# Linked Open Data and Knowledge Graph

Paper title: Knowledge Graph Completeness: A Systematic Literature Review

Habiba Naeem

# Agenda

1. About
2. Introduction
3. Motivation
4. Completeness and its types
5. Research question
6. Methodology
7. Metrics, Approaches and Challenges
8. Tool Analysis
9. Challenges
10. Conclusion
11. Future work

# About

- **Title:** Knowledge Graph Completeness: A Systematic Literature Review.
- **Authors:** Subhi Issa, Onaopepo Adekunle, Faycal Hamdi, Samira SI- Said Cherfi, Michel Dumontier and Amrapali Zaveri
- **Published In:** IEEE Access

  **Publication Date:** 02 February 2021
- Open Source

1. The Problems addressed
2. Approaches and metrics proposed
3. Tools developed to assess completeness

# Introduction

1. The paper evaluates the quality of Knowledge Graphs (KGs), focusing specifically on completeness as a key quality dimension.
2. Completeness is essential for assessing fitness for use in applications.
3. The study performs a **Systematic Literature Review (SLR)** to analyze **56 articles** on KG completeness, unifying and formalizing terminologies.
4. Seven types of completeness are identified, including three not recognized in earlier surveys.
5. Nine tools for assessing KG completeness are reviewed.

The aim is to provide researchers and data curators with a deeper understanding of completeness and encourage new approaches.

# Motivation

- Modern web technologies like RDF allow publishing vast amounts of interconnected data.
- However, Knowledge Bases often face problems with missing data, known as "completeness issues."
- Completeness is crucial because it influences other aspects like accuracy, timeliness, and consistency.
- Solving these issues makes Knowledge Graphs more useful and reliable for various applications.

# Completeness in Knowledge Graph

- Completeness is a key measure of data quality, indicating the amount of information present in a dataset.

- Incomplete data can result in inaccurate analysis and affect timeliness in decision-making.

# Research Question

How can we assess the completeness of Knowledge Graphs, considering various types and approaches?

**Sub-questions:**

1. What types of completeness exist for Knowledge Graphs?
2. What approaches and metrics are used to measure completeness?
3. What data completeness problems are being discussed?
4. What tools are available to detect completeness?

# Methodology

- Exp. 1: ((“Knowledge Graph”) OR (“Linked Data”)) AND (quality OR assessment OR evaluation OR methodology OR measuring OR completeness)

- Exp. 2: (“Linked Open Data”) AND (quality OR assessment OR evaluation OR methodology OR measuring OR completeness)

- Exp. 3: (KG OR LOD) AND (quality OR assessment OR evaluation OR methodology OR measuring OR completeness)

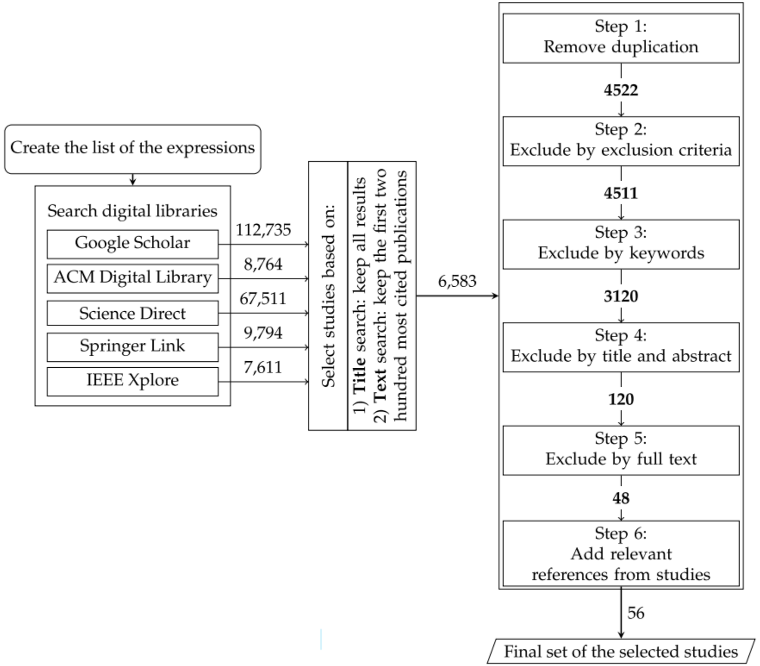


**FIGURE 2.** **Overview of the systematic literature review methodology.**

# Types of completeness

Schema

Ensures all required classes and properties are present in the schema.

