

Assessing the Impact of Quality-by-Design on Knowledge Graph Construction from Open Government Data

Roxana Martínez*

Universidad Argentina de la Empresa, Buenos Aires, Argentina

Abstract

Nowadays, the use of Open Government Data (OGD) has increased the availability of public information for reuse by both private companies and public entities, while also enhancing transparency and data-driven decision-making to derive value-added insights. However, transforming heterogeneous open datasets into reliable, reusable knowledge remains a significant challenge due to structural inconsistencies, formatting issues, incomplete meta-data, semantic heterogeneity, and limited interoperability. Consequently, knowledge graphs warrant attention, as they offer a suitable representation for integrating heterogeneous sources and facilitating semantic access to organizational and public knowledge; nevertheless, their effectiveness depends on the quality of the underlying data and the semantic models employed. This article investigates the impact of incorporating Quality by Design (QbD) principles during the construction of knowledge graphs from open government data. Unlike other approaches that assess quality only after data publication or graph construction, this work aims to identify the quality requirements necessary during the initial stages of semantic modeling and data transformation. Furthermore, the research adopts an experimental approach comparing knowledge graphs built with and without explicit quality requirements, evaluating aspects such as semantic consistency, completeness, interoperability, query correctness, and knowledge reusability. This study uses open government data as a case study to analyze whether quality requirements incorporated during knowledge construction contribute to generating more reliable and reusable knowledge assets. The contribution of this work lies in an empirical evaluation of the relationship between QbD and knowledge graph quality, as well as a quality-oriented approach to facilitating the transformation of open government data into reusable semantic knowledge.

Keywords

Knowledge graphs, Open government data, Quality by Design, Data quality, Knowledge representation

1. Introduction

Today, Open Government Data (OGD) [1, 2] initiatives have increased the availability of public information to foster transparency, accountability, research, innovation, and data-driven decision-making. Governments are increasingly publishing datasets via open data portals, facilitating access to information that can be reused by citizens, organizations, researchers, and public institutions [3, 4]. However, the availability of data does not guarantee that it can be seamlessly transformed into reliable, reusable information. Accompanying this are numerous issues—such as heterogeneous formats, incomplete metadata, inconsistent representations, semantic ambiguities, and limited interoperability—that can hinder the integration and meaningful reuse of public information [5]. Consequently, a major challenge lies not only in making government data available but also in ensuring that the resulting knowledge can be reliably represented [6], interpreted, queried, integrated, and reused.

In this context, Linked Open Data (LOD) [7, 8] technologies offer mechanisms to address some of these challenges by representing resources and relationships using machine-readable semantic structures [9]. In particular, the Resource Description Framework (RDF) [10] enables data to be represented as interconnected statements, while vocabularies and ontologies provide semantic structures that facilitate interoperability among heterogeneous information sources [11]. Knowledge graphs build upon these principles by representing entities and their relationships as interconnected knowledge

ICAIW 2026: Workshops at the 9th International Conference on Applied Informatics 2026, October 13–17, 2026, Orlando, United States

*Corresponding author.

✉ mariarmartinez@uade.edu.ar (R. Martínez)

ORCID 0000-0001-9318-6411 (R. Martínez)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

structures, thereby enabling semantic queries and facilitating knowledge integration, discovery, and reuse [12]. From a knowledge management perspective, this transformation allows for the organization and representation of information from heterogeneous data sources [8], with the aim of facilitating its interpretation and reuse.

However, constructing a knowledge graph from open government data depends heavily on data quality, decisions made during semantic modeling, and the transformation process itself. Data quality issues that are not identified or addressed during these stages can propagate to the resulting knowledge graph and compromise the reliability of the knowledge. For instance, incomplete information can lead to incomplete knowledge representations; inconsistent values can result in semantically inconsistent entities or relationships; and inadequate vocabulary mappings can limit interoperability with other knowledge sources. Therefore, quality assessment should not be viewed merely as a downstream activity but rather should be integrated into the initial stages of the knowledge construction process.

This perspective aligns with the “Quality by Design” (QbD) principle [13], which holds that quality requirements should be considered from the earliest phases of a process rather than addressed as an afterthought [14]. When applied to knowledge graph construction [15], the QbD approach implies that quality requirements must be considered during data preparation, semantic modeling, transformation, and knowledge representation. This approach can help prevent quality issues from propagating to the resulting knowledge asset, thereby enhancing its reliability [16, 17].

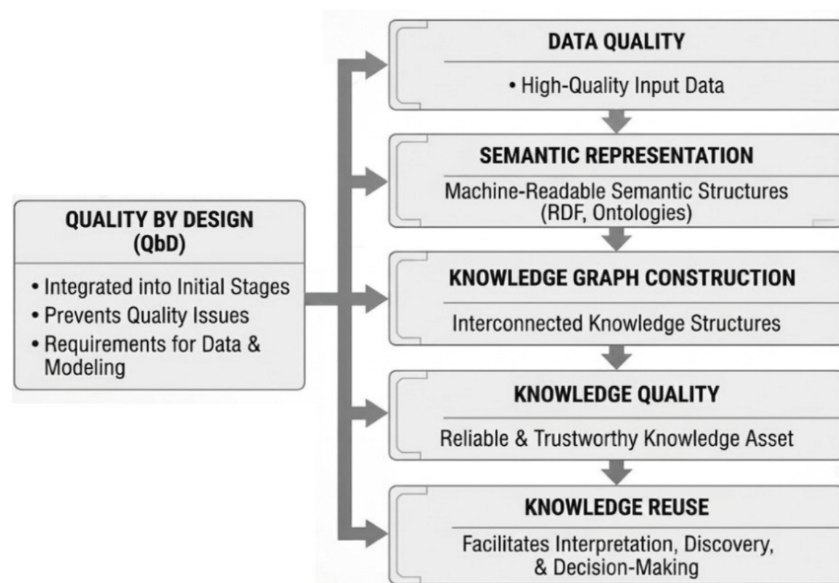


Figure 1: Proposal conceptual framework for Knowledge QbD in Open Government Data pipeline: from raw data integration to reliable knowledge reuse.

Figure 1 illustrates that, to systematically address the challenges of data heterogeneity and semantic ambiguity in OGD, this study demonstrates an interaction between knowledge graph engineering and quality assurance. Following the QbD paradigm, quality requirements are not treated as a post-hoc assessment but are instead natively integrated into the initial stages of data preparation and modeling [18, 19].

