Quantifying Knowledge Wealth to Reveal Knowledge Gaps: A Case Study of Wikidata

Authors[1]

[1]*Affiliations*

Abstract

The increasing volume of data has amplified the demand for structured, machine-readable information, making knowledge graphs essential for organizing and storing knowledge. Understanding and ensuring the quality of knowledge graphs is essential to improving their fairness and usability. One key aspect of knowledge graph quality is knowledge wealth, which represents the amount of information an entity possess. A higher knowledge wealth suggests a more informative and robust knowledge graph, whereas a low wealth level indicates gaps and potential biases. However, there is no standardized method to quantify and analyze knowledge wealth. This study introduces a novel framework for quantifying knowledge wealth and assessing knowledge gap within knowledge graphs, and focuses on Wikidata as a case study. The proposed approach formalizes three notions of knowledge wealth based on property cardinality, property type, and link direction, and employs statistical measures and visualizations, including metrics such as the Gini coefficient and Lorenz curves, to analyze the distribution of knowledge wealth across entity classes. We also develop a Python-based implementation that supports both analysis and insight generation. Through empirical evaluation, we reveal significant knowledge gaps across gender (males vs. female entities) and geographic regions (western vs. non-western entities). We also examine how different definitions of knowledge wealth influence inequality measurements. Additionally, we uncover a structural gap across property types, where certain properties (e.g., object, literal, ID) contribute unequally to entity wealth, leading to systemic imbalances. Our findings offer valuable insights for identifying underrepresented entities and support data enrichment efforts aimed at closing knowledge gaps in open KGs.

Keywords

Knowledge graph, knowledge wealth, knowledge gap, data quality, Wikidata

1. Introduction

Wealth is the abundance of valuable possessions or money, or the state of having this abundance [1]. In economics, wealth is defined as the current market value of all assets owned by individuals or households, calculated as total assets minus liabilities [2]. When considering individual (human) wealth, the most common measure is net worth. One important application of this measure is in official asset reporting, where many governments require public officials to disclose their financial for transparency and accountability. For instance, the United States of America mandates a public financial disclosure report, while the European Union requires a declaration of interests from its commissioners. In Indonesia, this is done through the *La-*

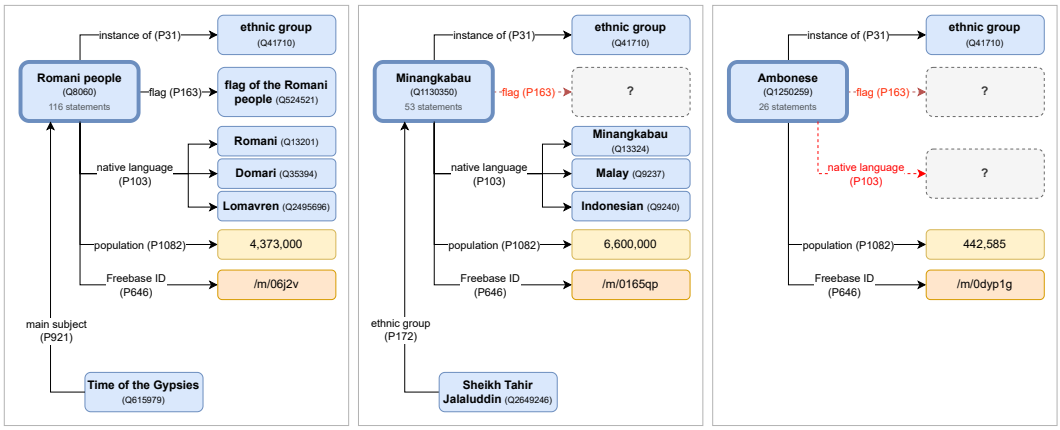


Figure 1: Data on Wikidata about the Romani people, the Minangkabau, and the Ambonese retrieved on 17th of February. This figure illustrates how entities of the same class can have variations in their representation, including differences in *(i)* the properties associated with them, *(ii)* the data types of property values (e.g., objects vs. literals vs. IDs), *(iii)* the cardinality of properties (single vs. multiple values), and *(iv)* the direction of information association (whether entities appear in the subject or object position in RDF triples).

poran Harta Kekayaan Penyelenggara Negara[1] (LHKPN), which includes components such as immovable assets (e.g., land and buildings), movable assets (e.g., vehicles), securities, cash and its equivalents, receivables, and liabilities. On the other hand, in knowledge graphs (KGs), wealth can be defined as the amount of information (in terms of properties or links) an entity possesses. Just like the different components of wealth in LHKPN, the wealth of an entity in a KG can be calculated based on different types of properties or links, and with different notion of calculation model.

Figure 1 shows three Wikidata entities of the type ethnic group: the Romani people, the Minangkabau, and the Ambonese, along with some information about them. For example, the entity Romani people includes information such as its class, flag, native languages, population, and associated Freebase ID. We can observe that a type of information can have a single value, such as the property *flag* for each entity, which is exactly one (or zero, if the information does not exist), or multiple value, such as the property *native languange*. Both of these properties have values that are other Wikidata entities. Additionally, there are other types of properties, such as *population*, which has a static numerical value, and *Freebase ID*, which provides a string used to identify the entity in the Freebase database.

While the previous examples focus on information with outgoing links, we can also consider the opposite perspective by examining incoming links. This approach does not directly show the possessions of an entity but rather indicates its popularity by showing how often it is mentioned elsewhere. This is illustrated by the Romani people being mentioned in *Time of the Gypsies* and the Minangkabau being associated with Sheikh Tahir Jalaluddin.

[1]State Official Wealth Report, a mandatory annual report that requires government officials to report their assets to Indonesia's Corruption Eradication Commission (KPK)

The example in Figure 1 provides a clear picture of how entities in Wikidata can have different kinds and amounts of information, which may lead to knowledge imbalance. If this issue is left unaddressed, it can be problematic for anyone utilizing open KGs as a data source. Data users may draw invalid inferences and conclusions based on incomplete or imbalanced data, such as the Minangkabau and Ambonese are less important than the Romani. This imbalance becomes even more striking when considering that the Minangkabau population (6.6 million) is larger than the Romani (4.3 million), yet the Minangkabau entry is less developed—missing basic properties like a flag, and having fewer than half the number of statements (53 vs. 116). If contributors to open KGs cannot identify which entities or classes are lacking information, efforts such as editathons may not be effective, potentially widening the gap between information-rich and information-poor entities. For example, contributors might prioritize enriching the Romani people's data while overlooking gaps in the Minangkabau entry, leaving the *flag* property in the Minangkabau empty despite the well-documented existence of the Marawa flag of Minangkabau[2].

Existing approaches often focus on (...), lacking a way of quantifying the amount of information contained in KGs. However, quantification of the amount of information is important to ensure more effective effort in tackling the imbalance and incompleteness issue within KGs. From the previous example, we already see the wealth of an entity from three different views: wealth based on the cardinality of property, wealth based on types of property, and wealth based on the direction of link. In addressing the aforementioned gap, our study proposes a formal model to define the knowledge wealth in the RDF knowledge graph.

Specifically, we focus on three key contributions: *(i)* introducing three notions of quantifying knowledge wealth for knowledge graphs and demonstrating how they can be used to further characterize knowledge wealth; *(ii)* implementing the formal and insight model using Python and making it accessible for broader use; and *(iii)* conducting a knowledge gap analysis on Wikidata classes, highlighting how knowledge wealth varies across gender and regional groups, how different definitions of wealth impact inequality measures, and how structural composition of knowledge wealth across entity classes.

The remainder of this paper is organized into the following sections: Section 2 reviews related work on data completeness and knowledge gaps in knowledge graphs, highlighting existing methodologies and their limitations; Section 3 introduces the proposed knowledge wealth analytics framework, detailing its formal model, insight model, and Python-based implementation; Section 4 presents a comprehensive knowledge gap analysis on Wikidata, covering disparities across gender and regional groups, the influence of different wealth defintions on inequality, and structural differences in knowledge composition across classes; and Section 5 discusses the broader implications of the findings, potential improvements, and future research directions before concluding the study.

2. Related Work

In this section, we present related work in knowledge graph (KG) population, KG completeness, and knowledge gaps. KG population concerns adding more information about entities in a KG,

[2]See the Wikipedia entry on Marawa, the traditional flag of the Minangkabau people: https://id.wikipedia.org/wiki/Marawa

thus enriching the knowledge wealth of such entities. KG completeness deals with ensuring whether sufficient information in a KG is available for the task at hand. Intuitively, when a KG possesses more wealth, it has a better chance of being complete for a given task. Knowledge gaps, on the other hand, focus more on measuring whether knowledge wealth in a KG is accumulated in an imbalanced manner.

KG Population. Mihindukulasooriya [3] introduced a tool and a method to populate Wikidata with scholarly information from DBLP, particularly co-authors and proceedings. Furthermore, Mihindukulasooriya et al. [4] argued that Wikidata offers a sustainable approach to providing structured scholarly data and that they developed an LLM-based method to populate Wikidata with conference metadata from unstructured sources. They analyzed 105 Semantic Web-related conferences and improved the description of over 6000 Wikidata entities. Bolinches and Garijo [5] made available in Wikidata links between software and their related articles through their SALTbot tool. In addition to creating a new software item in Wikidata (if not there yet), SALTbot will add properties such as "main subject software" and also its inverse, "described by source article". These initiatives demonstrate that improving the quantity in Wikidata is crucial and that measuring how such an improvement makes a difference (by computing the wealth difference of the before-and-after state) is, therefore, a great addition to such efforts.

KG Completeness. Wang and Strong [6] devised a conceptual framework of data quality, highlighting related aspects such as completeness and the appropriate amount of data. Similarly, Zaveri et al. [7] conducted a systematic review of Linked Data (LD) quality dimensions, focusing on aspects like relevancy (R2) and completeness. Wisesa et al. [8] contributed by developing a tool to profile attribute completeness in Wikidata, aligning with initiatives aimed at assessing data quality. Issa et al. [9] provided a comprehensive review of knowledge graph completeness research, identifying seven distinct types of completeness. Additionally, Luthfi et al. [10] introduced a SHACL-based method to profile completeness in knowledge graphs, further expanding the methodologies available for evaluating data quality. In [11], Xue and Zou surveyed related work on KG quality management. They discovered that most of the work focused on accuracy and completeness. In contrast, no mention was made about the work in KG gaps/imbalances.

