are reviewed triennially. As such, the dataset provides a high-level, regionally grounded perspective on anticipated infrastructure delivery needs.

3.2. Stage 2: Data Quality Classification

The second stage involved a systematic categorisation and comparative evaluation of the selected construction pipeline datasets using Wang and Strong's multidimensional data quality framework. This framework was selected for its ability to comprehensively

assess the reliability, usability, and decision-support capacity of data systems. Given the structural and content heterogeneity across the four datasets, the evaluation focused on four interrelated dimensions: accessibility, intrinsic quality, contextual relevance, and representational clarity. Accessibility was assessed by examining whether each dataset was publicly available or restricted through subscription, as well as by evaluating the mode of access, specifically, whether the data were disseminated through push mechanisms or retrieved through pull-based systems. These aspects were crucial for determining the openness and reach of each dataset among diverse stakeholder groups. Intrinsic quality was examined through three sub-criteria: believability, accuracy and objectivity, and usability. Believability was assessed based on the credibility of data sources, while accuracy and objectivity were evaluated by reviewing the transparency and robustness of the data collection and compilation methodologies. Usability was analysed in terms of dataset scope, including whether both public and private sector projects were captured, whether future pipeline projections and historical records were included, and how frequently the datasets were updated to ensure data currency. Contextual relevance was determined by reviewing the richness and consistency of project-specific attributes such as project name, ownership, location, type of work, development stage, approval status, procurement and delivery methods, funding information, estimated costs, and scheduling data. In addition, dataset completeness was examined by analysing the number of reported fields and the proportion of fully populated, partially filled, or missing entries. Representational clarity was assessed by considering the structure and usability of the datasets, including file format, interface accessibility, logical organisation, and the overall ease with which data could be interpreted and applied by end-users. Together, these assessments offered a structured and integrated basis for understanding the relative strengths and limitations of each dataset and informed the development of a harmonised minimum dataset framework in the next stage of the study.

3.3. Stage 3: Data Assessment Criteria

The categorised data was then analysed and assessed using six comparison criteria, all of which were derived from the literature and tailored to suit the characteristics of construction pipeline datasets. These criteria include (1) information depth and precision (granularity), (2) data completeness, (3) terminology consistency and semantics, (4) format and usability, (5) update frequency (timeliness), and (6) stakeholder relevance (fitness-for-use). The choice of the criteria was grounded in established literature, as shown in Table 1.

The criteria shown in the table, selected for their theoretical and empirical relevance, were operationalised using a five-point scoring scale. A broad scoring system (range 1–5) was utilised to support transparency and replicability. Table 2 shows the developed scoring framework with defined thresholds for each criterion based on the nature and characteristics of the data fields. Each criterion was evaluated through multiple sub-components: granularity (12 attributes), completeness (4 attributes), semantics (7 attributes), format (3 attributes), timeliness (1 attribute), and stakeholder relevance (6 attributes). All the datasets were independently evaluated by two researchers across all six comparison criteria identified in Table 1. Numerical scores were used to support comparative interpretation and synthesis. Detailed discussions to reconcile any differences between the two researchers' scoring outputs were conducted to reach consensus and enhance the inter-rater reliability.

Table 1. Comparison Criteria Selection.

Comparison Criterion	Brief Description	Supporting Literature
Information Depth and Precision (Granularity)	Level of detail is provided for project types, lifecycle phases, and locations.	[27]
Data Completeness	Presence of key fields, blank and missing data across columns.	[14,15,20]
Terminology Consistency and Semantics	Clarity, consistency, and classification of terminology and field names.	[26,28]
Dataset Format and Usability	File structure, interpretability, filterability, and date formats.	[19,26]
Update Frequency (Timeliness)	How frequently is the dataset updated (e.g., quarterly, biannually)?	[12,21]
Stakeholder Relevance (Fitness-for-Use)	The extent to which the dataset supports different stakeholder needs	[12,16,22]

Table 2. Scoring Scale applied to four construction pipeline datasets using six core criteria.

Score = 1 (Very Poor)	Score = 2 (Poor)	Score = 3 (Moderate)	Score = 4 (Good)	Score = 5 (Excellent)
Information Depth and Precision Assesses the extent and granularity of information provided across project attributes such as type of work, lifecycle phases, geographic location, procurement method, and financial values.				
Minimal identifiers only (e.g., project name). No lifecycle, location, or classification details.	Broad or vague labels. No structured detail or defined terms.	Some detailed fields, like location or phase, are inconsistently applied.	Detailed info across most attributes, including lifecycle, sub-region, and categories.	High-resolution coverage across all fields with structured naming, defined lifecycle phases, and precise procurement and cost bands.
Data Completeness Measures the proportion of fields populated and the presence of blank or missing data.				
<20% populated. Severe gaps limit usability.	20–39% populated. Many key fields are missing.	40–59% populated. Partial reliability.	60–79% populated. Mostly complete.	>80% populated. Strong analytical readiness.
Terminology Consistency and Semantics Evaluates the clarity and standardisation of labels, categories, and field names.				
Terms are ambiguous or inconsistent. Conflicting field usage.	Partial labelling but classification mismatches persist.	Some structured naming, with gaps or inconsistencies.	Mostly consistent terms and field names.	Fully standardised terms (e.g., ANZSIC06, RIBA), interpretable across datasets.
Dataset Format and Usability Evaluates whether the dataset is machine-readable, filterable, and easy to use.				
PDF or image formats only. No filtering or export.	Usable formats exist, but they are poorly structured.	Excel/CSV with basic formatting.	Well-structured Excel or Tableau with filter features.	Interactive dashboards, downloadable formats, and API access.

Table 2. *Cont.*

Score = 1 (Very Poor)	Score = 2 (Poor)	Score = 3 (Moderate)	Score = 4 (Good)	Score = 5 (Excellent)
Update Frequency (Timeliness) Reflects how current and frequently refreshed the dataset is.				
No updates in 3+ years or unknown cycle.	Updated every 2–3 years only.	Annual or biannual updates.	Quarterly updates with regularity.	Monthly or continuous updates with push notifications.
Stakeholder Relevance (Fitness-for-Use) Evaluates how well the dataset serves diverse users across the public and private sectors.				
Developed for internal use only.	Limited scope, applicable to narrow segments.	Serves some users with limited lifecycle visibility.	Serves most users and includes private/public sector projects.	Broad coverage of stakeholders and phases, usable for planning, delivery, and policy.