This issue becomes evident in the context of government open data, where various agencies generate datasets using distinct structures, vocabularies, metadata practices, and publication procedures. In such environments, constructing a knowledge graph requires more than the mere technical conversion of datasets into RDF format [20, 21]; it entails making decisions regarding entity identification, relationship representation, vocabulary alignment, the management of incomplete or inconsistent information, and the incorporation of semantic constraints into the resulting representation. While previous research has independently addressed Linked Data quality assessment, semantic interoperability, knowledge graph construction [22], and knowledge management, there remains a need to better understand the

relationship between the quality requirements incorporated during graph construction and the quality of the resulting knowledge [23]. This study addresses precisely that research gap. Based on the foregoing, this research addresses the following question:

RQ1: How does the incorporation of “Quality-by-Design” principles during the construction of knowledge graphs affect the semantic quality and reusability of open government data?

To answer this question, the study proposes an experimental comparison between two knowledge graph construction approaches. The first approach represents a conventional construction process in which quality is primarily assessed after the knowledge representation has been generated. The second incorporates explicit quality requirements throughout the construction process, following “Quality-by-Design” principles. Subsequently, the resulting knowledge graphs are evaluated using a multidimensional knowledge quality model comprising five dimensions: semantic consistency, completeness, interoperability, query correctness, and reusability. The research is based on the following hypothesis:

H1: Knowledge graphs constructed by applying “Quality-by-Design” principles achieve higher levels of knowledge quality than those developed without incorporating quality considerations.

Therefore, the main contribution of this article is not the development of yet another knowledge graph construction technique, but rather the empirical evaluation of the effect that a quality-oriented construction approach has on the resulting knowledge. The study brings a quality-centric perspective to knowledge graph construction by linking data quality requirements with knowledge management objectives. Specifically, it: (i) defines a multidimensional model to assess knowledge quality in knowledge graphs derived from open government data; (ii) operationalizes “Quality-by-Design” principles within the knowledge graph construction process; (iii) establishes an experimental comparison between conventional construction approaches and quality-oriented ones; and (iv) analyzes the resulting differences regarding semantic coherence, completeness, interoperability, query correctness, and reusability.

The remainder of the article is organized as follows. Section 2 presents the conceptual foundations of open government data, linked open data, knowledge graphs, knowledge management, and Quality-by-Design. Section 3 reviews research related to linked data quality, knowledge graph construction, and quality-oriented knowledge management, identifying the research gap addressed by this study. Section 4 describes the research methodology, including research questions, hypothesis, experimental design, and evaluation criteria. Section 5 presents the proposed QbD approach for knowledge graph construction. Section 6 describes the case study based on open government data. Section 7 presents experimental evaluation and results. Section 8 discusses the findings and their implications for knowledge management. Section 9 addresses threats to validity, and Section 10 presents the conclusion and future lines of research.

2. Conceptual Foundations

2.1. Knowledge Graphs as Knowledge Assets

A knowledge graph is a structured representation in which entities, attributes, and relationships are explicitly represented and interconnected. Beyond merely representing individual facts, knowledge graphs offer a graph-based structure for organizing heterogeneous information regarding semantically identifiable entities. This feature enables access to knowledge through relationships, rather than exclusively through isolated records or attributes. From a knowledge management perspective, a knowledge graph can therefore be viewed as a knowledge asset rather than simply a data structure. It possesses the capacity to represent domain knowledge in a way that allows it to be queried, interpreted, linked to other sources, and reused by various users. This approach highlights the relationship between the technical generation of graph triples and the characteristics of the knowledge they represent.

The most compelling advantage of knowledge graphs is that they are constructed from heterogeneous sources. A single entity may be represented in different ways across datasets; relationships may carry diverse meanings, and certain properties might be missing or represented using incompatible vocabularies. Consequently, the resulting graph embodies the semantic decisions made during its

construction. Based on the foregoing, this study considers the knowledge graph to be the knowledge representation layer situated between the original open government data and its subsequent use, distinguishing the quality of the original data from the quality of the knowledge representation generated from it. The quality of this layer is evaluated based on the extent to which the represented knowledge is coherent, sufficiently complete, interoperable, effectively consulted, and reusable.

2.2. Knowledge Representation and Semantic Constraints

Knowledge representation involves formalizing concepts, entities, properties, and relationships so that they can be processed by computational systems. In knowledge graph environments, this representation typically relies on semantic technologies such as RDF, RDF Schema, ontologies, controlled vocabularies, and associated constraint languages. Consequently, the selection and application of semantic models entail design decisions that directly influence the resulting knowledge representation. These decisions encompass identifying classes and properties, defining relationships between entities, reusing existing vocabularies, aligning concepts from diverse sources, and representing constraints within the graph. Semantic constraints are particularly significant because syntactic validity does not guarantee semantic validity. A graph may conform to the RDF data model yet contain relationships incompatible with the intended conceptual model. Similarly, an entity might be represented by a syntactically valid identifier but be associated with an inappropriate class or property. This research treats semantic constraints as mechanisms for expressing the expected conditions of the knowledge representation; these conditions may relate to cardinality, data type restrictions, class membership, property usage, mandatory relationships, controlled values, and other domain-specific requirements. Constraint validation provides a mechanism to determine whether the generated graph conforms to the intended representation. In the proposed approach, these constraints are not limited to the final evaluation of the knowledge graph; rather, they are considered during the construction process, allowing quality requirements to influence modeling and transformation decisions before the final knowledge asset is generated. This distinction is fundamental to the experimental design presented in Section 4: the conventional approach and the “Quality-by-Design” approach differ primarily in the timing and way quality requirements are incorporated into the construction process, rather than in the graph technology itself.

2.3. Knowledge Quality

The quality of a knowledge graph cannot be characterized solely by evaluating the quality of its source data. While data quality assessment provides established dimensions for evaluating Linked Data, the construction of a knowledge graph introduces considerations requiring additional factors to be considered, such as semantic issues and usage aspects. It is worth noting that previous research has identified a wide range of quality dimensions for Linked Data, such as completeness, consistency, validity, and interoperability, demonstrating that quality involves multiple analytical perspectives [24]. More recent studies have extended this view specifically to knowledge graphs, proposing quality dimensions and metrics that consider not only the graph’s internal characteristics but also its suitability for specific applications and tasks [25]. Building on these approaches, this research conceptualizes “knowledge quality” as the quality of the knowledge asset resulting from the semantic transformation and representation of heterogeneous data. Rather than replicating an existing quality framework, this work focuses on five dimensions that are particularly relevant to the construction and use of knowledge graphs derived from open government data: $KQ = \{SC, C, I, QC, R\}$ where SC represents semantic consistency, C completeness, I interoperability, QC query correctness, and R reusability. The five dimensions are defined as follows.