Knowledge Gaps. Ramadhana et al. [12] designed a tool for analyzing knowledge imbalances. The tool featured property quantification for classes and Gini index analysis. Nevertheless, the tool focused only on Wikidata and did not offer fine-grained measures for wealth and imbalances. Ramadizsa et al. [13] introduced the concept of gap properties that helps to characterize class-level knowledge gaps within knowledge graphs. The framework adapts association rule mining and empirically analyzes property gaps among various Wikidata classes. Our work complements both of the work with more general yet more detailed wealth analysis over KGs, including Wikidata. Furthermore, in [14], Abián et al. examined gaps in Wikidata content, analyzing edit metrics (contributions to Wikidata) in relation to corresponding Wikipedia pageviews (user needs). Their findings suggest that gaps in gender and recency are not driven by internal factors within Wikidata. However, gaps related to socio-economic factors may be partially endogenous

to Wikidata.

3. Knowledge Wealth Framework

In this section, we define the formal framework for measuring knowledge wealth in knowledge graphs (KGs). We start by introducing the underlying data model based on RDF structure and how entities are represented through triples. Then, we formalize different notions of knowledge wealth using property-based metrics. Lastly, we present an insight model that supports both quantitative analysis and qualitative interpretation of wealth distributions across entity classes.

3.1. Wealth Formal Model

Knowledge graphs follow Resource Description Framework (RDF) [15] as a means of data organization. Without loss of generality of how the form of the URIs is, data is stored in the form of triple (s, p, o); a combination of a subject s, a predicate p, and an object o which can be visualized as nodes and directed-arc diagrams. For example, the statement "William Shakespeare's notable work is Romeo and Juliet" in human-readable URIs is mapped to the triple (*WilliamShakespeare*, *notableWork*, *RomeoAndJuliet*). Likewise, the statement in Wikidata, which uses ID-based URIs, is mapped to (*Q692*, *P800*, *Q83186*).

There are 3 kinds of nodes: IRIs, literals, and blank nodes. A triple is in the form of $(s, p, o) \in (I \cup B) \times I \times (I \cup B \cup L)$ where I is the node with type IRIs, B is the node with type of blank node, and L is the node with type of literals. In this study, we omit the usage of blank node, as this adds complexity to the analysis and may result in quality issue.

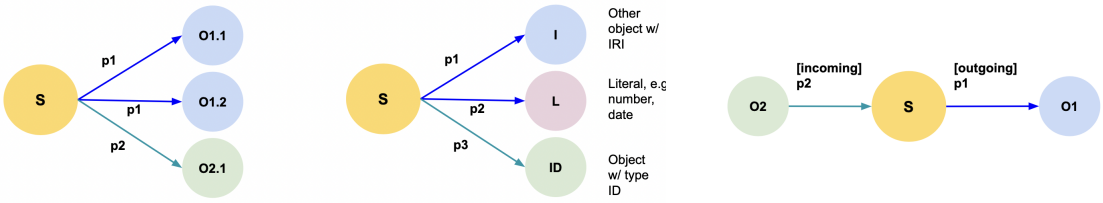
3.1.1. Entity-Level Wealth: Knowledge Wealth Type and Definition

Let s be any entity in a knowledge graph G. We quantify the wealth of entity s in G as the amount of information about s available in G. Thus, the knowledge wealth of an entity is defined by the number of properties associated/linked to it. For example, the wealth of William Shakespeare (Q692) in Wikidata counts all triples describing Q692 in Wikidata, including those detailing his family, occupation, image, and so on.

There are several notion on how to calculate the knowledge wealth of an entity: (1) wealth based on the (non-)uniqueness of individual property; (2) wealth based on type of property; and (3) wealth based on the direction of the link. The wealth of s with regard to graph G for each wealth category is denoted by W, formalized and explained as follows.

Wealth based on the (non-)uniqueness of individual properties. The first measure of the knowledge wealth of s is bag of properties—the cardinality of set of all triples that has s in their subject position. In this definition, the triples (s, p_1, o_1) and (s, p_1, o_2) account for a wealth of 1 each, thus both have a total of 2. Let $N_{bag}(s, G)$ be a set that comprises all pair of predicate/property and object (p, o) that is connected to s. Then $W_{bag}(s, G)$ is the cardinality of $N_{bag}(s, G)$.

$$N_{bag}(s, G) = \{(p, o) | (s, p, o) \in G\}$$
$$W_{bag}(s, G) = |N_{bag}(s)|$$



(a) Illustration of bag of properties and set of properties

(b) Illustration of 3 types of property: object, literal, ID

(c) Illustration of incoming link vs. outgoing link

Figure 2: Three simple graphs

Another way of measuring the wealth is by counting the number of distinct properties describing the entity, or set of properties. By this way, we are capturing the variety of information about an entity. In contrast to bag of properties, in set of properties (s, p_1, o_1) and (s, p_1, o_2) would be regarded as the "same" information because of the identical property p_1, thus they only account for a total wealth of 1. Let $N_{set}(s, G)$ be a set that comprises all predicate/property p that is connected to s. Then $W_{set}(s, G)$ is the cardinality of $N_{set}(s, G)$.

$$N_{set}(s, G) = \{p | \exists o, (s, p, o) \in G\}$$

$$W_{set}(s, G) = |N_{set}(s, G)|$$

By the above definition, the wealth of entity s in Figure 2a is 3 and 2, using bag of properties and set of properties respectively.

Wealth based on type of property. Object properties are other entities besides s that is connected with s through a property p. Wealth of s using bag of properties with only considering the object properties is defined as:

$$W_{bag,object}(s, G) = |\{(p, o) | ((s, p, o) \in G) \cap (o \in I)\}|$$

Literal properties are non-object properties that is connected with s through a property p. Wealth of s using bag of properties with only considering the literal properties is defined as:

$$W_{bag,literal}(s, G) = |\{(p, o) | ((s, p, o) \in G) \cap (o \in L)\}|$$

An external ID is a special type of string that is used to represent an entity in an external source. In Wikidata, an ID is identifiable by property type *wikibase:ExternalId*. Just like any other property, an ID is connected with s through a property p. Let $C_{ID,G}$ be a set comprising ID property in graph G. Wealth of s using bag of properties with only considering the ID properties is defined as:

$$W_{bag,ID}(s, G) = |\{(p, o) | ((s, p, o) \in G) \cap (o \in L) \cap (o \in C_{ID,G})\}|$$

Wealth based on the direction of the link. In outgoing link type of wealth, the properties that are used in the wealth calculation of an entity s are those obtained from link with outwards direction from that particular entity s; that is where s appears to be the subject in the set of triples in graph G. All types of wealth defined before use the notion of outgoing link.

In incoming link type of wealth, the properties that are used in the wealth calculation of an entity s are those obtained from link with inwards direction to that particular entity s; that is where s appears to be the object in the set of triples in graph G. To illustrate, let $N_{bag}(s)$ be a set that comprises all pair of object and predicate/property (o, p) that is connected to s in incoming direction to s i.e., $N_{bag}(s) = \{(o, p)|(o, p, s) \in G\}$. Then the wealth of s using bag of properties and the view of incoming link is notated as $W_{bag,incoming}(s, G)$, and equal to the cardinality of $N_{bag}(s)$.

$$N_{bag,incoming}(s, G) = \{(o, p)|(o, p, s) \in G\}$$
$$W_{bag,incoming}(s, G) = |N_{bag,incoming}(s, G)|$$

Looking at in Figure 2c, the wealth of entity s is 1 using outgoing link, which is from the triple $(s, p_1, o1)$. Its wealth is also 1 and using incoming link, which comes from the triple $(o2, p_2, s)$.

Each definition above can be used simultaneously. For example, the wealth of entity s using set of properties, calculating object and data but not ID properties, and using the direction of outgoing link is denoted by $W_{set,outgoing,(object \cup data) \cap ID^{\complement}}(s, G) = |N_{set,outgoing,(object \cup data) \cap ID^{\complement}}(s, G)|$ with $N_{set,outgoing,(object \cup data) \cap ID^{\complement}}(s, G) = \{p|\exists o, (s, p, o) \in G, \cap (o \in ((I \cup L) \cap C^{\complement}_{ID,G})\}$

3.1.2. Class-Level Wealth

In this study, we re-use the class model defined by Ramadizsa et al. [13]. A class is a group of entities that are the subject of the study. *Human*, *film*, and *taxon* are some examples of class. In general, entity s is an instance of class C is expressed by the triple (s, *instanceOf*, C) or (s, *type*, C). We can get a more narrow class inside the defined class by specifying additional conditions, each consisting of a particular property and value associated with it. Example of such conditions for human class is *gender* with associated value *male*, while example for a country would be *continent* with value *Asia*. For instance, the class of human with gender male that lived during English Renaissance is queried using ($?s$, {($?s$, *instanceOf*, *human*), ($?s$, *gender*, *male*), ($?s$, *timePeriod*, *EnglishRenaissance*)}).

Let C be a class that consists of m distinct entities s_1, s_2, ... s_m in graph G. We define T_C a multiset consisted of the wealth of each entity of C, i.e.,

$$T_C = \{W(s_1), W(s_2), W(s_3), ..., W(s_m)\}.$$

The overall wealth of class C can be be quantified using its constituent entities and be expressed as a function of T_C, denoted as $f(T_C)$, where f may represent statistical summaries such as the cardinality (i.e., entity count), mean, median, mode, or percentile of its entities's wealth. For example, let a class C consists of 4 entities s_1, s_2, s_3, and s_4 from Figure 3. If we use cardinality as the function f, then the wealth of class C is 4, corresponding to the number of entities it contains.

3.2. Insight Model

Exploratory Data Analysis (EDA): Descriptive Statistics Measures. Descriptive statistics is concerned with the description and summarization of data. It is a summary of a dataset that helps to describe features of data quantitatively (Ross, 2019). To have a general view of wealth distribution of a class, we use the following measures:

- measures of central tendency: mean, median, mode
- measures of frequency: count, cumulative frequency/percentage
- measures of position: quartile, percentile
- measures of dispersion: minimum, maximum, range, interquartile range, standard deviation, coefficient of variation, kurtosis
- measures of symmetry: skewness

Gini Coefficient. Gini coefficient is a metric used to measure the economic wealth gaps between countries. A study by Akbar (2020) utilized the Gini coefficient to measure the level of knowledge imbalance in knowledge graphs, particularly Wikidata classes. To calculate the imbalance level of a Wikidata class using Gini coefficient, the researcher started by calculating the number of properties of each entity of that particular class and storing them in an array. The array will then be sorted in descending order, from the smallest to the largest i.e., $y_i \geq y_{i+1} \forall i \in \{1, 2, ..., n\}$. The Gini coefficient will be calculated from the sorted array using the Gini coefficient formula below [16].

$$G = 1 - \frac{1}{n^2 \mu} \sum_{i=1}^{n} \sum_{j=1}^{n} Min(y_i, y_j)$$

$$G = 1 + \frac{1}{n} - \frac{1}{n^2 \mu} - (y_1 + 2y_2 + ... + ny_n)$$

In economics context, n is the size of population of a country, μ is the average income, and y is an array containing data of each country's income. However, in the context of knowledge graph, n is the number of entities in the class, μ is the average knowledge wealth of the entities, and y is an array containing data of each entity's wealth, sorted in descending order.