3.4. Stage 4: Cross-Dataset Comparative Analysis

A systematic comparison was carried out across the four selected construction pipeline datasets, guided by the six assessment criteria outlined in Stage 3. The analysis involved a detailed, side-by-side evaluation of all parameters listed in the comparison table, including aspects such as accessibility, data source reliability, contextual detail at the project level, field completeness, classification consistency, and ease of use. All datasets were evaluated and scored against the five-point scoring scale shown in Table 2 to ensure consistency and provide a common framework for identifying commonalities and variances. The objective was to uncover patterns, discrepancies, and gaps across datasets in how information is structured and aligned with sector needs. This comparative analysis underpins the findings presented in the next section and lays the groundwork for the synthesis undertaken in Stage 5.

3.5. Stage 5: Pattern Analysis and Framework Synthesis

Patterns of consistency and divergence across the four datasets were analysed to evaluate the significance of specific data fields. Commonalities were interpreted as indicators of core informational value and were used to establish the foundational structure of a harmonised minimum dataset. Scoring patterns from the scoring scale were used to identify which data attributes consistently achieved high completeness, clarity, and utility, indicating their suitability for inclusion in a baseline standard. Conversely, fields with persistent deficiencies across datasets highlighted standardisation gaps and deficiencies. The outputs from the cross-dataset analysis were compiled and synthesised to identify shared attributes and recurring variances.

3.6. Stage 6: Minimum Dataset Development

Informed by the results from Stage 5, a minimum dataset (MDS) was developed to represent the essential data fields commonly reported across pipeline sources. The MDS is organised into four thematic categories: (1) project identification, (2) project type and scope, (3) funding and delivery, and (4) scheduling and timing. Standardised definitions and reporting formats were introduced to address semantic inconsistencies and support interoperability and comparability between datasets.

4. Results

This section presents a systematic evaluation of construction pipeline datasets using the six-criterion scoring framework developed in the methodology. The analysis reveals significant performance disparities across granularity, completeness, semantics, format, timeliness, and stakeholder relevance dimensions, with implications for stakeholders relying on accurate, timely, and integrated infrastructure data. The findings highlight critical discrepancies between provider capabilities and market needs, with important implications for organisations seeking reliable construction pipeline intelligence.

4.1. Information Depth and Precision (Granularity)

Construction project datasets exhibit considerable variation in categorising and detailing work types. At one end of the spectrum, some datasets employ a highly granular approach, breaking down projects into as many as 16 distinct categories, while others use only two classification categories, with some using the owner's name as work type classification. A challenge arises in the representation of project location data. Certain datasets only specify broad geographic categories, such as cities or regions, whereas others offer varying degrees of granularity, from regional designations to exact coordinates.

The reported projects' lifecycles also varied significantly. While some datasets capture the full project lifecycle—including planning, procurement, and construction phases—others are more limited in scope. For instance, Dataset 3 focuses solely on the procurement stage, providing pipeline projections and tentative timelines without reporting any further phases. In contrast, the 10-Year Long-Term Plans (LTPs) released by local authorities tend to provide only high-level planning information, often lacking detailed project details and data. However, more details may be made available through other datasets, such as Dataset 1, once internal planning processes are finalised and the project has been confirmed. This fragmentation, along with the staggered release of information across different platforms, highlights the urgent need for a unified reporting framework to improve data consistency, enable cross-dataset integration, and support more effective sector-wide planning.

The way project funding status and estimated costs were reported varied widely across the datasets. Some datasets included information on whether funding had been confirmed, was still pending, or had already been secured. a few datasets appeared to only include projects that had already received funding approval. By contrast, high-level datasets like the 10-year plans from local authorities generally provided only the projected annual capital expenditure, without offering any indication of the certainty of funding. Three of the datasets grouped projects into value bands, but the number of bands and the size of their intervals differed significantly. Some datasets offered nine detailed categories, which allowed for more precise cost confidence, while others used fewer, broader bands. For example, a project valued at 1 billion dollars might be placed in a generic category like "over 25 million" in one dataset, whereas another would list it more accurately as "over 1 billion." Only one dataset, however, included a direct lump-sum estimate for the total expected cost of each project.

Dataset 1 reveals a fundamental mismatch between ambitious data collection goals and practical execution capabilities. While the dataset features the most comprehensive structural framework with 29 columns, four of these fields remain completely blank, indicating that the intended scope far exceeds what can be realistically achieved. This gap between design aspirations and operational reality not only undermines the dataset's completeness but also suggests that data collection efforts would benefit from more realistic scope planning that aligns field structures with actual captured capabilities. In contrast, Dataset 2 has a higher completion rate (0 blank columns), indicating a more realistic scope definition aligned with collection capacity. This pattern reveals that comprehensiveness

and completeness exist in tension, requiring careful balance in standardisation efforts. These variations are exacerbated by the granularity variations. More detailed categorisation systems entail greater data collection challenges, while simple data structures may achieve higher completion but jeopardise analytical depth.

To systematically evaluate the granularity performance, twelve key attributes were assessed using the scoring criteria illustrated in Table 2. Table 3 shows the detailed granularity assessment scores across the different relevant attributes and across the four datasets.

Table 3. Information Depth and Precision (Granularity) Scores.

Attributes	Dataset 1	Dataset 2	Dataset 3	Dataset 4
Project Identifiers	5/5	5/5	3/5	2/5
Agency Detail	4/5	4/5	3/5	3/5
Location Detail	5/5	4/5	2/5	3/5
Work Type Categories	5/5	4/5	1/5	3/5
Lifecycle Phases	5/5	5/5	2/5	1/5
Regulatory Approvals	1/5	2/5	1/5	1/5
Procurement Detail	5/5	1/5	1/5	1/5
Delivery Methods	5/5	1/5	3/5	1/5
Supplier Information	1/5	4/5	2/5	1/5
Funding Detail	4/5	1/5	3/5	3/5
Value Detail	3/5	5/5	3/5	4/5
Schedule Detail	5/5	1/5	2/5	1/5
Granularity	**4.00/5**	**3.00/5**	**2.17/5**	**2.00/5**

Dataset 1 scored the highest granularity score (4.00/5) across most information depth dimensions, particularly in work type categorisation (16 categories), location precision (coordinates included), and procurement detail. Dataset 2 demonstrated a balanced performance (3.00/5) with strong lifecycle coverage but limited procurement and delivery method detail. Datasets 3 and 4 scored significantly lower (2.17/5 and 2.00/5, respectively), reflecting their constrained scope and limited detail provision.