2.3.1. Semantic consistency

This refers to the degree to which the entities, classes, properties, and relationships represented in a knowledge graph conform to the semantic model. This dimension focuses on the absence of semantic contradictions and the proper use of conceptual structures defined by selected vocabularies or ontologies.

Consistency is a recognized dimension in Linked Data quality assessment and is particularly relevant in the construction of knowledge graphs, as semantically incompatible representations can arise even when the graph is syntactically valid [24]. Therefore, in this study, semantic consistency determines whether the generated knowledge complies with the semantic constraints established during the modelling process.

2.3.2. Completeness

Completeness refers to the extent to which the knowledge graph contains the information needed to properly represent the relevant entities, properties, and relationships of the domain. Completeness is a well-studied dimension in linked data and knowledge graph quality research [24, 25]. In this study, completeness is not understood simply as having many triples. Instead, it is evaluated according to the requirements defined for the intended knowledge representation. Therefore, missing required properties or relationships, as well as incomplete descriptions, may reduce the knowledge graph's ability to represent relevant domain knowledge.

2.3.3. Interoperability

Interoperability represents the knowledge graph's ability to integrate and exchange knowledge with other datasets, semantic models, vocabularies, and information systems. Interoperability is crucial for linked data, as the value of interconnected knowledge depends, in part, on the ability to establish meaningful links between independently generated resources [24]. In the proposed model, interoperability refers to the extent to which the semantic representation can connect with external knowledge resources using compatible vocabularies, identifiers, and semantic relationships. In this regard, this dimension extends the evaluation beyond the graph's internal correctness and considers its capacity to participate in a broader knowledge ecosystem.

2.3.4. Query Correctness

Query correctness refers to the extent to which queries executed on the knowledge graph return results that match the intended information requirements and the semantics represented in the graph. This dimension provides a task-oriented perspective for assessing knowledge quality. Recent research suggests that evaluating a knowledge graph only at the data level may not be sufficient, because its quality also depends on its ability to support practical tasks [26, 27]. Therefore, this study uses query execution to assess whether the semantic representation of the knowledge graph can provide correct answers to predefined information requirements. Query correctness is not considered a conventional Linked Data quality dimension adopted without changes. Instead, it is defined as a task-oriented dimension that connects the structural and semantic characteristics of the knowledge graph with its practical use.

2.3.5. Reusability

Reusability describes how effectively the knowledge represented in a knowledge graph can be reused by different users, applications, analytical processes, or integration scenarios. This concept is related to quality principles such as accessibility, interoperability, understandability, and the reuse of Linked Data [24]. However, in knowledge graph environments, reusability also depends on whether the resulting knowledge can support the intended uses. Recent studies highlight the importance of evaluating the usefulness of knowledge for downstream applications, rather than focusing only on the internal characteristics of the graph [26, 28]. Therefore, reusability can be considered a higher-level outcome that is influenced by the other dimensions of knowledge quality.

3. Research Gap and Contribution

Recent studies have proposed various approaches for constructing knowledge graphs from heterogeneous data, analyzing stages such as data preparation, ontology modeling, integration, graph generation, and validation. Hofer et al. [29] highlighted the need to better integrate these activities and quality assurance into the construction process. Similarly, Meckler [30] proposed a requirements-based construction procedure, using competency questions to guide knowledge graph development. While these studies demonstrate the importance of considering requirements during construction, they do not experimentally compare conventional construction strategies with quality-oriented ones.

Other approaches have focused on the automated, domain-specific construction of knowledge graphs. Zhang and Soh [31] demonstrated that the organization of the construction process can influence the quality of the generated knowledge, while Li et al. [32] incorporated iterative quality control into the construction of a medical knowledge graph. In another study by the same authors [33], the combination of knowledge graph construction and quality assessment in the cybersecurity domain is presented, and Feng et al. [34] used competency questions to guide the construction of ontology-based knowledge graphs. Although these approaches integrate quality-related activities into the construction process, they do not evaluate their effect on multiple dimensions of knowledge quality through a controlled comparison.

Based on what was previously mentioned, a research gap persists regarding the empirical effect of incorporating explicit quality requirements throughout the knowledge graph construction process. There is little evidence comparing knowledge graphs built with and without quality-oriented requirements under similar conditions or evaluating their impact across multiple quality dimensions.

Therefore, this study compares two Knowledge Graph construction strategies. The first uses a conventional construction process, while the second includes quality requirements during data preparation, semantic modelling, transformation, and validation. The resulting Knowledge Graphs are then evaluated using the five Knowledge Quality dimensions defined in Section 2.3: semantic consistency, completeness, interoperability, query correctness, and reusability.

This study focuses on knowledge graph construction, knowledge quality, and the QbD approach utilizing heterogeneous open government data. Rather than simply proposing a new quality metric or construction method, this research evaluates whether incorporating quality requirements during construction improves the quality of the resulting knowledge graph.

4. Research Methodology

4.1. Research Design

This research employs an experimental design to analyze whether incorporating QbD principles during knowledge graph construction contributes to improving the quality of the resulting knowledge.

To this end, two strategies are compared: a) a Conventional Construction Approach (CCA), in which data are transformed into a knowledge graph via a standard modeling and transformation process, with quality assessment performed after the graph is built; and b) a Quality by Design Approach (QbD-A), in which specific quality requirements are incorporated throughout the construction process, including data preparation, semantic modeling, data transformation, and validation.

Both approaches utilize the same government open data sources, domain, target semantic representation, and evaluation procedure. Therefore, the primary difference lies in the incorporation of quality requirements during the construction process. Figure 2 illustrates the design presented in this work. Both construction strategies originate from the same Open Government Data (OGD), referring to the raw datasets obtained from selected open data sources. The first branch follows the Conventional Construction Approach (CCA), wherein source data is transformed into a Knowledge Graph via a conventional construction process, yielding the KG-CCA (Knowledge Graph based on the Conventional Construction Approach). The second branch follows the proposed Quality by Design Approach (QbD-A), incorporating quality requirements throughout the construction process to produce the KG-QbD