For example, let's say we have a class that consists of 10 entities. After counting the number of properties of each entities (using the notion of bag of properties for wealth) and sorting them in descending order, we will have an array of $y = [10, 8, 8, 7, 4, 2, 2, 1, 1, 1]$. The length of the array is $n = 10$ and the average wealth is $\mu = 4.4$. Then, apply the Gini coefficient formula and we get $G = 1 + \frac{1}{10} - \frac{2}{10^2 \times 4.4}(1 \times 10 + 2 \times 8 + ... + 10 \times 1) = 0.414$

For another another example, let's say we have another array of 10 entities $z = [10, 9, 9, 9, 9, 9, 9, 9, 9, 5]$. By applying the same formula to z, we get a Gini coefficient value of 0.052.

The Gini coefficient has a value between 0 and 1. The higher the coefficient value, the greater the imbalance level. The value of 0 is achieved when all observed entities have the same amount of wealth. The value of 1 occurs when all income is owned solely by one entity and this phenomenon expresses full inequality.

Lorenz Curve. Lorenz curve is a graphical representation of wealth distribution and its inequality [17]. It shows how the wealth is cumulatively distributed, with data points sorted in ascending order from the poorest to the richest. In Lorenz curve, the horizontal axis represents the fraction of the population, and the vertical axis represents the cumulative wealth. Therefore, if the point $(x, y) = (30, 15)$ lies on the curve, then we can interpret that the bottom 30% of the population account for 15% of the total wealth in that population. The Lorenz curve is usually drawn along with a straight diagonal line with a slope of 1. This straight line represents perfect equality in wealth distribution, i.e., each individual in the observed population has equal wealth. The Lorenz curve itself is drawn below the straight line. The ratio of the area between the Lorenz curve and the straight line of perfect equality to the triangular area below the straight line, is the Gini coefficient.

3.3. Sample Application of Formal and Insight Model

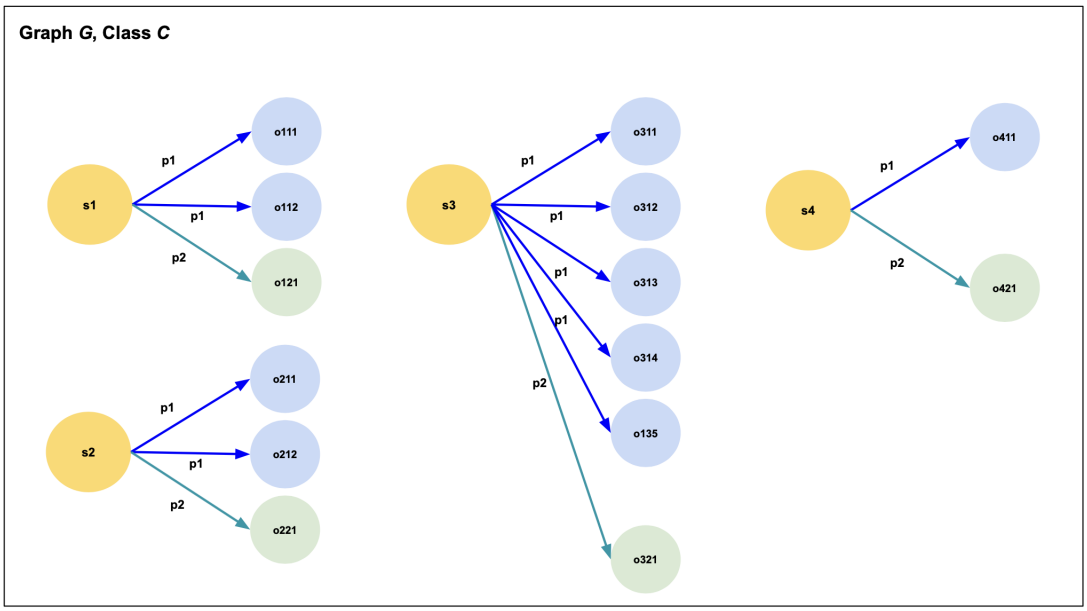


Figure 3: Sample knowledge graph G that contains class C with 4 entities

We provide small case in Figure 3 to illustrate how both models can be applied in quantifying knowledge wealth. In this example, we will focus on the notion using bag of properties with outgoing direction of link to quantify the wealth.

A graph G has a class C, which consists of four entities s_1, s_2, s_3, and s_4. Each entity has two distinct properties p_1 and p_2. For example, entity s_1 is linked by property p_1 to objects o_{111} and o_{112}, and by property p_2 to object o_{121}. Using the bag of properties and outgoing link direction, the wealth of entity s_1 is 3 (2 accounted for by p_1 and 1 by p_2). Similarly, the wealth of entities s_2, s_3, and s_4 is 3, 6, and 2, respectively. Table 1 provides a statistical summary describing the wealth of class C. Entities in class C have a mean wealth of 3.5, a median wealth of 3, a mode

wealth of 3, a minimum wealth of 2, and a maximum wealth of 6. Based on each individual entity's wealth, the imbalance measure of class C is quantified using the Gini coefficient, which has a value of 0.21.

Table 1: Statistical Summary of Wealth of Class C

Measure	Entity Count	Mean	Median	Mode	Minimum	Maximum	Gini
Value	4	3.5	3	3	2	6	0.21

This table shows some statistical measures to quantify the wealth of class C.

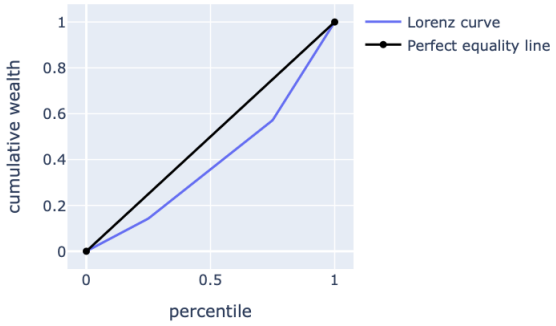


Figure 4: Lorenz curve of class C

In addition to the Gini coefficient, the Lorenz curve for the wealth of entities in class C is shown in Figure 4. This figure illustrates that the wealth distribution within class C is very close to the diagonal line of perfect equality, which aligns with the small Gini coefficient value of 0.21.

3.4. Python-Based Model Implementation

Our study provides a Python-based implementation library for both the formal model and the insight model. The library incorporates the three notions of knowledge wealth discussed in Subsection 3.1, as well as the insight model outlined in Subsection 3.2. While the model is platform-agnostic, i.e., it can be applied to any KG, we focus its implementation on Wikidata for demonstrating its use cases. Wikidata is specifically chosen because of the availability of Wikidata Query Service which facilitates structured queries over its data.

The notions of wealth are implemented at the query level, meaning that the filters and aggregations based on each wealth definition are directly embedded in SPARQL queries. The flow begins with defining class filters to specify the entities of interest. Next, SPARQL querying is performed using Wikidata Query Service to retrieve RDF triples matching the specified criteria and aggregate them according to the selected notion of wealth. The extracted data is then structured into pandas DataFrames, enabling further analysis through the insight model using Python libraries[3].

[3]Our library is available at https://github.com/nurulputri/paper-knowledge-wealth-framework

In our implementation, we only consider direct properties and exclude blank nodes to reduce complexity and maintain data quality. The complete flow of the formal model and insight model usage is illustrated in Figure 5.

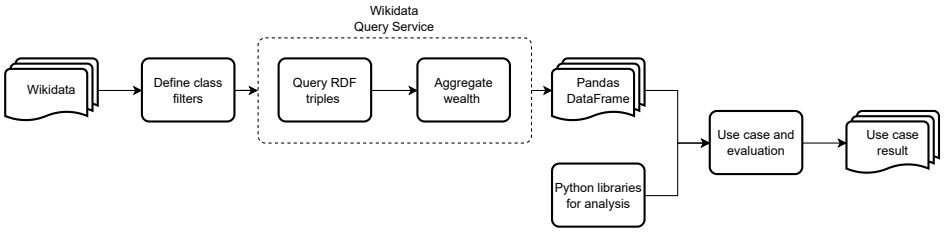


Figure 5: Model implementation and evaluation flow using Python Jupyter Notebook

4. Knowledge Gap Analysis and Evaluation

This section presents the analysis and evaluation of knowledge gaps using real-world use cases. The analyses focus on identifying disparities in knowledge representation. All data were obtained from Wikidata via the Wikidata Query Service, retrieved on March 2nd, 2025.

4.1. Group-Level Gap in Representation in Wikidata

In this subchapter, analysis is conducted to see whether any particular entity group in Wikidata is underrepresented compared to others. There are 2 analysis done: gender-based gap and regional gap.

Gender-Based Knowledge Gap: Male vs. Female Representation. Gender-based knowledge gap analysis in Wikidata will be performed on 10 Wikidata classes: computer scientist, American singer, American actress/actor, badminton player, businessperson, lawyer, American politician, American writer, American researcher, and American journalist. These classes are chosen to ensure variation in occupations, as gender representation differs across professions. Additionally, six of these classes are limited to American entities to manage scalability, as querying large, global datasets can lead to excessive query time and computational resource constraints. The United States, being a large and well-known country, is expected to provide a broadly reflective sample while keeping the analysis computationally feasible.

To analyze the gap, the first aspect that will be considered is the proportion of each gender in every class. We assumed that there are equal numbers of males and females in real-world and this will be the basis to determine if there is any gap in the data. Pearson's chi-square test (goodness-of-fit) is then performed to test the null and alternative hypotheses with significance level of α =5% as follows:

Gender-Based Knowledge Gap: Entity Count Gap (Pearson's chi-square test)	
H_0	The proportions of males and females in a particular class are equal to the assumed real-world proportion of 50%-50%.
H_1	The proportions of males and females in a particular class are not equal to the assumed real-world proportion of 50%-50%.
Insight 1:	Across all 10 classes, male entities significantly outnumber female entities in all classes. This suggests a systematic underrepresentation of female entities in Wikidata.