4.2. Data Completeness

Significant variances were perceived in the number of data fields or columns used in each dataset, highlighting discrepancies in the detail level and comprehensiveness of the datasets. For example, one dataset provided 29 data columns, while the others provided 14 and 9 columns. Among the 29 data columns reviewed, four were completely empty, with no entries recorded at all, and ten others had varying levels of missing data. In comparison, the remaining datasets were more complete, with none containing entirely blank columns and only one showing missing values in three columns. Other variances detected in Table 2 include the fact that not all datasets reported building consents and resource management consents.

Four key completeness dimensions were systematically evaluated using the scoring framework presented in Table 2. Table 4 presents the detailed completeness assessment across all datasets.

Datasets 2 and 3 achieved the highest completeness scores (4.25/5) through contrasting strategies. Dataset 2 had a higher scope, i.e., number of data fields or columns, but a lower focus on completeness when compared with dataset 3 (79% completion with 14 fields). Dataset 3 had a lower number of data fields or columns, but more focus on completeness (100% completion with 9 fields). Dataset 1's moderate score (3.00/5) reflects the ambitious scope challenges, with 34% of fields only partially populated despite comprehensive coverage. Dataset 4's poor performance (2.25/5) indicates systematic completion issues across all dimensions.

Table 4. Completeness Scores.

Attributes	Dataset 1	Dataset 2	Dataset 3	Dataset 4
Total number of columns	5/5 (29 columns)	4/5 (14 columns)	2/5 (9 columns)	2/5 (variable)
Blank field management	2/5 (14% blank)	5/5 (0% blank)	5/5 (0% blank)	2/5 (variable)
Partially populated columns	2/5 (34% partial)	4/5 (21% partial)	5/5 (0% partial)	2/5 (variable)
Overall completion rate	3/5 (52% complete)	4/5 (79% complete)	5/5 (100% complete)	3/5 (~60% complete)
Completeness	**3.00/5**	**4.25/5**	**4.25/5**	**2.25/5**

4.3. Terminology Consistency and Semantics

The studied datasets adopted a range of formats when presenting procurement related information, with considerable variance across the datasets. In some cases, the type of procurement, such as open tendering or direct award, was clearly stated, while in others, this information was missing. Project definition varied across the different datasets. All datasets assigned a project name, and some were assigned a project identification number. However, they were consistently different. Another example is the way work types were classified varied considerably across the datasets. Some providers used detailed classification systems with as many as 16 distinct categories, offering a clear breakdown of project types. Others relied on only two broad groupings, and in some cases, even used the project owner's name as a stand-in for the work type.

These semantic variations may inhibit data integration and render cross-comparison and triangulation challenging. Terminology variations, such as those seen in procurement scope, supplier information, and classification systems, highlight the need for standardisation and unified semantic approaches to enable effective and meaningful cross-comparison and triangulation.

To systematically assess semantic consistency challenges, seven key terminology and classification dimensions were evaluated using the scoring framework. Table 5 presents detailed semantic assessment across all datasets.

Table 5. Terminology Consistency and Classification Scores.

Attributes	Dataset 1	Dataset 2	Dataset 3	Dataset 4
Identifier consistency	3/5 (unique but different)	3/5 (unique but different)	2/5 (basic naming)	1/5 (name only)
Location format standards	4/5 (detailed, non-standard)	4/5 (street level detail)	2/5 (broad regions)	2/5 (suburb level)
Work type classification	3/5 (16 categories, non-ANZSIC)	3/5 (9 generic categories)	1/5 (2 basic categories)	2/5 (4–16 variable)
Lifecycle terminology	4/5 (clear phases)	4/5 (clear phases)	2/5 (procurement only)	1/5 (none reported)
Procurement standards	4/5 (4 clear methods)	1/5 (not reported)	1/5 (not reported)	1/5 (not reported)
Value structure consistency	2/5 (9 bands, no currency)	4/5 (lump sum estimates)	3/5 (7 moderate bands)	1/5 (no bands)
Date format standards	1/5 (quarters format)	4/5 (month/ year standard)	4/5 (month/ year standard)	1/5 (year only)
Semantics	**2.71/5**	**2.86/5**	**1.86/5**	**1.29/5**

The semantic assessment reveals standardisation challenges, with no dataset scoring more than 3.5/5. Dataset 2 performed the best (2.86/5) across the other datasets, through consistent date formats and value reporting, while Dataset 4 exhibited fundamental semantic inconsistencies (1.29/5) with varying terminology across local authorities.

4.4. Dataset Format and Usability

Dataset format and usability varied significantly across the sampled datasets, creating substantial barriers for data analysis and integration. Dataset 1 provided data in both Tableau and Excel formats, while Dataset 2 offered Excel and PDF formats, and Datasets 3 and 4 were only available in PDF format. The choice of file format directly impacted the usability of the datasets, as Datasets 1 and 2, available in Excel format, allowed users to easily apply data filters and create visualisations using standard spreadsheet software. In contrast, Datasets 3 and 4, provided only in PDF format, could not support data filtering or visualisation capabilities, significantly limiting their analytical utility. Dataset interpretability also varied considerably across the sample, with Datasets 1 and 2 considered easy to interpret, while Dataset 3 presented confusion between short-term and long-term project timelines. Dataset 4 exhibited the most significant interpretability challenges, with different semantics and nomenclature used across various local authorities, making cross-regional comparisons difficult. These format and usability differences create barriers for stakeholders attempting to integrate information from multiple sources, as users must adapt to different interfaces and interpretation methods, reducing efficiency in data analysis and comparison activities. Again, the identified variances in formats and usability contribute to further compounding the challenges of semantics and completeness variations discussed earlier. These variations could amount to a technical dimension and create systematic barriers extending beyond simple formatting differences and amounting to fundamental limitations.

To evaluate the technical barriers to data utilisation, three main format and usability attributes were systematically assessed using the scoring framework. Table 6 presents the detailed format assessment across all datasets.

Table 6. Dataset Format and Usability Scores.