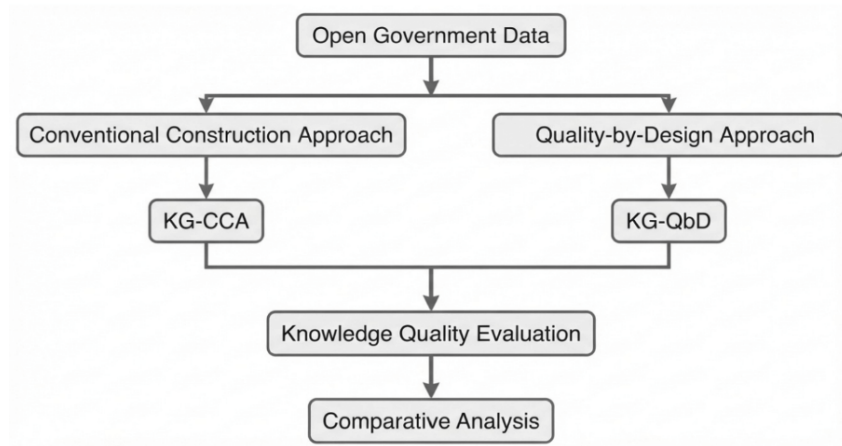


Figure 2: Experimental design for comparing Conventional and QbD Knowledge Graph construction approaches.

(Knowledge Graph based on Quality by Design). Both Knowledge Graphs represent the same domain and are subsequently evaluated using the same knowledge quality model and evaluation procedure, enabling a controlled comparison of the construction strategy’s impact on the resulting knowledge quality.

4.2. Research question and hypothesis

As introduced in Section 1, the experiment is guided by RQ1 and examined through H1, which are restated here only in relation to the five knowledge quality dimensions compared between KG-QbD and KG-CCA.

4.3. Experimental Variables

The independent variable is the Knowledge Graph construction strategy, with two levels:

- *CCA: Conventional Construction Approach.*
- *QbD-A: Quality-by-Design Approach.*

The dependent variable is Knowledge Quality (KQ), operationalized through the five defined dimensions: $KQ = \{SC, C, I, QC, R\}$, where SC, C, I, QC, and R are the five dimensions defined in Section 2.3.

On the other hand, Table 1 shows the different variables for this context.

Table 1

Table shows variables and operationalization.

Variable type	Variable	Operationalization
Independent	Construction strategy	CCA / QbD-A
Construction strategy	Knowledge Quality	SC, C, I, QC, R

The source data, domain, semantic representation, and evaluation procedure remain consistent across the CCA and QbD-A conditions.

4.4. Experimental Setting

The experimental units are KG-CCA and KG-QbD, generated from the same OGD source. This proposal uses the “Subte: GTFS” dataset [35, 36] published via Buenos Aires Data. This dataset provides

information on the Buenos Aires subway system in accordance with the GTFS (General Transit Feed Specification), including agencies, routes, trips, stop times, stops, service calendars, fares, and transfers. To ensure a proper comparison, the source dataset, data scope, domain, target entities and relationships, semantic representation, evaluation queries, quality dimensions, and evaluation procedures are kept constant for both construction conditions. The difference between the two conditions lies in how quality requirements are incorporated into the construction process. As previously mentioned, under the CCA condition, data are transformed following the conventional construction process and evaluated subsequently. Under the QbD-A condition, explicit quality requirements are defined prior to construction and incorporated into data preparation, semantic modeling, transformation, and validation activities. This setup allows for examining whether the construction strategy influences the quality of the resulting knowledge, while minimizing the influence of other experimental factors.

4.5. Experimental Procedure

The experimental procedure comprised five main stages: (1) preparation of the OGD source data; (2) construction of the knowledge graph using the Conventional Construction Approach (KG-CCA); (3) construction of the knowledge graph following the Quality-by-Design approach (KG-QbD), incorporating explicit quality requirements during the construction process; (4) evaluation of both knowledge graphs using the five Knowledge Quality dimensions and the defined competency questions; and (5) comparative analysis of the resulting quality scores. Both approaches used the same source data, target semantic representation, evaluation dimensions, validation rules, and correction criteria, with the main difference being when and how quality requirements were incorporated into the construction process.

5. Quality-by-Design Approach for Knowledge Graph Construction

This section presents the proposed Quality-by-Design (QbD-A) approach for constructing knowledge graphs from open government data (OGD). The approach is designed to establish a traceable relationship between construction activities and the knowledge quality dimensions defined in the previous section. The proposed QbD-A approach integrates quality requirements into key stages of knowledge graph construction: data preparation, semantic modeling, transformation, and validation. Quality requirements are defined prior to construction and subsequently used to guide decisions made at each stage. During the data preparation stage, source data is examined for structural inconsistencies, missing values, duplicate identifiers, and invalid references.

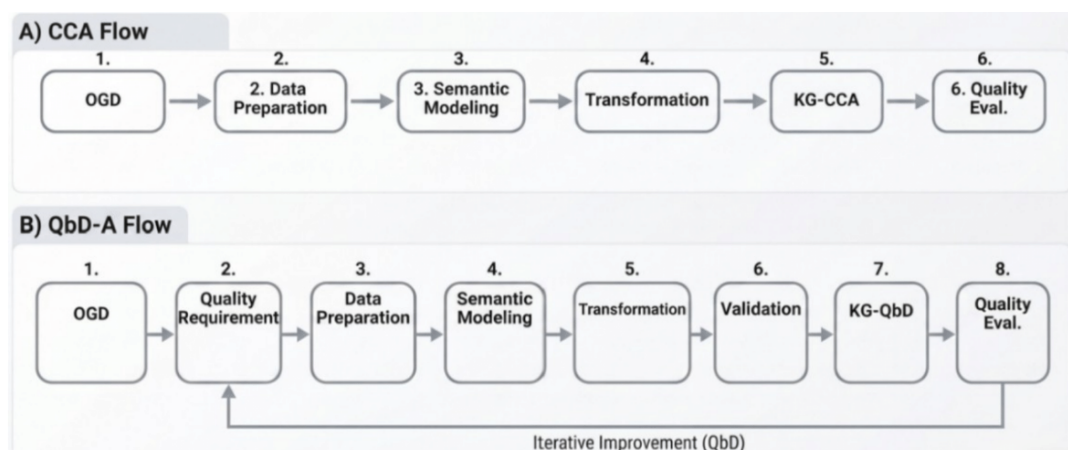


Figure 3: The resulting processes.

During semantic modeling, quality requirements are incorporated into the definition of entities, properties, and relationships. During transformation, these requirements guide identifier mapping,

property generation, data type assignment, and relationship creation. Finally, the validation stage verifies whether the resulting knowledge graph meets the quality constraints defined prior to the final assessment of knowledge quality. It is shown in Figure 3.