Table 2: Entity Count of 10 Wikidata Classes per Gender Category

Class Name	Entity	Male	Female	%Male	%Female	χ^2	p-value
American actress/actor	38655	21787	16868	0.563627	0.436373	625.96	3.78e-138
American journalist	18033	12402	5631	0.687739	0.312261	2542.36	0.0
American politician	96507	85800	10707	0.889055	0.110945	58430.57	0.0
American researcher	5233	3634	1599	0.694439	0.305561	791.37	4.06e-174
American singer	16038	9198	6840	0.573513	0.426487	346.69	2.23e-77
American writer	33554	19656	13898	0.585802	0.414198	988.10	6.95e-217
Computer scientist	19049	16180	2869	0.849388	0.150612	9301.42	0.0
Badminton player	25427	13493	11934	0.530656	0.469344	95.59	1.42e-22
Businessperson	76758	68583	8175	0.893496	0.106504	47540.67	0.0
Lawyer	94479	83147	11332	0.880058	0.119942	54587.73	0.0

This table shows the entity count of 10 Wikidata classes per Gender Category. Chi-square test result shows the significance of difference between the entity count of the two genders male and female.

From Table 2, we can see that there are more male entities than female entities in all of the classes. In terms of entity count, the gender gaps in some classes such as American singer, American actress/actor, badminton player, and American writer, are slim. The gender gaps in some other classes are huge, and it can be observed in the classes of computer scientist, businessperson, lawyer, American politician, journalist, and researcher. This phenomenon can also be easily identified through visualization, as exhibited in Figure 6a, where the histogram of the female subclass is much smaller compared to the male. Looking at the chi-square test result, as p-value is well below the chosen significance level, the null hypothesis is rejected in all classes. Hence, we considered the difference of entity count to be significant and conclude that the proportions of males and females in each Wikidata class are not the same as the assumed real-world proportion of 50%-50%.

However, it is arguable that, for some classes, the gap in entity count between both genders is expected because, in reality, there are more men than women in the workforce, especially in particular fields such as engineering. As a consequence, it is not reasonable if we expect to have an equal number of males and females entities in Wikidata. Therefore, entity count may not be a good measure of gap because of the nature of the data itself. To address this, we need to evaluate other metrics which can quantify the gap at entity-level.

The next metrics to be considered are the measures of central tendency and dispersion to see where the wealth distribution is concentrated and how the data spread.

Gender-Based Knowledge Gap: Measures of Central Tendency and Dispersion	
Insight 1:	For all 10 classes, female values for mean, median, and mode are generally lower than those of males. Only in the American Singer class does a female (Madonna) appear as the richest individual.
Insight 2:	Positive value of skewness and high value of kurtosis are observed across the board, indicating right-skewed distribution and frequent extreme outliers.

Table 3: Measures of Central Tendency of 10 Wikidata Classes per Gender Category

Class Name	Mean (o/m/f)	Median (o/m/f)	Mode (o/m/f)
American actress/actor	39.72/40.82/38.31	30.00/30.00/28.00	19/19/19
American journalist	31.30/33.12/27.29	24.00/25.00/21.00	14/15/14
American politician	19.15/19.31/17.90	15.00/15.00/15.00	9/9/12
American researcher	24.24/25.32/21.77	20.00/21.00/19.00	12/12/15
American singer	42.96/43.30/42.51	31.00/33.00/30.00	18/24/15
American writer	39.75/44.05/33.66	30.00/33.00/26.00	19/21/19
Computer scientist	24.30/24.65/22.33	19.00/19.00/18.00	8/8/11
Badminton player	21.90/21.64/22.19	16.00/16.00/16.00	14/14/14
Businessperson	17.06/16.98/17.73	13.00/13.00/13.00	10/10/9
Lawyer	22.58/23.16/18.33	19.00/19.00/15.00	14/16/12

This table shows the measures of central tendency of 10 Wikidata classes per gender category. Each measure will have 3 values: o (overall), m (male), and f (female).

Table 4: Measures of Dispersion and Symmetry of 10 Wikidata Classes per Gender Category

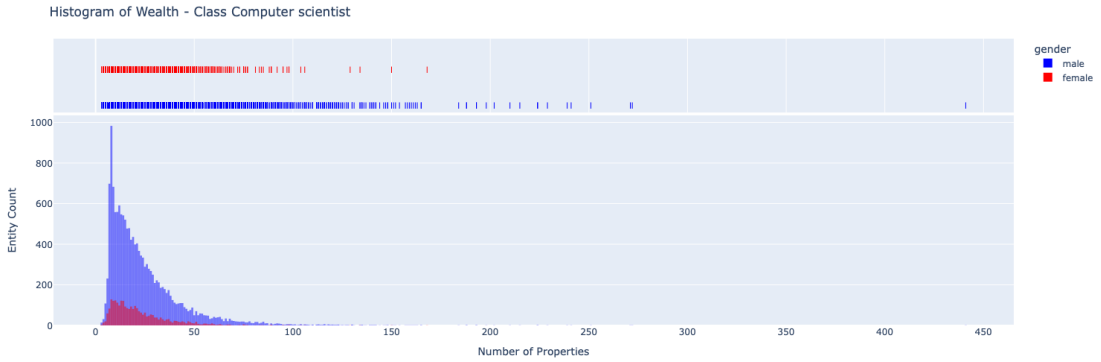
Class Name	Min (o/m/f)	Max (o/m/f)	Std. Deviation (o/m/f)	Skewness (o/m/f)	Kurtosis (o/m/f)
American actress/actor	4/4/4	703/585/703	35.15/34.47/35.97	4.22/3.65/4.88	30.55/22.97/39.11
American journalist	4/4/4	418/418/363	27.78/29.22/23.82	4.21/4.17/4.15	30.60/29.94/29.09
American politician	4/4/4	564/564/334	15.05/15.07/14.90	6.98/7.01/6.83	112.45/115.94/84.91
American researcher	4/4/4	225/225/195	17.16/18.46/13.46	3.86/3.75/3.74	25.74/23.07/31.65
American singer	4/5/4	703/585/703	40.36/36.61/44.90	4.15/3.33/4.65	29.66/19.78/33.57
American writer	4/4/4	564/564/435	35.82/39.48/28.82	3.63/3.37/4.02	21.27/18.08/27.34
Computer scientist	3/3/3	452/452/178	19.69/20.35/15.35	3.48/3.52/2.25	28.77/28.90/9.89
Badminton player	9/9/9	360/240/360	15.98/15.15/16.88	4.28/3.94/4.52	31.54/23.57/36.62
Businessperson	3/3/3	585/585/434	15.06/14.39/19.81	8.07/7.80/8.17	142.31/144.95/106.75
Lawyer	3/3/3	608/608/334	17.02/17.34/13.73	5.78/5.79/5.67	82.87/82.78/78.35

This table shows the measures of dispersion and symmetry of 10 Wikidata classes per gender category. Each measure will have 3 values: o (overall), m (male), and f (female).

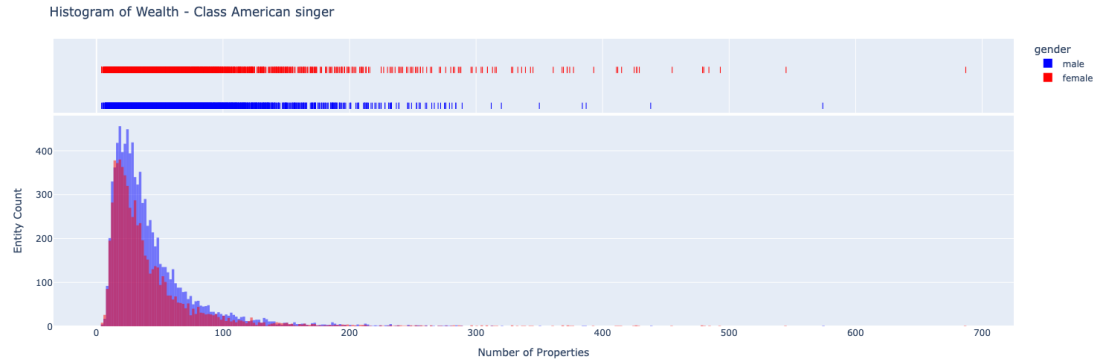
From Table 3, female entities generally have lower values of measure of central tendency (mean, median, mode). These characteristics can also be observed from the histogram in Figure 6a: female histograms' peak and dense area are located on the left of the male's. However, there are some classes in which the richest entity is a female. An example for this is the class of American Singer, which is shown by Figure 6b. Though the value of mean, median, and mode of count of properties are lower for female compared to male, the richest entity on that class is a

female entity *Madonna* (Q1744) with bag of property count of 703, with a significant difference with *Michael Jackson* (Q2831) with bag of property count of 585.

From Table 4, we also observed positive values of skewness (skewness > 0) and high kurtosis values (kurtosis > 3) in all classes, denoting the wealth distribution is right-skewed and leptokurtic. The high skewness values indicate that a small number of individuals accumulate disproportionately large property counts. Moreover, high kurtosis values across all genders suggest wealth distributions with heavy tails, meaning that extreme outliers are relatively frequent. Furthermore, we saw higher variability among males, as reflected in their generally wider range of property counts and higher standard deviations. These statistical properties point toward a more unequal wealth distribution among male entities, while females tend to cluster within a lower and more compressed wealth range.



(a) Histogram and Marginal Distribution Plot of Wealth for Class Computer Scientist



(b) Histogram and Marginal Distribution Plot of Wealth for Class American Singer

Figure 6: Histogram of wealth

At a glance we saw female classes are poorer compared to the male classes. To test this, we used t-test and Welch's test. First, we performed F-test to check if the male and female classes have equal variance. The result of F-test is then used to determine the appropiate test to be used in each class. Those with equal variance, i.e., if the p-value is more than 0.05, will use t-test; otherwise Welch's test is used. Then, we performed the appropiate tests to verify the null and

alternative hypotheses with significance level of $\alpha = 5\%$ as follows:

Gender-Based Knowledge Gap: Mean of Wealth Gap (two-sided t-test and Welch's test)	
H_0	The means of wealth of males and females in a particular class are equal.
H_1	The means of wealth of males and females in a particular class are not equal.
Insight 1:	9 out of 10 classes shows unequal mean, with 7 of them are in favor of males.