Attributes	Dataset 1	Dataset 2	Dataset 3	Dataset 4
File Format Quality	5/5 (Tableau + Excel)	4/5 (Excel + PDF)	1/5 (PDF only)	1/5 (PDF only)
Data Filtering Capability	5/5 (easy filtering)	4/5 (Excel filtering)	1/5 (no filtering)	1/5 (no filtering)
Interpretability	5/5 (easy interpretation)	5/5 (easy interpretation)	2/5 (timeline confusion)	1/5 (varying terminology)
Format and Usability	**5.00/5**	**4.33/5**	**1.33/5**	**1.00/5**

The format and usability scores show a clear division between high-performing and low-performing datasets. Datasets with easily versatile and easy-to-read and manipulate formats (machine-readable), such as Datasets 1 and 2, achieved higher scores (5.00/5 and 4.33/5), while PDF-only datasets, such as Datasets 3 and 4, yielded lower scores (1.33/5 and 1.00/5). This performance gap, 4 points across the datasets, represents the largest variation observed across all assessment criteria.

4.5. Update Frequency (Timeliness)

Timeframes and reporting of schedules differed across the sampled datasets. Generally, each dataset adopted its unique time reporting format, including quarters, months, years, and years only. Additionally, some datasets included timelines for each stage of a project's lifecycle, while others focused only on specific phases, such as tendering or contract dates. A few datasets simply listed the year, without offering any details on the phases involved. This wide variation in how time-related information is reported reduces confidence in the data and makes it harder to track construction activity accurately. It also limits the ability to develop coordinated market insights or align project timelines across different datasets.

To assess update frequency and timeliness of the datasets, one timeliness attribute was evaluated using the scoring framework introduced in Table 2. Table 7 presents the detailed timeliness assessment across all datasets.

Table 7. Update Frequency (Timeliness).

Attributes	Dataset 1	Dataset 2	Dataset 3	Dataset 4
Update frequency	4/5 (quarterly)	5/5 (monthly)	3/5 (bi-annual)	1/5 (3 years)
Timeliness	**4/5**	**5/5**	**3/5**	**1/5**

Dataset 2 achieved excellent performance (5/5) with monthly updates, while Dataset 4 scored very poor (1/5) with triennial update cycles. This 4-point performance gap creates a 36-fold difference in update frequency, where Dataset 2 provides 36 updates for every single update from Dataset 4. The substantial variation in update frequency fundamentally impacts dataset utility for real-time market intelligence and coordinated planning activities, with datasets updating anywhere from monthly to once every three years.

4.6. Stakeholder Relevance (Fitness-for-Use)

The accessibility dimension was studied regarding the dataset's availability and the data retrieval method. The intrinsic dimension was considered a believability/reliability parameter in terms of the information sources, the accuracy and objectivity of the dataset, and the methods used to compile the dataset. The usability spectrum was studied regarding the private sector inclusion, historical data availability, and future pipeline publishing.

To evaluate fitness-for-use across stakeholder needs, six stakeholder relevance attributes were assessed using the scoring framework presented in Table 2. Table 8 presents scores of Stakeholder Relevance category across all datasets.

Table 8. Stakeholder Relevance (Fitness-for-Use) Scores.

Attributes	Dataset 1	Dataset 2	Dataset 3	Dataset 4
Public Accessibility	5/5 (fully public)	3/5 (subscription required)	5/5 (fully public)	5/5 (fully public)
Data Retrieval Methods	3/5 (pull only)	5/5 (push + pull)	5/5 (push + pull)	3/5 (pull only)
Source Credibility	4/5 (agency submissions)	4/5 (multiple sources)	3/5 (internal planning)	3/5 (internal planning)
Private Sector Inclusion	1/5 (public only)	5/5 (public + private)	1/5 (public only)	1/5 (public only)
Historical Data Availability	1/5 (none available)	5/5 (available)	3/5 (on request)	1/5 (none available)
Multi-User Applicability	5/5 (multi-purpose)	5/5 (multi-stakeholder)	2/5 (limited scope)	1/5 (internal use only)
Stakeholder Relevance	**3.17/5**	**4.50/5**	**2.50/5**	**2.33/5**

Dataset 2 achieved the highest performance (4.50/5) through comprehensive multi-stakeholder design, including private sector coverage, historical data availability, and diverse retrieval methods, despite subscription barriers. Dataset 1 scored moderately (3.17/5) with strong multi-purpose applicability but limitations in private sector inclusion and historical data. Datasets 3 and 4 performed poorly (2.50/5 and 2.33/5, respectively) with narrow scope and limited user applicability.

4.7. Synthesis of Findings

The systematic scoring across all six criteria reveals significant performance disparities that highlight a fragmented information landscape, also observed qualitatively. Table 9 presents a summary of the quantitative cross-dataset analysis across all assessment dimensions.

Table 9. Summary of Quantitative Cross-Dataset Analysis.

Assessment Criterion	Dataset 1	Dataset 2	Dataset 3	Dataset 4
Granularity	4.00/5	3.00/5	2.17/5	2.00/5
Completeness	3.00/5	4.25/5	4.25/5	2.25/5
Semantics	2.71/5	2.86/5	1.86/5	1.29/5
Format and Usability	5.00/5	4.33/5	1.33/5	1.00/5
Timeliness	4/5	5/5	3/5	1/5
Stakeholder Relevance	3.17/5	4.50/5	2.50/5	2.33/5
Overall Score	**3.65/5 (73%)**	**3.99/5 (80%)**	**2.52/5 (50%)**	**1.64/5 (33%)**

The four studied datasets may be purposely developed and individually logical. However, when considered collectively and comprehensively, a fragmented information landscape emerges. This is evident from the qualitative and quantitative cross-dataset analysis conducted. This demonstrates the need for a standardised minimum dataset that can accommodate diverse stakeholder needs and enable cross-dataset integration, triangulation and comparison. Table 10 displays the results and analysis attained from comparing the four datasets.

Table 10. Comparison of Project Data from Selected Providers.

Category	Parameters	Data Attributes	Dataset 1	Dataset 2	Dataset 3	Dataset 4
Accessibility Parameters	Accessibility	Availability	Public	Paid Subscription	Public	Public
		Data Retrieval	Pull (manual retrieval)	Push/Pull	Push/Pull	Pull
Intrinsic Data Quality Parameters	Believability	Information Sources	Agency Submissions	Open (all Sources)	Internal Planning Unit	Internal Planning Units
	Accuracy/Objectivity	Data Collection Methods	Agency Updates	Surveys and Direct Contact	Internal Plans	Internal Plans
	Usability	Inclusion of Private Sector Projects	No	Yes	No	No
		Forward Pipeline Projects	Yes	Yes	Yes	Yes
		Historical Project Records	No	Yes	Available on request	No
Contextual Data Quality Parameters	Comprehensiveness	Identifiers	Project name, identification number, and short Description	Project name, identification number, and short Description	Name, Scope (New vs. Redevelopment)	Name
		Procuring Agency	Yes	Yes	Owner is the procuring agency	Owner is the procuring agency
		Location	Region, City, Suburb, Coordinates	Provides the region, Suburb and Street address	Provides a broad region	Provides suburb and area
		Project Work Type (as defined by the source)	Includes 16 categories, such as transport, housing, energy, water, health, education, defence, justice, science, waste, emergency management, commercial, community infrastructure, and communications.	Adopts nine broad categories, including residential, civil, commercial, industrial, education, health, energy, sports, and mixed-use projects.	New construction and redevelopment of existing buildings	Depending on the issuing authority, categories range from 4 to 16 and typically include areas like water services, transport, footpaths and roads, parks and community amenities, waste and environmental management, and core engineering services.