For the experimental case study, examples of quality requirements include: valid and unique identifiers, complete mandatory attributes, valid references between related entities, compatible semantic relationships, consistent use of identifiers and vocabularies, and the ability to answer predefined competency questions.

6. Experimental Case Study

This work focuses on the “Subte: GTFS” dataset [35, 36], which describes the Buenos Aires subway system using the General Transit Feed Specification (GTFS); it provides structured information regarding agencies, routes, trips, stops, stop times, service calendars, fares, and transfers. It was selected because it constitutes a real-world source of open government data comprising multiple related entities and explicit links, elements ideally suited for semantic transformation into a knowledge graph. The dataset serves as a common input for both experimental conditions. Consequently, neither CCA nor QbD-A modifies the source data; rather, the experimental difference lies in how quality requirements are addressed during the transformation of that data into knowledge. The main entities considered in the case study are Agency, Route, Trip, Stop, Stop Time, and Service, along with the relationships necessary to represent their semantic associations. This scope provides a sufficiently connected structure to evaluate semantic coherence, completeness, interoperability, query correctness, and reusability.

Two knowledge graphs are constructed from the same GTFS source: KG-CCA and KG-QbD. In the CCA condition, source data are transformed into RDF following the target semantic model, while quality checks are applied primarily after the graph has been built. In the QbD-A condition, quality requirements are incorporated into the data preparation, semantic modeling, transformation, and validation phases, prior to the final evaluation of knowledge quality. For KG-CCA, the construction process focuses on representing source entities and their relationships according to the selected semantic model. For KG-QbD, this transformation is complemented by explicit quality requirements. The requirements are translated into validation rules that can be automatically verified during the construction process. For example, a Trip must reference a valid Route and Service, while a StopTime must reference a valid Trip and Stop.

Furthermore, mandatory properties must be present, and identifiers must remain consistent throughout the transformation. The validation results provide evidence as to whether the construction process meets predefined requirements and identify potential quality issues prior to the final assessment of knowledge quality. The quality requirements are linked to the five dimensions of the Knowledge Quality model.

Table 2

Table shows quality requirements.

Dimension	Main quality requirement	Evaluation focus
SC	Valid classes, properties, identifiers, and relationships	Compliance with the semantic model
C	Presence of mandatory entities, attributes, and relationships	Coverage of required knowledge
I	Consistent identifiers and compatible vocabularies	Ability to connect with external resources
QC	Valid semantic relationships and competency questions	Correct answers to predefined queries
R	Consistent and sufficiently complete representation	Support for alternative uses and queries

Table 2 summarizes the relationship between the construction requirements and the evaluation dimensions.

This analysis considers competency questions to represent the intended information requirements. Examples include identifying the routes served by a stop, determining the stops associated with a trip, retrieving the service linked to a trip, and identifying the sequence of stops for a specific trip. These questions determine whether semantic representation provides the information required by users and applications. Thus, query correctness complements structural validation by evaluating the knowledge graph from a usage-oriented perspective.

7. Experimental Evaluation and Results

7.1. Experimental Dataset and Evaluation Setup

The experimental evaluation was conducted using the Subte: GTFS dataset published by the Government of the City of Buenos Aires [35, 36]. The dataset describes the Buenos Aires subway transportation network and includes information about agencies, routes, trips, stops, stop times, service calendars, pathways, and transfers. The supplied GTFS feed contains the core files required to represent the transit service, including `agency.txt`, `routes.txt`, `trips.txt`, `stops.txt`, `stop_times.txt`, and `calendar.txt`. The GTFS specification defines the identifiers and foreign-key relationships connecting these files, such as the relationship between trips and routes, trips and services, and stop times and trips and stops [GTFS Reference]. The version of the feed used in this experiment identifies the publisher as GP-SBASE and reports feed version 2.0, with a service period from May 1, 2014 to December 31, 2019. The supplied dataset contains 1 agency, 8 routes, 48 trips, 708 stops, 714 stop-time records, 3 service definitions, 967 pathway records, and 112 transfer records.

Two knowledge graphs were considered: KG-CCA, constructed using the Conventional Construction Approach, and KG-QbD, constructed following the Quality-by-Design approach. Both approaches use the same source dataset, domain, semantic representation, entity scope, and final evaluation procedure. The distinction between them concerns the point at which quality requirements are incorporated into the construction process. For KG-CCA, the source data are transformed into a semantic representation without explicitly incorporating the quality requirements as construction constraints. Quality is subsequently assessed on the resulting representation. For KG-QbD, the same requirements are incorporated during data preparation, semantic mapping, relationship construction, and validation. In particular, identifier uniqueness, mandatory attributes, foreign-key consistency, and semantic relationships are explicitly checked before and during graph construction.

To avoid confounding the comparison with different validation opportunities, both approaches were evaluated using the same final validation rules and the same correction criteria. Therefore, the experiment compares the effect of incorporating quality requirements during construction rather than comparing a validated graph with an unvalidated graph.

The experimental scope focuses on the following entities and relationships: Agency, Route, Trip, Stop, Stop Time, Service, Pathway, and Transfer. The evaluated relationships include Route–Agency, Trip–Route, Trip–Service, StopTime–Trip, StopTime–Stop, Pathway–Stop, and Transfer–Stop relationships. These relationships were selected because they represent the principal identifier-based connections among the GTFS entities included in the experimental dataset. The evaluation follows the five quality dimensions introduced in Section 4: Semantic Consistency (SC), Completeness (C), Interoperability (I), Query Correctness (QC), and Reusability (R).

7.2. Operationalization of Quality Metrics

To make the evaluation reproducible, each quality dimension was operationalized as a normalized score between 0 and 1.

Semantic Consistency (SC) measures whether the relationships represented in the knowledge graph

are semantically valid with respect to the source-data constraints and the selected semantic representation:

$$SC = \frac{N_{\text{valid semantic relationships}}}{N_{\text{evaluated semantic relationships}}} \quad (1)$$

The evaluation considered the relationships between routes and agencies, trips and routes, trips and services, stop times and trips, stop times and stops, pathways and stops, and transfers and stops. The GTFS specification defines these relationships through foreign identifiers and corresponding primary identifiers [35, 36].

Completeness (C) measures the proportion of required information successfully represented in the knowledge graph:

$$C = \frac{N_{\text{present required elements}}}{N_{\text{required elements}}} \quad (2)$$

The analysis considers the required identifiers and mandatory relationship attributes defined by the GTFS specification for the files included in the experimental scope.