Table 5: F-Test, t-Test, and Welch's Test Result of 10 Wikidata Classes

Class Name	F-Test statistic	F-Test p-value	t-Test statistic	t-Test p-value	Welch's Test statistic	Welch's p-value
American actress/actor	0.92	0.00	-	-	6.93	4.27e-12
American journalist	1.50	1.00	13.12	4.05e-39	-	-
American politician	1.02	0.94	9.11	8.46e-20	-	-
American researcher	1.88	1.00	6.92	4.91e-12	-	-
American singer	0.66	0.00	-	-	1.19	0.23
American writer	1.88	1.00	26.44	1.98e-152	-	-
Computer scientist	1.76	1.00	5.83	5.80e-09	-	-
Badminton player	0.81	0.00	-	-	-2.75	0.01
Businessperson	0.53	0.00	-	-	-3.34	0.00
Lawyer	1.60	1.00	28.47	1.69e-177	-	-

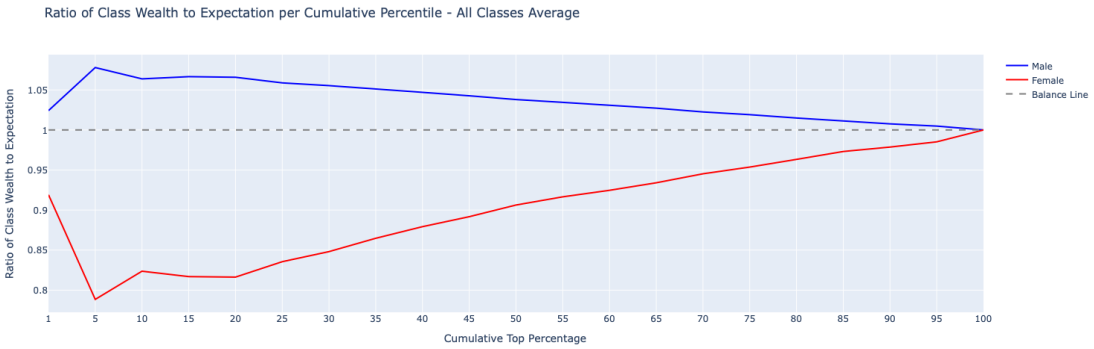
This table shows the result of F-test, t-test, and Welch's test of 10 Wikidata classes to compare the significance difference between the males and females within each class.

Based on F-test on the 10 classes, variance between male and female subclasses was found to be unequal ($p < 0.05$) in 4 classes, requiring the use of Welch's test. From the test results in Table 5, we rejected the null hypothesis in 9 out of 10 class–American singer being the only exception with p-value > 0.05. Among them, 7 classes' means are in favor of male. The other 2 classes, badminton player and businessperson, deviate from this trend, where Welch's test results indicate that the female subclasses actually have a higher mean wealth. The extremely small p-values in classes such as American writer and lawyer indicate a very strong statistical signal that the mean wealth differs significantly between genders. Finally, we may conclude that female classes are more likely to have smaller means than male classes, although certain classes may reflect localized or domain-specific representation imbalances. These results reinforce our earlier observations that female entities tend to be poorer in terms of knowledge wealth.

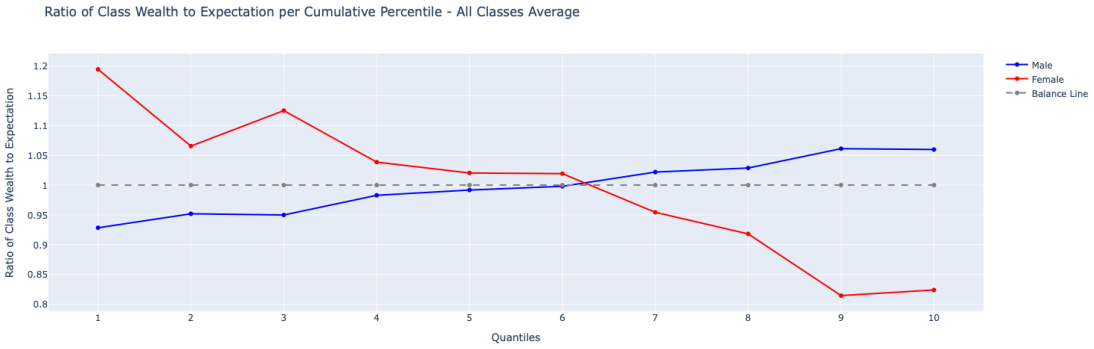
Entity count alone might not be a reliable measure of gap because the nature of the data itself. For instance, in real world, the number of male computer scientists is greater than that of female computer scientists. Therefore, expecting an equal number of male and female entities in Wikidata for this class would be unreasonable. Similarly, measures of central tendency—such as the mean or median—may offer a general overview but fail to capture the distributional nuances of individual entities within a class. To address this, a more holistic metric is needed.

Here, we introduced a new measure: the ratio of top $x\%$ representation (of male/female) relative to expectation. The value of expectation of a gender in a class is equal to the percentage of that particular in the class. Top $x\%$ male relative to expectation is the ratio of percentage of male entities in the top $x\%$ to the expectation. Similarly, top $x\%$ female relative to expectation is the ratio of percentage of female entities in the top $x\%$ to the expectation.

Gender-Based Knowledge Gap: Relative Expectation Ratios	
Insight 1:	A ratio value of 1 indicates a balanced distribution, where the group's actual share of wealth aligns with its expected share under equal representation.
Insight 2:	On average across all 8 classes, the bottom six quantiles (i.e., the lower wealth segments) are predominantly composed of female entities, while the top four quantiles (i.e., the wealthier segments) are predominantly male.
Insight 3:	At the individual class level, 6 out of 8 classes show male dominance in the upper wealth quantiles. The exceptions are the businessperson and badminton player classes.



(a) Ratio of Class Wealth to Expectation per Cumulative Top Percentage - All Classes Average



(b) Ratio of Class Wealth to Expectation per Quantile - All Classes Average

Figure 7: Ratio of each gender wealth to expectaion

When the shape of distribution of male and female of a class is the same (in other word, the wealth is distributed equivalently to male and female entities), then the value of top $x\%$ relative to expectation should be 1 for both male and female subclasses. A value higher than 1 indicates domination by that particular gender. Conversely, a value lower than 1 indicates underrepresentation.

From Figure 7 the value of ratio between top $x\%$ potion to the expectation in the above

tables, we can see that on average, the rich entities are dominated by male. Exceptions are held for 2 classes, that is classes businessperson, and badminton player. In the businessperson class, female dominance in the top quantile (Q10) creates an exception. However, quantiles 2–9 are predominantly male-dominated, albeit only slightly and very close to the balance line. Meanwhile, the lowest quantile (Q1) is significantly female-dominated. The badminton player class deviates from all other classes by exhibiting a distinct zig-zag pattern across quantiles, indicating alternating dominance between males and females, without a clear upward or downward trend.

Moreover, as we set bigger portions (higher percentage), the gap of ratio between the two ender in each class decreases i.e. the value of top $x\%$ relative to expectation of both genders converge to 1.

Regional Knowledge Gap: Western vs. Non-Western Representation. Regional knowledge gap analysis in Wikidata will be performed on 5 Wikidata classes: computer scientist, singer, memorial, university, and river. For each class, we collected the data for the western portion from 8 countries: Canada, France, Germany, Ireland, Poland, Switzerland, the United Kingdom (UK), and the United State of America (USA). For the non-western portion, we also selected 8 countries: China, Egypt, India, Indonesia, Japan, Morocco, Nigeria, and South Africa. The selected countries represent different continents to capture a diverse geographic distribution. Additionally, we focused on large and well-known countries to ensure the dataset is representative, as data from smaller countries may not provide meaningful insights due to limited coverage in Wikidata.

To analyze the gap, the first aspect that will be considered is the proportion of each regional category in every class. We assumed that there are equal numbers of western and non-western and this will be the basis to determine if there is any gap in the data. Pearson's chi-square test (goodness-of-fit) is then performed to test the null and alternative hypotheses with significance level of $\alpha = 5\%$ as follows:

Regional Knowledge Gap: Entity Count Gap (Pearson's chi-square test)	
H_0	The proportions of western and non-western entities in a particular class are equal.
H_1	The proportions of western and non-western entities in a particular class are not equal.
Insight 1:	Across all 5 classes, the proportions of western and non-western entities differ significantly. Western entities dominate in 4 out of 5 classes, with the only exception being the university class, which is skewed toward non-western entities.

Table 6: Entity Count of 5 Wikidata Classes per Regional Category

Class Name	Entity	Western	Non-western	%Western	%Non-western	χ^2	p-value
Computer scientist	6063	5446	617	0.90	0.10	3846.15	0.0
Singer	43240	31039	12201	0.72	0.28	8206.99	0.0
Memorial	4011	3836	175	0.96	0.04	3341.54	0.0
University	6124	2398	3726	0.39	0.61	287.98	1.37e-64
River	125567	70059	55508	0.56	0.44	1686.20	0.0

This table shows the entity count of 5 Wikidata classes per regional category. Chi-square test result shows the significance of difference between the entity count of the two regions.

From Table 6, we can observe that the proportion of western entities exceeds that of non-western entities in 4 out of the 5 classes. In the memorial and computer scientist classes, the gap is particularly striking, with western entities making up over 90% of the total population. Similarly, in the singer and river classes, western entities also represent the majority. The university class is the only exception, where non-western entities slightly outnumber western ones, comprising 61% of the total. Based on the results of the chi-square test, the p-values in all five classes are well below the significance threshold ($\alpha = 0.05$), leading us to reject the null hypothesis. This indicates that the differences in entity proportions between western and non-western entities are statistically significant and not due to random variation.

However, similar to entity count in gender-based knowledg gap analysis, it is debatable that some of these gaps are expected. For instance, the dominance of western representation in the computer scientist or memorial classes is not necessarily a result of data entry errors or systemic bias within Wikidata alone. Instead, it may mirror general historical and sociopolitical factors—such as unequal access to documentation, global academic visibility, and archival infrastructure. Conversely, the university class—where the number of non-western entities are higher—might be a reflection of population-driven realities, as countries like China, India, and Indonesia possess massive populations and have established a vast number of higher education institutions to serve their demographic needs. The greater number of non-western universities represented in Wikidata is probably a result of this population scale. Due to this, raw entity count alone may not be an ideal measure of gap. To more accurately assess representation and equity, we must consider other metrics that analyze wealth distribution and prominence at the entity level.