Table 10. *Cont.*

Category	Parameters	Data Attributes	Dataset 1	Dataset 2	Dataset 3	Dataset 4
Contextual Data Quality Parameters	Comprehensiveness	Project Lifecycle Phases Reported	Planning, Procurement, Construction	Planning, Procurement, Construction	Procurement	None reported
		Planning and Regulatory Approvals	No	Partial/Inconsistent	No	No
		Procurement	Yes (Direct, Limited, Open, Selective)	Not reported	Not reported	Not reported
		Delivery Method	Design and Build, Early Contractor Involvement, Alliance, Construction, only, Public–Private Partnerships	No	Traditional, Design and Build, Early Contractor Involvement	No
		Supplier information for awarded projects	No	Yes	Preferred Supplier may be identified	No
		Funding	Funding status includes confirmed and to-be-confirmed sources	No	Includes approved projects only	High-level annual capital expenditure
		Preliminary Estimated Value	No	Yes	No	Yes
		Value-bands of projected costs	9	5	7	None
		Value Band Interval Width	Detailed value bands: under NZD 1 M, NZD 1–5 M, NZD 5–25 M, NZD 25–50 M, NZD 50–100 M, NZD 100–250 M, NZD 250–500 M, NZD 500 M–NZD 1 B, over NZD 1 B	Broad value bands: < NZD 1 M, NZD 1–5 M, NZD 5–10 M, NZD 10–25 M, and < NZD 25 M	Moderate value bands: VB1 < NZD 1.5 M, VB2 NZD 1.5–3.5 M, VB3 NZD 3.5–7 M, VB4 NZD 7–14 M, VB5 NZD 14–21 M, VB6 NZD 21–30 M, VB7 > NZD 30 M	No

Table 10. *Cont.*

Category	Parameters	Data Attributes	Dataset 1	Dataset 2	Dataset 3	Dataset 4
Contextual Data Quality Parameters	Comprehensiveness	Schedules Reported	Business Case. Procurement, Construction (Start and End of each phase)	Not specified	Tender, Contract (start date)	No
	Completeness	Number of Columns	29	14	9	NOT SPECIFIED
		Blanks	4	0	0	NOT SPECIFIED
		Partially Populated Columns	10	3	0	NOT SPECIFIED
		Update Frequency	Three months	Update Subscription	Bi-Annual	3 Years
Representational Data Quality Parameters	Ease of Use and Presentation Format	Format	Tableau/Excel	Excel/PDF	PDF	PDF
		Date Formats	Quarters	Month/Year	Month/Year	Year
		Ease of Data Handling/Filtering	Easy to apply data filters and visualisation on MS Excel	Easy to apply data filters and visualisation on MS Excel	Filtering not supported in current format	Filtering not supported in current format
		Interpretability	Easy	Easy	Confusion between overlapping time horizons	Varying terminology across local authorities

5. Discussion

This section interprets the comparative findings through the lens of Wang and Strong's data quality dimensions, offering insights into recurring patterns, structural inconsistencies, and the implications for standardisation

5.1. Data Quality Framework Analysis

Several variances and differences were identified across the four dimensions studied, as shown in Table 10. The quantitative cross-dataset scores in Table 9 further confirm this, with dataset performance ranging from 1.64/5 to 3.99/5. This highlights significant performance gaps and illustrates the extent of data fragmentation. The type of data fields reported across the four studied datasets varied significantly. The construction sector is often challenged with difficulties in acquiring the required project information and data from one source and may be required to navigate through different platforms, synthesise, and triangulate information. The broad definitions of data quality, such as "fitness for use" by [18] or to the satisfaction of customers by [11], allow for numerous interpretations. Given the vastness of the construction sector and the numerous specialities involved, what is deemed fit for use by one provider may not align with others. No dataset achieved 5.0/5.0 across all dimensions, with semantic consistency particularly poor across all datasets, with scores below 3.0/5. This underscores the need for a minimum dataset (MDS) that captures consistent, standardised baseline information to serve the sector's preliminary requirements. As the number of data collectors or providers increases, the variances and differences are expected to increase [23].

5.2. Granularity and Completeness Variations

One of the main differences identified was related to the aims, objectives, scopes, and perspectives of the different data collectors or data sources, as depicted by [23]. As illustrated in the results section, some data providers were private business entities, while others were government entities. This is evidenced by the wide variation in data columns, ranging from 9 to 29 fields, with granularity scores reflecting this scope difference (2.00/5 to 4.00/5) as shown in Table 3. Some take a comprehensive approach and try to capture more breadth, while others focus on reliability over scope breadth. Even within the same data field, the granularity of reporting varied. For example, some datasets adopted a detailed approach in reporting the work types, listing up to 16 distinct work types, while others adopt a much broader classification, offering only a limited description of the project scope, such as "new build" or "redevelopment". This reveals a lack of consensus on the data structure for the required type and number of fields/columns [26].

Completeness scores varied greatly across the datasets, with focused datasets like Dataset 3 scoring as high as 4.25/5 (100% completion) while comprehensive approaches like Dataset 1 scored lower at 3.00/5 (52% completion). Over-ambition and a very broad scope could have contributed to this. Furthermore, the provider's operational background, such as private business versus government entity, could guide the data provider's interest in reporting newly initiated or announced projects to the sector, which may be eager for upcoming opportunities. In contrast, Dataset 1 seemed to pursue full reporting of all lifecycles of projects in their efforts to serve the government, sector, and pipeline providers.