Interoperability (I) measures the proportion of the evaluated source concepts and relationships that can be represented consistently using the selected semantic model while preserving their identifiers and intended relationships:

$$I = \frac{N_{\text{successfully mapped concepts} \wedge \text{relationships}}}{N_{\text{evaluated concepts} \wedge \text{relationships}}} \quad (3)$$

Query Correctness (QC) evaluates whether the knowledge graph can correctly answer the competency questions defined for the experiment:

$$QC = \frac{N_{\text{correctly answered competency questions}}}{N_{\text{evaluated competency questions}}} \quad (4)$$

The following competency questions were evaluated:

1. Which routes serve a given stop?
2. Which stops are associated with a given trip?
3. Which service calendar is associated with a given trip?
4. What is the ordered sequence of stops for a given trip?

Finally, Reusability (R) measures whether the resulting representation can be used in independent information-retrieval scenarios without requiring reconstruction of the original tabular relationships:

$$R = \frac{N_{\text{successful reuse scenarios}}}{N_{\text{evaluated reuse scenarios}}} \quad (5)$$

The reuse scenarios correspond to the competency questions and to reverse navigation across the main relationships of the graph, such as navigating from a stop to its trips and routes or from a trip to its associated stops and service.

All metrics are normalized to [0,1], where 1 indicates complete compliance. Scores were calculated as the proportion of satisfied criteria using the same dataset, semantic representation, validation conditions, and evaluation criteria for both approaches. So, the reported scores reflect compliance with the predefined requirements within this case study and should not be interpreted as universal measures of knowledge graph quality.

7.3. Dataset Validation Results

Before constructing the knowledge graphs, the source dataset was inspected to establish the baseline conditions under which the two construction approaches were compared. The validation focused on identifier uniqueness, mandatory references, foreign-key consistency, and the ordering of stop sequences. The analysis identified no duplicated identifiers for the principal entities considered in the experimental scope. The primary identifiers of agencies, routes, trips, stops, services, and pathways were unique within their corresponding files. For `stop_times.txt`, the composite key defined by GTFS, consisting of `trip_id` and `stop_sequence`, was also unique for the evaluated records. The principal foreign-key relationships were then evaluated according to the relationships defined by the GTFS specification. The validation covered 3,690 relationship assertions: 8 Route–Agency references, 48 Trip–Route references, 48 Trip–Service references, 714 StopTime–Trip references, 714 StopTime–Stop references, 1,934 Pathway–Stop endpoint references, and 224 Transfer–Stop endpoint references. All evaluated references were found to be valid; no orphan references were detected.

In addition, the `stop_sequence` attribute was evaluated for the 48 trips represented in the feed. The sequence values were checked to verify that the ordering of stops within each trip was valid. No duplicated or non-monotonic stop sequences were identified. These validation results establish that the supplied GTFS feed presents a high level of structural consistency for the relationships included in the experimental scope. Consequently, no source-level structural defects were available to distinguish the two construction approaches through the final quality scores. This characteristic of the dataset is explicitly considered when interpreting the comparative results in the following sections. The validation therefore serves two purposes in the experiment. First, it establishes the empirical baseline against which KG-CCA and KG-QbD are evaluated. Second, it prevents introducing artificial defects solely to obtain a measurable difference between the approaches, preserving the validity and reproducibility of the comparison.

7.4. Experimental Results

Table 3 shows that the two conditions differ primarily in the timing of the introduction of quality requirements, rather than in whether or not they are introduced. The results show that both approaches produced representations satisfying all the evaluated quality criteria for this dataset. In particular, the semantic relationship checks did not identify invalid references, the required information was represented, the evaluated concepts and relationships could be mapped consistently, all four competency questions were correctly answered, and the evaluated reuse scenarios were supported.

Table 3
Quality evaluation of KG-CCA and KG-QbD.

Quality dimension	KG-CCA	KG-QbD	Absolute difference
SC	1.00	1.00	0.00
C	1.00	1.00	0.00
I	1.00	1.00	0.00
QC	1.00	1.00	0.00
R	1.00	1.00	0.00

The results show that both KG-CCA and KG-QbD achieved complete compliance across the five evaluation dimensions, with all four competency questions correctly answered. Thus, both approaches reached equivalent final quality under the selected conditions. However, these results do not establish equivalence in development effort, error-detection timing, or maintenance cost. For structurally consistent source data, the main difference between the approaches may therefore lie not in final quality,

but in when quality requirements are incorporated during the construction lifecycle. Further empirical evaluation is needed to assess this potential distinction, as process-level indicators were not collected.

7.5. Construction-Process Analysis

Although the final quality scores were equivalent, the analysis provides evidence of the different roles played by quality requirements in the two construction strategies.

In KG-CCA, quality requirements are primarily treated as evaluation criteria applied to the resulting representation. In KG-QbD, the same requirements are treated as construction constraints and validation conditions throughout the transformation process.

The result illustrates that QbD may provide process-oriented benefits even when no difference in final quality scores is observed. By incorporating quality requirements as explicit construction constraints, QbD enables earlier detection of potential violations. The experiment therefore demonstrates the operationalization of the proposed quality dimensions on a real open-government GTFS dataset.

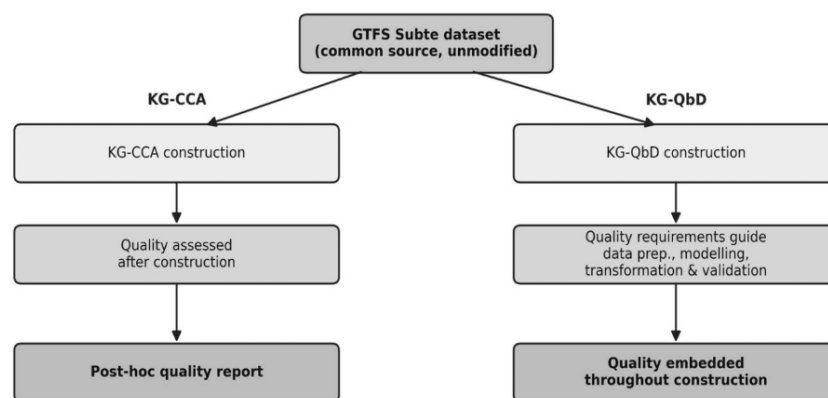


Figure 4: Point at which quality requirements are introduced in KG-CCA vs. KG-QbD.

In contrast, because the evaluated source data contain no detected violations in the selected structural relationships, this single case does not provide sufficient evidence to attribute an improvement in final knowledge quality exclusively to QbD. Further experiments using datasets with heterogeneous quality levels or controlled, transparently documented construction defects are required to measure whether earlier quality intervention produces measurable improvements in final graph quality.