The next step in our analysis is examining the measures of central tendency and dispersion to evaluate where the knowledge wealth is concentrated and how it varies between western and non-western entities.

Regional Knowledge Gap: Measures of Central Tendency and Dispersion	
Insight 1:	In 4 out of 5 classes, western entities have higher mean, median, and mode values than non-western entities, indicating greater concentration of knowledge wealth.
Insight 2:	Western entities generally exhibit higher standard deviations, skewness, and kurtosis—suggesting greater variability, more frequent outliers, and stronger right-skewed distributions.

Table 7: Measures of Central Tendency of 5 Wikidata Classes per Regional Category

Class Name	Mean (o/w/n)	Median (o/w/n)	Mode (o/w/n)
Computer scientist	35.00/35.87/27.24	29.00/29.00/22.00	21/15/16
Singer	34.99/39.14/24.43	25.00/29.00/18.00	15/18/14
Memorial	11.04/11.04/11.13	9.00/9.00/9.00	9/9/9
University	23.11/31.61/17.63	17.50/24.00/16.00	6/6/6
River	7.69/8.55/6.60	7.00/7.00/6.00	7/7/7

This table shows the measures of central tendency of 5 Wikidata classes per regional category. Each measure will have 3 values: o (overall), w (western), and n (non-western).

Table 8: Measures of Dispersion and Symmetry of 5 Wikidata Classes per Regional Category

Class Name	Min (o/w/n)	Max (o/w/n)	Std. Deviation (o/w/n)	Skewness (o/w/n)	Kurtosis (o/w/n)
Computer scientist	4/4/5	441/441/145	25.15/25.67/18.15	3.02/3.02/2.37	20.90/20.83/8.49
Singer	3/4/3	687/687/379	33.27/36.60/18.94	4.49/4.28/3.15	35.56/31.09/21.59
Memorial	2/3/2	142/142/52	5.89/5.82/7.28	8.50/8.95/2.98	143.57/156.15/12.79
University	2/2/2	234/234/166	20.00/25.88/12.25	2.37/1.65/2.22	10.34/5.33/12.70
River	2/2/2	452/452/148	5.24/6.46/2.68	21.71/19.54/14.67	1152.84/868.56/465.42

This table shows the measures of dispersion and symmetry of 5 Wikidata classes per regional category. Each measure will have 3 values: o (overall), w (western), and n (non-western).

From Table 7, we observed that in 4 out of 5 classes, western entities consistently show higher values across all measures of central tendency (mean, median, and mode) compared to non-western entities. This suggests that western entities tend to be richer in knowledge wealth. The memorial class is the only exception, where both groups exhibit nearly identical values. A particularly extreme disparity is observed in the university class, where the mean property count for western entities (31.61) is nearly double that of non-western entities (17.63), highlighting a strong central tendency skew. This extreme pattern can also be seen in the singer class.

Turning to the measures of dispersion and symmetry in Table 8, we saw that standard deviation is generally higher for western entities across all classes except memorial, suggesting greater wealth variability among western entities. Furthermore, in the university class, western entities show a standard deviation of 25.88, more than double that of non-western entities at 12.25. We also observed positive skewness and high kurtosis across all classes and both regions, indicating that knowledge wealth distributions are right-skewed and leptokurtic. However, western entities often exhibit higher values, implying more extreme outliers and longer tails. For example, in the river class, the skewness and kurtosis for western entities are 19.54 and 868.56, respectively, compared to 14.67 and 465.42 for non-western entities.

Taken together, these statistics suggest that western entities tend not only to accumulate more knowledge wealth on average but also to dominate the extreme high end of the distribution. In contrast, non-western entities tend to cluster around lower values with tighter distribution, reinforcing regional disparity in knowledge representation.

At a glance, western entities appear to be richer compared to their non-western counterparts. To statistically verify this, we conducted a series of mean comparison tests between the two groups. We began by applying the F-test to assess whether the western and non-western subsets within each class have equal variance. The result of the F-test determines the appropriate test to be used: if the p-value is greater than 0.05, indicating equal variance, we use the standard t-test; otherwise, we apply Welch's test, which does not assume equal variance.

Regional Knowledge Gap: Mean of Wealth Gap (two-sided t-test and Welch's test)	
H_0	The means of wealth of western and non-western entities in a particular class are equal.
H_1	The means of wealth of western and non-western entities in a particular class are not equal.
Insight 1:	In 4 out of 5 classes, western entities have significantly higher mean knowledge wealth than non-western entities.

Table 9: F-Test, T-Test, and Welch's Test Result of 5 Wikidata Classes

Class Name	F-Test statistic	F-Test p-value	T-Test statistic	T-Test p-value	Welch's Test statistic	Welch's p-value
Computer scientist	2.00	1.00	8.13	5.10e-16	-	-
Singer	3.73	1.00	42.22	0.0	-	-
Memorial	0.64	0.00	-	-	-0.16	0.88
University	4.46	1.00	28.40	8.23e-167	-	-
River	5.83	1.00	66.91	0.0	-	-

The result of the F-test in Table 9 shows that only one class—memorial—has unequal variance ($p < 0.05$), requiring the use of Welch's test. For the remaining four classes, the t-test was applied. From the test results, we rejected the null hypothesis in 4 out of 5 classes, indicating that the difference in mean wealth between western and non-western entities is statistically significant. The only exception is the memorial class, where Welch's test yields a p-value of 0.88, suggesting no significant difference in mean. Among the classes with significant results, the direction of difference consistently favors Western entities. This can be seen clearly in classes such as university and singer, where the t-statistics are noticeably high (28.40 and 42.22, respectively) and the p-values are zero. These findings align with our previous observations that western entities dominate in terms of quantity and average information richness. We conclude that western entities are more likely to have higher mean knowledge wealth than non-western ones, although exceptions like the memorial class suggest that certain domains may present a more balanced knowledge distribution.

As with gender-based knowledge gap, measuring regional gap in Wikidata cannot rely solely on entity counts. The number of entities associated with western and non-western regions may reflect real-world disparities in documentation or notability, making raw counts an unreliable indicator of gap. Similarly, measures of central tendency—such as the mean or median—fail to capture how representation is distributed across the full spectrum of prominence

or wealth. To address these limitations, we applied the same metric introduced in the gender-based gap analysis: the relative expectation ratio. This measure compares a region's share of entities within a given quantile to its expected share under equal representation. A ratio of 1 indicates proportional representation, while values above or below 1 signal overrepresentation or underrepresentation, respectively. This allows for a more nuanced, quantile-level understanding of how western and non-western entities are distributed in knowledge graphs like Wikidata.

Regional Knowledge Gap: Relative Expectation Ratios	
Insight 1:	On average across all 5 classes, the less prominent segments are dominated by non-western entities, while the more prominent ones are dominated by western entities. The middle quantiles exhibit a zig-zag pattern, reflecting inconsistent representation between the two groups.
Insight 2:	At the individual class level, 2 out of 5 classes display a particularly strong western dominance. The rests exhibit a zig-zag trend.

On average across all classes, the lower four quantiles (i.e., the less prominent or "poorer" segments) are predominantly composed of non-western entities, while the top three quantiles (i.e., the more prominent or "wealthier" segments) are dominated by western entities. The middle quantiles exhibit a zig-zag pattern, reflecting inconsistent representation between the two groups. At the individual class level, the computer scientist and singer classes display a particularly strong western bias. In both classes, non-western entities are consistently overrepresented in the lower quantiles (Q1–Q4), while western entities are consistently overrepresented in the upper quantiles (Q5–Q6). This pattern suggests a clear stratification, with non-western entities disproportionately concentrated in less prominent positions. The memorial, university, and river classes, however, deviate from this pattern and exhibit a zig-zag trend across quantiles. This suggests a more unstable distribution of western and non-western entities across the quantiles.

4.2. Impact of Wealth Type on Knowledge Inequality

Impact of Wealth Type on Knowledge Inequality	
Insight 1:	Wealth based on the bag of properties is always higher than that based on the set of properties, leading to higher value of Gini coefficient.
Insight 2:	Wealth measured using incoming links shows significantly higher inequality than outgoing links. Lorenz curves confirm that very few entities have incoming links, while most have none.
Insight 3:	In general, the order of Gini coefficient values from smallest to largest is: object properties, literal properties, and ID properties.

In this subchapter, analysis is done to see how each wealth type affects the level of knowledge inequality of Wikidata classes. There are 2 ways this is done–quantitatively using Gini coefficient and qualitatively using Lorenz curve. The analysis is performed on 8 Wikidata classes, in which 4 of them are human-related class while the other 4 are not.

(a) Ratio of Class Wealth to Expectation per Cumulative Top Percentage - All Classes Average

(b) Ratio of Class Wealth to Expectation per Quantile - All Classes Average

Figure 8: Ratio of each regional wealth to expectaion

Table 10: Knowledge Wealth Type on Gini Coefficient

Class Name	Gini Bag	Gini Set	Gini Object	Gini Pure Literal	Gini ID	Gini Outgoing	Gini Incoming
American researcher	0.33	0.31	0.26	*0.65	0.50	0.33	0.77
American singer	0.40	0.39	0.29	0.36	0.53	0.40	0.82
Badminton player	0.29	0.14	0.30	*0.14	0.60	0.29	0.68
Computer scientist	0.41	0.37	0.36	*0.64	0.56	0.41	0.81
Historical painting	0.23	0.15	0.25	0.26	0.45	0.23	0.87
Memorial	0.22	0.19	0.20	0.31	0.41	0.22	0.99
Sci-fi book	0.30	0.21	0.31	0.33	0.44	0.30	0.82
University	0.44	0.41	0.39	0.49	0.53	0.44	0.91

This table shows the comparison of Gini coefficient of 8 Wikidata classes

When looking at the notion of wealth using the characteristics of (non-)uniqueness of individual properties, it is intuitive that the measure of the bag of properties will always give higher (or at least, equal) amount of wealth compared to the measure of set. Set of property will have an upper bound of number of unique property, while the bag of properties does not have

any upper bound. Moreover, using the bag of properties, a large number of triples having the same property may inflate the wealth substantially–though this is not necessarily a problem nor an advantage. This characteristics has a direct impact on inequality measure and it is well depicted on the value of Gini coefficient. From Table 10, in all clasess, the Gini coefficient using the bag of properties is always higher than of set of properties.

Using the notion of wealth by type of property, in general the smallest Gini coefficient value comes from wealth using object properties, followed by literal properties, with ID properties having the highest value. This condition holds for the four non-human classes that are being observed. However, there are anomalies in the human classes, as illustrated in Figure 9.