5.3. Terminology Consistency and Semantics

Table 5 reveals that the terminology and semantics scores varied greatly across the datasets, with scores ranging from 1.29/5 to 2.86/5 and with no dataset scoring above 3.00/5, reflecting major standardisation challenges. The use of different semantics, language and reporting formats was also evidenced qualitatively [26,28,30]. For example, each

provider reported projects using different naming systems, with some assigning their project identification numbers, creating difficulties in cross-checking the same project over multiple databases. Furthermore, instances of different formats in reporting the same data values were witnessed [25]. For example, the difference in date formats reveals how the organisational inclinations motivate their reporting behaviours. Monthly formats work well for operational teams tracking project progress, while quarterly reporting suits long-term planning. However, these different approaches create real problems when trying to analyse project timelines across multiple datasets. The objectives of the data providers could also influence this, similar to the example of managers versus traders highlighted by Wang, Kon and Madnick [25].

This variability could also be attributed to the differences in the understanding of the typologies and structures of the collected data [29]. None of the data providers adopted the semantics and structure of the standardised classification/categorisation systems such as StatsNZ [8] or RIBA [7], as suggested by Willemse, Welzen and Mols [26]. For example, in the project work type fields, some datasets take a detailed approach, listing up to 16 distinct work types, while others adopt a much broader classification, offering only a limited and generalised description of the project scope, such as "new build" or "redevelopment". Though the generally known construction work types include categorisations such as residential, non-residential, or civil work categories, some datasets used categorisation terminology related to the client's affiliation, such as "defence" or "justice". While this offers specificity, it may hinder interpretability and cross-dataset comparison. This reflects how organisations design their categories around their own needs and priorities, rather than thinking about how the data might be used across the whole sector.

5.4. Format, Accessibility, and Timeliness Challenges

Accessibility is considered one of the main merits of databases, allowing for more data integration from different sources [26]. The format and accessibility, and timeliness scores revealed the highest variation across the datasets with scores ranging from 5.00/5 to 1.00/5 as shown in Table 6, while timeliness scores varied from 1/5 to 5/5 as shown in Table 7, creating significant operational disparities. Only Dataset 2 exhibited a payment subscription module, as shown in Table 10. This limits the usefulness and availability of the dataset for the wider sector. Data is commonly retrieved through "pull" methods, where users consult the respective data providers' websites and can download the required information. Datasets 2 and 3 offer a "push" service to disseminate updates to subscribed parties. However, the available data formats differed significantly, causing variations across the "representational" data quality dimensions. For example, Dataset 1 offers online Tableau graphs and tables, with the possibility of providing Excel files biannually. Tableau formats may present challenges to some users when downloading and manoeuvring the data. Datasets 3 and 4 offer their data in PDF file formats, making it difficult to manoeuvre or manage compared to Microsoft Excel formats. The lack of compatibility across data provider platforms reduces the satisfaction of the end users, and lowers the compatibility and interoperability across computerised systems, thus reducing the overall quality of the available data [29]. Update frequency variances further exacerbate the format and accessibility issues. Dataset 2 was updated monthly while Dataset 4 was updated every three years, implying that some datasets have much more current information than others. Without standardised data structures, decision-makers have to work with fragmented information, which can result in misaligned planning and duplicated efforts across agencies. The findings thus highlight a pressing need for coordinated national data governance in infrastructure planning.

5.5. A Conceptual Minimum Dataset Approach

Given the identified differences, variances, and challenges within the sampled datasets, the quantitative assessment provides clear evidence that no standardised minimum dataset is being adopted in New Zealand, with overall performance scores ranging from 1.64/5 to 3.99/5 and no dataset achieving excellence across all dimensions. The vast construction domain includes numerous specialities with different "fit-for-use" criteria, contemporary priorities, social and environmental considerations, construction methods, and technical data requirements. Thus, a conceptual minimum dataset approach is proposed in this study to adequately describe construction projects uniformly by providing the basic and preliminary data and information requirements of end users. The scoring patterns showed that even high-performing areas, such as project identification, which scored 5/5 across most datasets, provide a foundation for standardisation, while consistently underperforming areas, such as regulatory approvals, scoring between 1.00/5 and 2.00/5 across all datasets, highlight challenges that require attention. Adopting standard structures, semantics, and formats could enhance the proposed minimum dataset framework and offer a practical alternative towards better integrated and standardised project data [26,27,30].

The conceptual framework is guided by the common and recurring "contextual" data fields within the studied datasets. As an initial step, the "contextual" data within the four studied datasets was broken down into three distinct domains, including "Project Preambles", "Project Characteristics", and "Project Budget and Time", with each domain containing further subcategories as shown in Table 11. A brief definition and the suggested standardised format are proposed to remove confusion and ambiguities.

Table 11. Proposed Minimum Dataset for Construction Activities in New Zealand.

Category	Subcategory 1	Subcategory 2	Definition	Standardisation Recommendation
Preambles	Identity	Project Name Unique ID	Assign a consistent and unique project name that remains the same throughout the project's entire lifecycle	Standardise project names and IDs as assigned by owners
	Location	Region City Suburb	Provide the precise location of the project	Use standard location/ address formats
	Owner	Owner Name	State the identity of the procuring organisation	
Characteristics	Type	Heavy and Civil Non-Residential Residential Services [8]	Identify the specialisation of sector organisations	Use standard classification such as ANZSIC06 by StatsNZ [8]

Table 11. *Cont.*

Category	Subcategory 1	Subcategory 2	Definition	Standardisation Recommendation
Characteristics	Phase	Initial Concept Concept Approvals Design Procurement Construction Completion Maintenance [7]	Reflect the current phase of the project	Use standard phases such as RIBA [7] standards
	Priority	ASAP Scheduled Flexible	Indicate the urgency and priority of the project	
	Funding	Funded Unfunded	Express the funding status of the project	
	Delivery	Traditional Design and Build Direct managed ECI PPP	State the delivery method adopted	Use conventional delivery methods such as MBIE [33]
Budget and Time	Value	Estimated Value Awarded Value	Provide an estimate of the construction-only or contracted value. Expressed in New Zealand Dollars	Use estimated value rather than value bands.
	Time	Start Date Duration	Related to the construction phase but may report other phases. Expressed in year and quarter	Use standard date formats such as Month/Year and use Month for durations.

The Preambles category focuses on the identification information of a project. This includes a unique project name and ID to the project, ensuring consistent identification throughout its lifecycle. The same nomenclature and identification number are proposed to be adopted as presented in the initial building permit/consent application. The Location subcategory captures the precise geographical details of the project site, including the region, city, and suburb, providing a clear understanding of the project's physical location. The Owner subcategory identifies the procuring organisation or the project owner, establishing a key stakeholder in the project.