Property

Validates the presence of specific property values.

Population

Measures how well the dataset covers real-world objects.

Interlinking

Checks whether instances in the dataset are linked to equivalent instances in other datasets.

Currency

Examines how property values evolve over time.

Metadata

Observes whether sufficient metadata is available about the dataset.

Labelling

Ensures all entities have clear, human- and machine-readable labels.

# Completeness illustrated using Einstein example

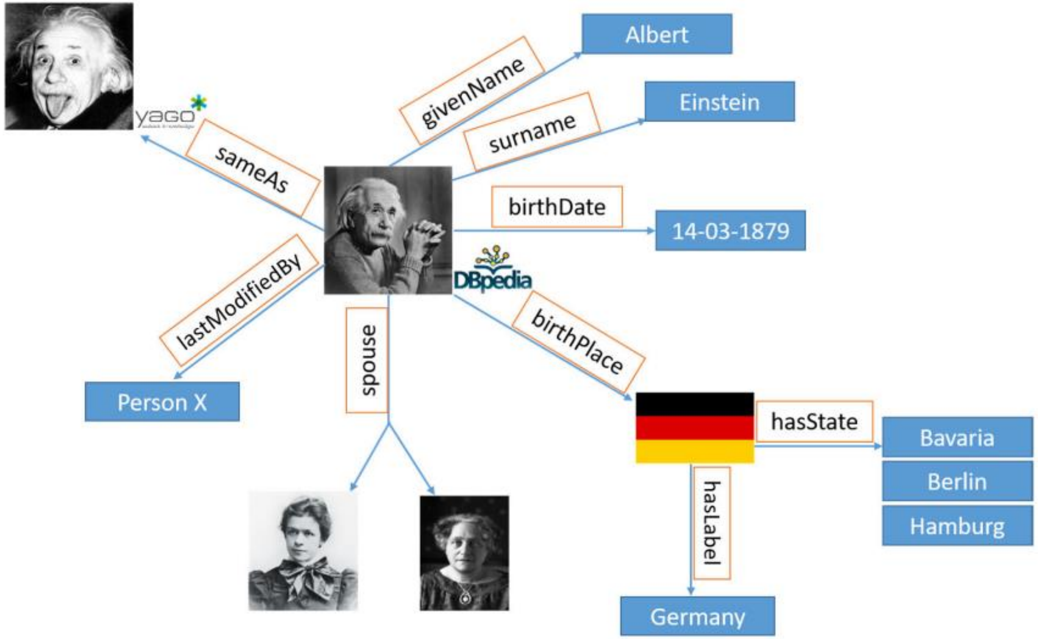
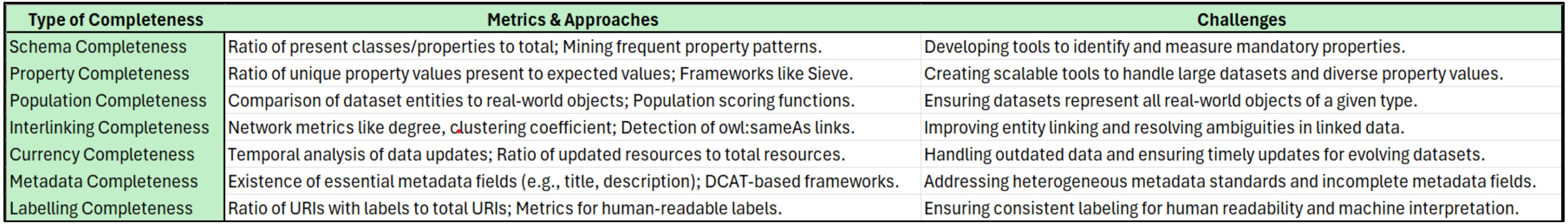