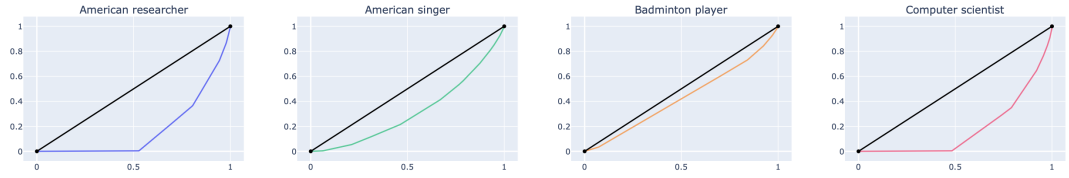


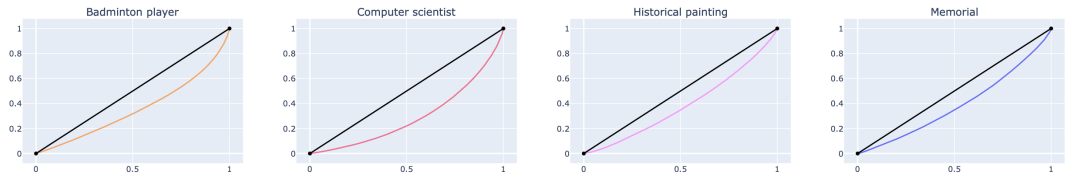
Figure 9: Lorenz curve of wealth using literal properties

In the classes of American researchers and computer scientists, the Gini coefficients for wealth using pure literal properties are significantly higher than those of other classes. This is due to approximately 50% of the entities in these classes having 0 pure literal properties–not even basic human-related literal properties such as *date of birth* (P569). For example, an American researcher *John Heidemann* (Q29354581) has a total wealth of 61 (using the bag of properties) and is among the top 4% entities in his class, yet this entity has 0 literal properties. In contrast, only 818 out of 16150 (just about 5%) entities of American singer class have 0 literal properties. Moreover, an instance of American singer, *Kris Allen* (Q216927), who also has a total wealth of 61 (using the bag of properties), has 6 literal properties.

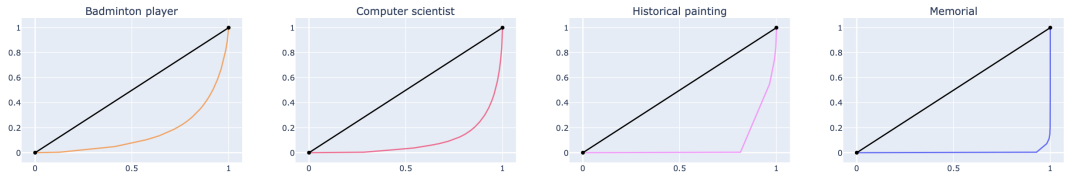
Another anomaly is observed in the class of badminton player where the Gini coefficient for wealth using pure literal is significantly lower than that of other classes. Upon inspection, three main reasons for this anomaly were identified. First, there are no entities in this class with 0 literal properties; all entites have at least 1 literal. Second, the range of wealth in this class is smaller, with a minimum value of 1 and maximum value of 10, in comparison to the classes of American researcher, American singer, and computer scientist which all have a minimum value of 0 and maximum values of 16, 26, and 54, respectively. This smaller range means the difference between the poorest and richest entities in the class of badminton player is minimal, and a small wealth range generally leads to a low Gini coefficient because inequality is limited by the narrow spread of wealth. The third reason is the large number of entities in this class, which totals 25,402. With such a high population size, the wealth of each individual—whether the poorest, middle, or richest entity—contributes only a tiny fraction to the total wealth of the class. Combined with the small wealth range, this further reduces the inequality measure, resulting in a low Gini coefficient.

Using the notion of wealth by the direction of the link, the Gini coefficient when using incoming link is always higher than using outgoing link. By inspecting the Lorenz curve, we can see that most entities do not have any incoming link, and only the small percentage of

entites has some incoming link. Figure 10 shows the comparison of Lorenz curve of knowledge wealth based on the direction of the link from 3 Wikidata classes. The difference between the two is very significant, because in Figure 10a the Lorenz curves are closer to the perfect equality line, meanwhile in Figure 10b the diagonal and the Lorenz curve almost form a right triangle which is very close to maximum inequality.



(a) Lorenz curve of wealth using outgoing link



(b) Lorenz curve of wealth using incoming link

Figure 10: Comparison of Lorenz curve of wealth based on the direction of link

4.3. Structural Gap Across Property Types in Knowledge Wealth

Structural Gap Across Property Types in Knowledge Wealth	
Insight 1:	The usage of literal properties is lower than of object and ID.
Insight 2:	Generally, the majority of knowledge wealth comes from internal resources (object and literal types).

In this subchapter, analysis is done to see the contribution of different property types to the knowledge wealth of Wikidata classes. The analysis is performed on 8 classes, in which 4 of them are human-related class while the other 4 are non-human. For each class, knowledge wealth is computed using both the bag of properties and the set of properties approaches, separately applied to object, literal, and ID property types, as well as their combinations. This allows us to observe structural disparities in how different types of information contribute to the overall wealth of entities.

Two methods of averaging are used in this analysis—global percentage and average contribution per individual entity. The first method is done by calculating the total sum of each property across all entities and then dividing it by the grand total of all properties combined, providing a holistic view of each property's overall contribution. The second method is done by first determining the percentage contribution of each property within each individual entity and then averaging these percentages across all entities, ensuring that every entity is equally

represented regardless of its scale. While the global percentage method highlights absolute contributions, the average per-entity method captures relative contributions within each entity, making it more robust to variations in scale and less sensitive to outliers since every data point contributes equally to the final result.

From Table 11 and Table 12, generally, the usage of pure literal properties is lower than of object and ID. This aligns with the principle of knowledge graph to use URI for internal connections and ID for external resources.

We can say that object properties and literal properties combined represent wealth sourced internally from Wikidata, whereas ID properties originate from external sources. In general, more than half of the wealth is attributed to internal resources.

Table 11: Contribution of Property Type to Knowledge Wealth Using Bag of Properties

Class Name	% Object Bag	% Literal Bag	% ID Bag	% Object + Literal Bag
American researcher	59.55 / 65.65	3.32 / 2.8	37.12 / 31.55	62.88 / 68.45
American singer	40.44 / 48.42	7.47 / 8.48	52.09 / 43.1	47.91 / 56.9
Badminton player	79.34 / 78.96	10.14 / 11.68	10.52 / 9.36	89.48 / 90.64
Computer scientist	59.36 / 65.93	4.19 / 3.71	36.44 / 30.36	63.56 / 69.64
Historical painting	66.35 / 66.46	25.48 / 25.85	8.17 / 7.69	91.83 / 92.31
Memorial	62.82 / 64.16	21.0 / 20.4	16.18 / 15.44	83.82 / 84.56
Sci-fi book	62.96 / 62.62	14.44 / 15.78	22.59 / 21.6	77.41 / 78.4
University	37.59 / 44.79	18.11 / 18.31	44.3 / 36.9	55.7 / 63.1

This table shows the contribution of each property type of 8 Wikidata classes based on the notion of bag of properties. Each record has 2 values separated by "/". The first value is the contribution rate when calculated with global percentage method, and the second one is when using the average contribution per individual entity

Table 12: Contribution of Property Type to Knowledge Wealth Using Set of Properties

Class Name	% Object Set	% Literal Set	% ID Set	% Object + Literal Set
American researcher	51.25 / 59.69	3.97 / 3.29	44.78 / 37.02	55.22 / 62.98
American singer	33.91 / 43.58	8.1 / 9.18	57.99 / 47.24	42.01 / 52.76
Badminton player	71.75 / 74.13	13.79 / 13.9	14.46 / 11.97	85.54 / 88.03
Computer scientist	50.53 / 60.54	4.98 / 4.27	44.49 / 35.19	55.51 / 64.81
Historical painting	60.57 / 61.91	28.88 / 28.61	10.55 / 9.48	89.45 / 90.52
Memorial	60.06 / 62.01	22.41 / 21.59	17.54 / 16.4	82.46 / 83.6
Sci-fi book	59.36 / 61.91	13.86 / 14.59	26.78 / 23.5	73.22 / 76.5
University	33.93 / 42.74	17.67 / 18.32	48.4 / 38.95	51.6 / 61.05

This table shows the contribution of each property type of 8 Wikidata classes based on the notion of set of properties. Each record has 2 values separated by "/". The first value is the contribution rate when calculated with global percentage method, and the second one is when using the average contribution per individual entity

Typically, an ID property has exactly one associated value because it serves as a unique identifier in external resources, preventing ambiguity. In contrast, object properties can have multiple values. For example, the computer scientist *Noam Chomsky* (Q9049) has an ID property, *VIAF cluster ID* (P214), with exactly one value: *89803084*. At the same time, he has an object

property, *work location* (P937), with four values: *Pennsylvania* (Q1400), *Cambridge* (Q350), *Tucson* (Q18575), and *Massachusetts* (Q771). When using the bag of properties, *VIAF cluster ID* (P214) contributes 1 to the total wealth, while *work location* (P937) contributes 4. Under the set of properties, both properties contribute equally, each accounting for 1. This explains why the percentage contribution of object properties to wealth is smaller when using the set of properties compared to the bag of properties. Since the overall property count is lower in the set of properties, the percentage contribution of ID properties to wealth is correspondingly higher than in the bag of properties.

5. Discussion

Generality of Framework. The framework that is proposed in this study is applicable to any kind of knowledge graph. However, the library that we built for experimentation is specifically for Wikidata, because the query service, structure, and response is very unique for each knowledge graph. If further experimentation is to be conducted for other knowledge graph, then the library needs to be modified.

Scalability. In our Wikidata use case, we analyzed both human-related and non-human classes based on predefined filters. While analyzing the entire human class (i.e., all entities s that satisfy the triple $(s, instanceOf(P31), human(Q5))$) would provide valuable insights, we opted for smaller classes due to practical constraints. The primary limitation comes from the Wikidata Query Service itself, particularly its restrictions on query execution time.

The largest class we successfully retrieved in a single query was "American politicians" which contained 96,507 entities and required 34 seconds for query execution. However, when attempting to query significantly larger classes—such as all politicians without filtering by nationality—we encountered system limitations, including timeouts and out-of-memory errors.