The Characteristics category classifies projects based on a range of distinct attributes. The "Type" subcategory specifies the construction specialisation or sector required to deliver the project, such as Civil, Commercial, Residential, Environmental, Industrial,

Institutional, or Utilities, allowing for a clear understanding of the project's nature and focus. It is proposed that standardised classification standards, such as ANZSIC06 by Statistics New Zealand, be adopted to overcome any ambiguities or confusion [8]. With Terminology Consistency and Semantics scoring below 3.0/5 across all datasets, adopting these standards could significantly improve the quality and clarity of the datasets. The "Phase" subcategory indicates the current stage or phase of the project, ranging from Initial Concept to Maintenance, providing insight into the project's progress and development. Standard and default construction phases, such as those proposed by RIBA, are suggested to be adopted [7]. The "Priority" subcategory communicates the urgency and priority of the project, whether it requires immediate action (ASAP), follows a predetermined schedule, or allows for flexibility in its execution. The "Funding" subcategory conveys the project's financial status, indicating whether its funding has been approved or remains unfunded. It is an essential indicator of the certainty and likelihood that the project will progress and provides timelines and expectation indications. The "Procurement" subcategory specifies the chosen procurement method for the project, such as open tender, negotiated, early contractor involvement, or collaborative, influencing the project's contracting and partnership strategies. It is proposed that standardised delivery models, such as those suggested by the Ministry of Business, Innovation & Employment in New Zealand, be used [33].

The "Budget and Time" category encapsulates the financial and temporal aspects of the project. This addresses the value reporting inconsistencies identified in the semantic analysis, where different datasets used incompatible approaches, such as detailed bands vs. lump sums vs. no values. The "Value" subcategory includes the estimated value, which represents the construction estimate value based on the initial project designs and documents and the awarded Value of the project. The estimated and awarded value should express monetary values in terms of "currency", such as New Zealand Dollars, clearly understanding the project's financial scale and diminishing the uncertainty created by using generic value bands. The "Time" subcategory captures the start date, typically related to the construction phase. It can also include other phases for better planning coordination. Similarly, the projected duration of the project gives critical information required for decision-making and planning purposes. Using the common standard "month/year" as a standardised date format is proposed, and durations should be reported as "months", addressing the date format inconsistencies that contributed to poor semantic scores across datasets.

6. Conclusions

In response to the challenges posed by the varying quality and availability of project data, the study focused on proposing a conceptual model that could facilitate the identification of an appropriate structure and ontology for the publication of construction project data. The aim is to effectively recognise, describe and present a communicable dataset for the construction industry. The study has developed a conceptual framework that provides theoretical foundations for understanding minimum data requirements across diverse construction market contexts through systematic empirical analysis, using New Zealand as a representative case. While empirically grounded in New Zealand data, the framework's theoretical structure is designed to accommodate diverse regulatory environments, market structures, and stakeholder configurations found across international construction markets.

The study approach was to identify the critical data and information contributing to a well-defined and usable project-level dataset, which can serve as a foundation for effective decision-making, planning, and forecasting within the construction industry. The benefits of a standardised minimum dataset for project data are manifold. Firstly, it can

improve the overall quality of the data by ensuring that the essential project attributes are consistently captured and reported. This can enhance the reliability and usefulness of the data for decision-making, planning, and forecasting purposes. Secondly, the standardised minimum dataset allows for integrating and comparing data from different sources, enabling a more comprehensive and cohesive understanding of the construction industry's performance and trajectory. This can be particularly valuable for government agencies and industry organisations responsible for policy development, funding allocation, and strategic planning. Thirdly, the minimum dataset serves as a starting point for developing more advanced data collection and reporting frameworks within the construction industry and could pave the way for incorporating additional data attributes, integrating emerging technologies, and continuously improving data quality over time.

While this study provided valuable insights, it is essential to acknowledge its limitations. The standardised minimum dataset aims to satisfy the basic and preliminary requirements regarding the quantum of information required to describe a construction project adequately. However, construction professionals or end users may need to seek further granularity levels or data domains, depending on their requirements and the project life cycle. However, higher levels of granularity are often associated with higher costs incurred by the data providers, providing opportunities for the private data providers and businesses. Finally, minimal contextual information on the datasets was explored other than publicly held information; hence, the study was unable to compile information that could give more insights into peculiarities, such as operational limitations/constraints, customer specificities and requirements. Therefore, future studies could benefit from triangulating other methods to provide a more comprehensive understanding of the challenges associated with construction projects' data quality. Future studies should also consider validating and piloting the MDS framework with the stakeholders and industry for further refinements and fine-tuning.

Finally, the proposed MDS may face several challenges in implementation. One significant challenge that may arise is technological incompatibility, such as organisations using advanced systems while others rely on basic PDF reports. Furthermore, securing institutional buy-in may be equally difficult, as government agencies and private companies have different priorities and no clear ownership of the standardisation process. The MDS would work best as a baseline standard that builds on existing systems rather than replacing them entirely. Gradually rolling out the proposed MDS and embedding it within procurement requirements may facilitate its adoption and operationalisation. This would also aid in converging the different levels of granularity adopted currently over the course of time.