**FIGURE 1. Example of Knowledge Graph instance illustrating various types of completeness.**

# Metrics, Approaches and Challenges



| Type of Completeness | Metrics & Approaches | Challenges |
|---|---|---|
| Schema Completeness | Ratio of present classes/properties to total; Mining frequent property patterns. | Developing tools to identify and measure mandatory properties. |
| Property Completeness | Ratio of unique property values present to expected values; Frameworks like Sieve. | Creating scalable tools to handle large datasets and diverse property values. |
| Population Completeness | Comparison of dataset entities to real-world objects; Population scoring functions. | Ensuring datasets represent all real-world objects of a given type. |
| Interlinking Completeness | Network metrics like degree, clustering coefficient; Detection of owl:sameAs links. | Improving entity linking and resolving ambiguities in linked data. |
| Currency Completeness | Temporal analysis of data updates; Ratio of updated resources to total resources. | Handling outdated data and ensuring timely updates for evolving datasets. |
| Metadata Completeness | Existence of essential metadata fields (e.g., title, description); DCAT-based frameworks. | Addressing heterogeneous metadata standards and incomplete metadata fields. |
| Labelling Completeness | Ratio of URIs with labels to total URIs; Metrics for human-readable labels. | Ensuring consistent labeling for human readability and machine interpretation. |

# Tools Analysis

The study identified **nine tools** used for assessing Knowledge Graph completeness. These tools are categorized as automatic or semi-automatic based on their operation and focus on different types of completeness.

Sieve

Assesses schema, property, and interlinking completeness using a scoring function.

Loupe

Inspects datasets with a visual explorer, focusing on schema, property, and population completeness.

Luzzu

Customizable quality assessment for schema, property, interlinking, metadata, and labelling completeness.

Link-QA

Automatically assesses interlinking completeness using network metrics.

LiQuate

Uses Bayesian networks for population and interlinking completeness.

# Tools Analysis

LODsyndesis — Detects interlinking completeness using lattice-based algorithms..

Slint+ — Finds owl:sameAs links between datasets to check interlinking completeness.

LODQM — Automatically evaluates schema and property completeness with goal-question-metric (GQM) approaches.

KBQ — Assesses currency completeness through temporal analysis of dataset changes.

# Challenges

1. **Open World Assumption:** Metrics struggle with undefined data.
2. **Lack of Gold Standards:** No complete reference datasets.
3. **Data Maintenance:** Errors propagate across linked datasets.
4. **Tool Limitations:** Many tools require manual setup; some lack support.
5. **Diverse Dimensions:** Completeness interacts with other quality metrics.
6. **Scalability Issues:** Difficulties handling large, dynamic datasets.
7. **Metadata Gaps:** Incomplete metadata affects usability and discoverability.
8. **Question Answering:** Needs robust metrics for accurate filtering.

# Conclusion

- Linked Open Data (LOD) principles are widely applied in domains like life sciences, media, medicine, and e-government.
- High-quality data is crucial as poor data quality has significant economic impacts (e.g., $3.1 trillion/year in the US).
- The study focuses on **completeness**, a vital dimension of Knowledge Graph quality.
- Analyzed **56 studies**, classified **7 types of completeness**, and identified metrics and tools for assessment.
- Addressed the main research question by summarizing approaches, identifying problems, and exploring gaps in LD completeness research.
- Provides a foundational document for researchers in Knowledge Graph completeness.

# Future Work

- **Expand Search Scope:** Include additional keywords (e.g., RDF dataset, RDF graph) and synonyms like "coverage."
- **Investigate Other Dimensions:** Explore data quality dimensions such as **accuracy** and **timeliness**.
- **Improve Search Strategies:** Develop a broader search strategy to capture completeness with alternative terms (e.g., schema, ontology completeness).
- **Replicate for Other Dimensions:** Apply the review methodology to other data quality dimensions for in-depth analysis.
- **Encourage Further Research:** Serve as a guide for future researchers to address gaps in Knowledge Graph quality.