7.6. Interpretation of the Results

The experimental results show that the proposed five-dimensional Knowledge Quality model can be operationalized on a real open-government GTFS dataset through structural validation, semantic relationship checks, competency questions, and reuse-oriented scenarios.

Figure 5 illustrates the traceability established in the QbD-A approach between quality requirements, construction activities, validation rules, and the resulting Knowledge Quality dimensions. This traceability represents a process-oriented characteristic of QbD-A rather than evidence of higher final quality, which was not observed in the present experimental case.

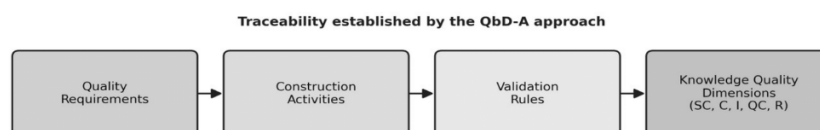


Figure 5: Traceability model linking quality requirements, construction activities, validation rules, and Knowledge Quality dimensions in QbD-A.

However, KG-CCA and KG-QbD achieved identical final quality scores, consistent with the high structural consistency of the source data. Thus, incorporating quality requirements earlier did not produce a measurable difference in this case. The results support the operationalization of the proposed framework but do not confirm H1 for this dataset. Further experiments with heterogeneous quality conditions and documented transformation risks are needed to assess whether QbD-A improves final knowledge quality.

8. Discussion

The case study shows that Quality-by-Design principles can be incorporated into knowledge graph construction by explicitly linking quality requirements with data preparation, semantic modeling, transformation, validation, and assessment activities. In the evaluated case, however, KG-CCA and KG-QbD achieved identical scores across all five Knowledge Quality dimensions. This result is explained by the high structural consistency of the selected GTFS dataset and the use of the same final validation rules and correction criteria for both approaches.

These findings indicate that incorporating quality requirements during construction did not produce a measurable improvement in final knowledge quality for this dataset. Nevertheless, the QbD approach provides explicit traceability between quality requirements, construction decisions, validation rules, and the resulting quality dimensions. This process-oriented characteristic may become particularly relevant when datasets contain structural inconsistencies, incomplete relationships, heterogeneous formats, or other quality risks that can affect the resulting knowledge graph. These findings are consistent with previous research emphasizing the importance of requirements, construction processes, and quality assessment in knowledge graph engineering [29] – [34]. The results also reinforce the distinction between the quality of the original data and the quality of its semantic representation. Although the source dataset showed high structural consistency, the evaluation framework enabled the assessment of how such properties were preserved in the resulting knowledge graphs. Therefore, the contribution of this study is not evidence that QbD always produces higher-quality knowledge graphs, but rather an empirical operationalization of the relationship between quality requirements, construction activities, validation rules, and knowledge quality. Given the single-dataset setting, further experiments with heterogeneous datasets and controlled quality variations are required to determine whether these process-oriented benefits translate into measurable improvements in final knowledge quality.

9. Threats to Validity

It is worth noting that this study has certain limitations. First, the experimental case study relies on a single open government data domain, the Buenos Aires subway GTFS dataset [35, 36], meaning the findings cannot be generalized to other datasets or government domains without further experimentation. Second, the evaluation depends on the knowledge quality dimensions and requirements defined in this study; employing alternative quality models or domain-specific requirements could yield different results. Third, the comparison focuses on the construction strategy while keeping the source data, semantic representation, and evaluation procedure constant; while this strengthens internal validity, other factors inherent to knowledge graph engineering could influence the resulting quality. Finally, reusability depends on the users and intended applications; therefore, it must be interpreted in the context of the study's use cases and competency questions, rather than as an absolute property of the graph. Conducting future experiments that incorporate additional datasets, domains, quality models, and independent evaluators would strengthen the external validity of the proposed approach.

10. Conclusions and Future Work

This study investigated the incorporation of Quality-by-Design principles into the construction of knowledge graphs from open government data. Using a five-dimensional Knowledge Quality model

comprising semantic consistency, completeness, interoperability, query correctness, and reusability, the study compared KG-CCA and KG-QbD generated from the same Subte [35, 36]. The experimental results showed that both construction approaches achieved equivalent final quality scores for the selected dataset. The structural validation identified no duplicated primary identifiers, no invalid foreign-key references, and no violations in the evaluated stop sequences. Consequently, no measurable difference in final knowledge quality was observed between the two approaches. The main contribution of the study is therefore the operationalization of a traceable relationship among quality requirements, construction activities, validation rules, and knowledge quality dimensions. The results also highlight an important condition for evaluating Quality-by-Design: when the source data are already highly consistent, incorporating quality requirements earlier in the construction process may not result in higher final quality scores. Its potential benefits should therefore be evaluated across datasets presenting different quality conditions and transformation risks. The findings should be interpreted as evidence from a single case study rather than as confirmation that QbD-A systematically produces higher-quality knowledge graphs. In particular, the present experiment does not confirm H1 for the selected dataset. Additional empirical studies are required to determine under which data-quality conditions and construction scenarios QbD-A produces measurable improvements. Future work will also evaluate the approach across additional domains and more heterogeneous data conditions.

Declaration on Generative AI

During the preparation of this work, the author used Google Translate (Google) for text translation; ChatGPT (OpenAI) for literature review support, drafting and refinement of selected sections, paraphrasing and rewording, improvement of writing style, and grammar and spelling checking; and Gemini (Google) for the generation of selected illustrative images included in the work. All AI-assisted content and images were reviewed, edited, and validated by the author. The authors retain full responsibility for the scientific content of the work, including the research design, data analysis, interpretation of results, and conclusions.