Author Contributions: Conceptualization, E.E. and S.W.; Methodology, E.E., J.O.B.R. and T.P.; Formal analysis, E.E., J.O.B.R., T.P. and T.D.M.; Investigation, E.E., J.O.B.R. and T.P.; Resources, E.E.; Data curation, E.E.; Writing—original draft, E.E. and T.D.M.; Writing—review & editing, E.E., J.O.B.R., T.P. and S.W.; Supervision, J.O.B.R. and S.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the CanConstructNZ research programme, funded through the Ministry of Business, Innovation and Employment (MBIE) Endeavour Fund, Contract Number: MAUX2005. The programme is administered by the School of Built Environment at Massey University.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in the study are included in the article; further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Fernz, B. Importance of Open Data to Infrastructure Planning, Procurement, and Delivery. Global Infrastructure Hub. 2022. Available online: https://www.gihub.org/articles/importance-of-open-data-to-infrastructure-planning-procurement-and-delivery/ (accessed on 15 May 2025).
2. Scottish Procurement and Property Directorate. *Construction Pipeline Forecast Tool*; Scottish Government: Scotland, UK, 2024.
3. Moshood, T.D.; Rotimi, J.O.B.; Shahzad, W. Examining Infrastructure Pipelines Information for Their Relevance in Construction Organizations' Strategic Decision-Making. In *Advances in Engineering Management, Innovation, and Sustainability, Proceedings of the 13th International Conference on Engineering, Project, and Production Management, Auckland, New ZealandShow,29 November–1 December 2023*; Springer Nature: Cham, Switzerland, 2023; pp. 173–195.
4. Ministry of Business, Innovation & Employment. National Construction Pipeline Report. 2022. Available online: https://www.mbie.govt.nz/building-and-energy/building/supporting-a-skilled-and-productive-workforce/national-construction-pipeline-report/ (accessed on 26 September 2022).
5. Workforce Information Platform (WIP). About WIP: Infrastructure. Available online: https://wip.org.nz/about-wip?industry_group=Infrastructure (accessed on 1 May 2025).
6. Moshood, T.D.; Rotimi, J.O.; Shahzad, W. Enhancing construction organizations' performance through strategic decision-making: Unveiling the mediating role of quality of information. *Int. J. Organ. Anal.* **2024**. [CrossRef]
7. RIBA. *Plan of Work 2020*; RIBA: London, UK, 2020.
8. Stats NZ. *ANZSIC06 V1.0 (Class) to ANZSIC06 Output CategoriesV1.0.0*; Stats NZ: Auckland, New Zealand, 2007.
9. Yin, R.K. *Case Study Research and Applications: Design and Methods*; Sage Publications: Thousand Oaks, CA, USA, 2017.
10. Miles, M.B.; Huberman, M.; Saldaña, J. *Qualitative Data Analysis: A Methods Sourcebook*, 4th ed.; SAGE: Los Angeles, CA, USA, 2020.
11. Alizamini, F.G.; Pedram, M.M.; Alishahi, M.; Badie, K. Data quality improvement using fuzzy association rules. In Proceedings of the 2010 International Conference on Electronics and Information Engineering, Kyoto, Japan, 1–3 August 2010; pp. V1-468–V1-472.
12. Batini, C.S. *Data and Information Quality: Dimensions, Principles and Techniques*; Springer: Cham, Switzerland, 2018.
13. Rafi, S.; Yu, W.; Akbar, M.A.; Alsanad, A.; Gumaei, A. Multicriteria based decision making of DevOps data quality assessment challenges using fuzzy TOPSIS. *IEEE Access* **2020**, *8*, 46958–46980. [CrossRef]
14. Pipino, L.L.; Lee, Y.W.; Wang, R.Y. Data quality assessment. *Commun. ACM* **2002**, *45*, 211–218. [CrossRef]
15. McGilvray, D. *Executing Data Quality Projects: Ten Steps to Quality Data and Trusted Information (TM)*; Academic Press: Cambridge, CA, USA, 2021.
16. Jesiļevska, S. Data quality dimensions to ensure optimal data quality. *Rom. Econ. J.* **2017**, *20*, 89.
17. Fang, Y.; Zhu, H.; Zeng, Y.; Ma, K.; Wang, Z. Perceptual quality assessment of smartphone photography. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 3677–3686.
18. Wang, R.Y.; Strong, D.M. Beyond accuracy: What data quality means to data consumers. *J. Manag. Inf. Syst.* **1996**, *12*, 5–33. [CrossRef]
19. Vetrò, A.; Canova, L.; Torchiano, M.; Minotas, C.O.; Iemma, R.; Morando, F. Open data quality measurement framework: Definition and application to Open Government Data. *Gov. Inf. Q.* **2016**, *33*, 325–337. [CrossRef]
20. Cichy, C.; Rass, S. An overview of data quality frameworks. *IEEE Access* **2019**, *7*, 24634–24648. [CrossRef]
21. Loshin, D. *Enterprise Knowledge Management: The Data Quality Approach*; Morgan Kaufmann: San Francisco, CA, USA, 2001.
22. Cappiello, C.; Francalanci, C.; Pernici, B. Data quality assessment from the user's perspective. In Proceedings of the 2004 International Workshop on Information Quality in Information Systems, Paris, France, 18 June 2004; pp. 68–73.
23. Watts, S.; Shankaranarayanan, G.; Even, A. Data quality assessment in context: A cognitive perspective. *Decis. Support Syst.* **2009**, *48*, 202–211. [CrossRef]
24. Yousif, O.S.; Zakaria, R.B.; Aminudin, E.; Yahya, K.; Mohd Sam, A.R.; Singaram, L.; Munikanan, V.; Yahya, M.A.; Wahi, N.; Shamsuddin, S.M. Review of big data integration in construction industry digitalization. *Front. Built Environ.* **2021**, *7*, 770496. [CrossRef]
25. Wang, R.Y.; Kon, H.B.; Madnick, S.E. Data quality requirements analysis and modeling. In Proceedings of the IEEE 9th International Conference on Data Engineering, Vienna, Austria, 19–23 April 1993.
26. Willemse, L.; Welzen, P.C.; Mols, J. Standardisation in data-entry across databases: Avoiding Babylonian confusion. *Taxon* **2008**, *57*, 343–345.
27. Woodburn, M.; Paul, D.L.; Addink, W.; Baskauf, S.J.; Blum, S.; Chapman, C.; Grant, S.; Groom, Q.; Jones, J.; Petersen, M.; et al. Unity in variety: Developing a collection description standard by consensus. *Biodivers. Inf. Sci. Stand.* **2020**, *4*, e59233. [CrossRef]
28. Madnick, S.; Zhu, H. Improving data quality through effective use of data semantics. *Data Knowl. Eng.* **2006**, *59*, 460–475. [CrossRef]

29. Siegel, M.; Madnick, S.E. Context interchange: Sharing the meaning of data. *ACM SIGMOD Rec.* **1991**, *20*, 77–78. [CrossRef]
30. Chapman, A.D. *Principles of Data Quality*; GBIF: Copenhagen, Denmark, 2005.
31. Mohammed, S.; Ehrlinger, L.; Harmouch, H.; Naumann, F.; Srivastava, D. Data quality assessment: Challenges and opportunities. *arXiv* **2004**, arXiv:2403.00526.
32. Jabareen, Y. Building a conceptual framework: Philosophy, definitions, and procedure. *Int. J. Qual. Methods* **2009**, *8*, 49–62. [CrossRef]
33. Ministry of Business, Innovation & Employment. *Developing Your Construction Procurement Strategy: Construction Procurement Guidelines*; Ministry of Business, Innovation & Employment: Wellington, New Zealand, 2019.