References

- [1] Open Government Partnership, Open data, 2026. URL: <https://www.opengovpartnership.org/open-data/>.
- [2] Organisation for Economic Co-operation and Development (OECD), Open government data, 2026. URL: https://www.oecd.org/en/publications/open-government-data_5k46bj4f03s7-en.html.
- [3] Open Government Partnership, Data dashboard, 2026. URL: <https://www.opengovpartnership.org/data-dashboard/>.
- [4] Open Definition, The open definition, 2026. URL: <http://opendefinition.org/>.
- [5] L. Jing, Q. Tianshu, H. Qing, Open government data and green economic growth, *Economic Analysis and Policy* 85 (2025) 2207–2218.
- [6] Z. Hu, P. Sun, Does open government data improve public service quality? a quasi-natural experiment based on the launch of open government data platforms in china, *Public Performance & Management Review* 49 (2026) 510–536.
- [7] S. von Mering, S. Leachman, J. Santos, H. M. Meudt, Wikidata for botanists: benefits of collaborating and sharing linked open data, *Annals of botany* 136 (2025) 491–511.
- [8] L. John, K. E. Farfar, S. Auer, O. Karras, Scimantify-a hybrid approach for the evolving semantification of scientific knowledge, in: *International Conference on Web Engineering*, Springer, 2025, pp. 217–225.
- [9] V. de Boer, S. B. A. Shoilee, Linked open data for cultural heritage, in: *The Palgrave Encyclopedia of Cultural Heritage and Conflict*, Springer, 2025, pp. 1–7.
- [10] Q. Li, S. Kim, L. Zaslavsky, T. Cheng, B. Yu, E. E. Bolton, A resource description framework (rdf)

- model of named entity co-occurrences in biomedical literature and its integration with pubchemrdf, *Journal of cheminformatics* 17 (2025) 79.
- [11] S. Hayat, N. F. Warraich, M. N. Akhtar, S. Batool, F. S. Hashmi, A systematic review of linked data and rdf adoption in libraries: Benefits, challenges and best practices, *ASSAJ* 4 (2025) 2180–2192.
- [12] A. S. Lippolis, G. Lodi, A. G. Nuzzolese, The water health open knowledge graph, *Scientific Data* 12 (2025) 274.
- [13] R. Duan, Z. Jiang, C. Ke, H. Cui, B. Sun, A knowledge graph-based intelligent inference method for sustainable design requirements across the product lifecycle, *Journal of Engineering Design* (2026) 1–37.
- [14] Y. Huang, G. Wang, R. Wang, T. Peng, H. Li, Y. Yan, A case-based knowledge graph with reinforcement learning for intelligent design approach of complex product, *Journal of Engineering Design* 36 (2025) 1451–1478.
- [15] G. Chen, X. Xiahou, J. Li, L. Chen, Y. Zou, S. Zhou, Knowledge graph-driven question answering for prefabricated building quality management through natural language processing and transfer learning, *Engineering, Construction and Architectural Management* 33 (2026) 6887–6913.
- [16] M. A. Parvez, Analytical quality by design (aqbd) a comprehensive review of modern analytical method development, *World Journal of Pharmaceutical Sciences* (2026).
- [17] N. N. Pavan Kumar, K. Chandrasekhar, Quality by design (qbd) approach to se-hplc method development for cetuximab monoclonal antibodies, *Preparative Biochemistry & Biotechnology* 56 (2026) 1255–1267.
- [18] A. K. Singh, Y. Vashitha, A. Jain, G. Das Gupta, S. K. Verma, A review on advancement in analytical quality by design (aqbd), *Current Analytical Chemistry* 21 (2025) 901–922.
- [19] C. Guerra-García, A. Nikiforova, S. Jiménez, H. G. Perez-Gonzalez, M. Ramírez-Torres, L. Ontañón-García, Iso/iec 25012-based methodology for managing data quality requirements in the development of information systems: Towards data quality by design, *Data & Knowledge Engineering* 145 (2023) 102152.
- [20] E. Rajabi, R. Midha, J. F. de Souza, Constructing a knowledge graph for open government data: the case of nova scotia disease datasets, *Journal of biomedical semantics* 14 (2023) 4.
- [21] A. Quarati, R. Albertoni, Linked open government data: Still a viable option for sharing and integrating public data?, *Future Internet* 16 (2024) 99.
- [22] T. Lin, Q. Chen, G. Cheng, A. Soylu, B. Ell, R. Zhao, Q. Shi, X. Wang, Y. Gu, E. Kharlamov, Acordar: a test collection for ad hoc content-based (rdf) dataset retrieval, in: *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2022, pp. 2981–2991.
- [23] B. Xue, L. Zou, Knowledge graph quality management: a comprehensive survey, *IEEE transactions on knowledge and data engineering* 35 (2022) 4969–4988.
- [24] A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, S. Auer, Quality assessment for linked data: A survey: A systematic literature review and conceptual framework, *Semantic web* 7 (2015) 63–93.
- [25] E. Huaman, Steps to knowledge graphs quality assessment, *arXiv preprint arXiv:2208.07779* (2022).
- [26] X. Li, Y. Zhou, K. Zhu, F. Li, H. Yang, Y. Ren, Knowledge graph quality evaluation under incomplete information, in: *International Conference on Advanced Data Mining and Applications*, Springer, 2025, pp. 145–159.
- [27] Y. Zhang, G. Xiao, A novel customizing knowledge graph evaluation method for incorporating user needs, *Scientific Reports* 14 (2024) 9594.
- [28] S. Seo, H. Cheon, H. Kim, D. Hyun, Structural quality metrics to evaluate knowledge graphs, *arXiv preprint arXiv:2211.10011* (2022).
- [29] M. Hofer, D. Obraczka, A. Saeedi, H. Köpcke, E. Rahm, Construction of knowledge graphs: Current state and challenges, *Information* 15 (2024) 509.
- [30] S. Meckler, Procedure model for building knowledge graphs for industry applications, *arXiv preprint arXiv:2409.13425* (2024).

- [31] B. Zhang, H. Soh, Extract, define, canonicalize: An llm-based framework for knowledge graph construction, in: *Proceedings of the 2024 conference on empirical methods in natural language processing*, 2024, pp. 9820–9836.
- [32] X. Li, Y. Yuan, Y. Yang, Y. Guan, H. Wang, J. Jiang, H. Shi, X. Liu, Quality-controllable automatic construction method of chinese knowledge graph for medical decision-making applications, *Information Processing & Management* 62 (2025) 104148.
- [33] H. Li, Z. Shi, C. Pan, D. Zhao, N. Sun, Cybersecurity knowledge graphs construction and quality assessment, *Complex & Intelligent Systems* 10 (2024) 1201–1217.
- [34] X. Feng, X. Wu, H. Meng, Ontology-grounded automatic knowledge graph construction by llm under wikidata schema, *arXiv preprint arXiv:2412.20942* (2024).
- [35] Subterráneos de Buenos Aires (SBASE), Subte: Gtfs, 2026. URL: <https://data.buenosaires.gob.ar/dataset/subte-gtfs>.
- [36] Transitland, Gtfs feed: f-69y7-recursosdatabuenosairesgobar, 2026. URL: <https://www.transit.land/feeds/f-69y7-recursosdatabuenosairesgobar/versions/3c17116d0885c269a8adade6f170877dfe3d6bf1>.