For Wikidata or other KGs developer, a potential solution to overcome these limitations is to host them on a more reliable infrastructure with optimized indexing and caching mechanisms. This would allow for faster query execution and improved scalability for large-scale data retrieval.

That said, it is important to note that the analysis performed in this study is non-transactional, meaning it does not require real-time execution. Instead, it is designed for periodic analyses (e.g., monthly or quarterly), where latency constraints and query limitations are less critical. Given this characteristic, our current approach remains viable despite the performance limitations of the Wikidata Query Service.

Issue of Incoming Link. Wikidata is an entity-centric knowledge graph which means in the editathon efforts, the subject s is always be the subject of interest and the starting point to be edited by contributors. It is very unlikely that the opposite approach is done, that is, an object (o) is given and a pair subject and property (s, p) is to be searched. Not only from the contributors, the platform itself does not support the later view. This explains the phenomenon that we observed in Subsection 4.2 regarding outgoing and incoming link. With existing subject-centric point of view, the number of new outgoing link introduced to the knowledge graph will grow

in a much higher rate as opposed to incoming link. As a result, incoming link will be very rare, or even only present in certain entites.

Weight of Property Set of properties might be more suitable if the main concern is the presence of properties, instead of the abundance of information it contains. We will take *William Shakespeare* (Q692) in Wikidata as an example. It is reasonable if property like *date of birth* (P569) to be treated using set of properties, but for a well-known playwright and poet, we shall expect the property *notable work* (P800) to incorporate all or most of his well-known works. If only few works registered in G despite he has dozens of works, then we might conclude the entity is poor. For this case, treating the property using the notion of set is not preferable because it will fail to capture the aforementioned poor condition, while blatantly using bag of properties might skew the wealth amount. Due to this, the notion of (non-)uniqueness can be extended to a weighted form. The weight is given independently for each property and can be defined in such a way that is most appropriate for the nature of the property.

Another phenomena that should also be considered is that an entity might have several occurrences of the same property, but this property might just be 'trivial'. This is similar to a document having a lot of 'the' or 'a' (stopwords). Conversely, an entity might just have a single occurrence of a property, but it is a non-trivial one. Perhaps, the property is a highly relevant one for the entity's class. For example, *time period* (P2348) for *William Shakespeare* (Q692) will be a highly relevant one for prominent poets. However, the property *sibling* (P3373) might not be too relevant for Shakespeare's career. One idea to handle such trivial is to regard those porperties using the notion of set of properties–thus we care only on its existence rather than its actual count–or, we can use threshold function to set a tolerance for how much of such trivial we allow.

Multiclass Entity and Misclassification. In data classification, entities may belong to multiple classes or be misclassified due to inconsistencies in data sources. For instance, the richest entity in the computer scientist class based on pure literal properties count, *Bruno Darcet* (Q71055086), has a total wealth of 54, which accounts from 1 value in *date of birth* (P569) and 53 values in *Elo rating* (P1087). The same case is also observed in the second richest entity, *Daniel José Queraltó* (Q23906907). From their profiles, we may say these 2 entities are not actually computer scientists, highlighting an occupational misclassification issue. A similar problem occurs with the classification of American researcher, where the wealthiest by pure literal count, *Alexander Julien* (Q61249179), is not a university researcher. In the case of badminton players, *Dirk Stikker* (Q194654), the richest individual in the class based on ID properties count, is more of a politician with badminton affiliations rather than a prominent player. These examples illustrate that an entity can belong to multiple classes, or be entirely misclassified. These classification inconsistencies reveal data quality issues, necessitating careful data querying, cleaning, and validation to ensure accuracy prior to analysis.

6. Conclusions

In this study, we introduce a novel framework for analyzing knowledge wealth in knowledge graphs, using Wikidata as a case study. By drawing an analogy between financial wealth and information richness, we develop the formal model of 3 notions of knowledge wealth based on property cardinality, property type, and link direction. Complementing this, we introduce an insight model that includes statistical summaries, Gini coefficients, and Lorenz curves, providing interpretable diagnostics of wealth distribution. We make the Python-based implementation library for both the formal model and the insight model available at https://github.com/nurulputri/paper-knowledge-wealth-framework.

We demonstrate how these formal definitions, when analyzed through the insight model, yield varying inequality levels, as reflected in measures such as the Gini coefficient and Lorenz curves. Our proposed metric—relative expectation ratio—offers a powerful and intuitive way to assess representation bias across different entity groups. Unlike raw entity counts or central tendency measures, this metric captures the distributional imbalances of information across quantiles, providing a more nuanced understanding of how knowledge is allocated within Wikidata.

Future work may explore several things in order to improve the knowledge wealth framework. At the moment, our evaluation is limited to Wikidata. However, the framework itself is designed to be general, and our Python library is openly available. It can be extended to other KGs by simply modifying the SPARQL endpoint and adapting the query structure, making it feasible to assess the framework's generalizability and to further explore the overall quality of other KGs. Another promising direction is to explore *property-weighted knowledge wealth*, where not all properties are treated equally. In the current framework, each property contributes uniformly to an entity's wealth, regardless of its importance or informativeness. Incorporating weights—based on factors such as semantic significance, frequency of use, or relevance to the class—could offer a more nuanced and realistic estimation of an entity's information richness. Last but not least, it is also of our interest to explore the use of machine learning algorithms to classify entities into "rich" and "poor" groups based on their knowledge wealth. This goes beyond simply sorting entities and selecting the top $x\%$ or bottom $y\%$, as different distribution shapes across classes may imply different thresholds or grouping patterns. Such clustering could assist contributors in identifying which segments require additional enrichment efforts, enabling more targeted and efficient improvements to KGs completeness.

Acknowledgments

TBD

References

[1] Wealth, N. meanings, etymology and more, https://www.oed.com/dictionary/wealth_n, n.d. Accessed: 2025-03-19.

[2] E. Saez, G. Zucman, Wealth Inequality in the United States since 1913: Evidence from Capitalized Income Tax Data, The Quarterly Journal of Economics 131 (2016). URL: https://doi.org/10.1093/qje/qjw004. doi:10.1093/qje/qjw004.

[3] N. Mihindukulasooriya, DBLP to wikidata: Populating scholarly articles in wikidata, in: L. Etcheverry, V. L. Garcia, F. Osborne, R. Pernisch (Eds.), Proceedings of the ISWC 2024 Posters, Demos and Industry Tracks: From Novel Ideas to Industrial Practice co-located with 23nd International Semantic Web Conference (ISWC 2024), Hanover, Maryland, USA, November 11-15, 2024, volume 3828 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024. URL: https://ceur-ws.org/Vol-3828/paper37.pdf.

[4] N. Mihindukulasooriya, S. Tiwari, D. Dobriy, F. Å. Nielsen, T. R. Chhetri, A. Polleres, Scholarly wikidata: Population and exploration of conference data in wikidata using llms, in: M. Alam, M. Rospocher, M. van Erp, L. Hollink, G. A. Gesese (Eds.), Knowledge Engineering and Knowledge Management - 24th International Conference, EKAW 2024, Amsterdam, The Netherlands, November 26-28, 2024, Proceedings, volume 15370 of *Lecture Notes in Computer Science*, Springer, 2024, pp. 243–259. URL: https://doi.org/10.1007/978-3-031-77792-9_15. doi:10.1007/978-3-031-77792-9_15.

[5] J. Bolinches, D. Garijo, Saltbot: Linking software and articles in wikidata, in: L. Kaffee, S. Razniewski, K. Alghamdi, H. Arnaout (Eds.), Proceedings of the Wikidata Workshop 2023 co-located with 22nd International Semantic Web Conference (ISWC 2023), Athens, Greece, November 13, 2023, volume 3640 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023. URL: https://ceur-ws.org/Vol-3640/paper12.pdf.

[6] R. Y. Wang, D. M. Strong, Beyond accuracy: What data quality means to data consumers, J. Manag. Inf. Syst. 12 (1996) 5–33. URL: https://doi.org/10.1080/07421222.1996.11518099. doi:10.1080/07421222.1996.11518099.

[7] A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, S. Auer, Quality assessment for linked data: A survey, Semantic Web 7 (2016) 63–93. URL: https://doi.org/10.3233/SW-150175. doi:10.3233/SW-150175.

[8] A. Wisesa, F. Darari, A. Krisnadhi, W. Nutt, S. Razniewski, Wikidata completeness profiling using prowd, in: M. Kejriwal, P. A. Szekely, R. Troncy (Eds.), Proceedings of the 10th International Conference on Knowledge Capture, K-CAP 2019, Marina Del Rey, CA, USA, November 19-21, 2019, ACM, 2019, pp. 123–130. URL: https://doi.org/10.1145/3360901.3364425. doi:10.1145/3360901.3364425.

[9] S. Issa, O. Adekunle, F. Hamdi, S. S. Cherfi, M. Dumontier, A. Zaveri, Knowledge graph completeness: A systematic literature review, IEEE Access 9 (2021) 31322–31339. URL: https://doi.org/10.1109/ACCESS.2021.3056622. doi:10.1109/ACCESS.2021.3056622.

[10] M. J. Luthfi, F. Darari, A. C. Ashardian, Sock: SHACL on completeness knowledge, in: V. Svátek, V. A. Carriero, M. Poveda-Villalón, C. Kindermann, L. Zhou (Eds.), Proceedings of the 13th Workshop on Ontology Design and Patterns (WOP 2022) co-located with the 21th International Semantic Web Conference (ISWC 2022), Online, October 24, 2022, volume 3352 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022. URL: https://ceur-ws.org/Vol-3352/paper1.pdf.

[11] B. Xue, L. Zou, Knowledge graph quality management: A comprehensive survey, IEEE Trans. Knowl. Data Eng. 35 (2023) 4969–4988. URL: https://doi.org/10.1109/TKDE.2022.3150080. doi:10.1109/TKDE.2022.3150080.

[12] N. H. Ramadhana, F. Darari, P. O. H. Putra, W. Nutt, S. Razniewski, R. I. Akbar, Usercentered design for knowledge imbalance analysis: A case study of prowd, in: V. Ivanova, P. Lambrix, C. Pesquita, V. Wiens (Eds.), Proceedings of the Fifth International Workshop on Visualization and Interaction for Ontologies and Linked Data co-located with the 19th International Semantic Web Conference (ISWC 2020), Virtual Conference (originally planned in Athens, Greece), November 02, 2020, volume 2778 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020, pp. 14–27. URL: https://ceur-ws.org/Vol-2778/paper2.pdf.

[13] M. Ramadizsa, F. Darari, W. Nutt, S. Razniewski, Knowledge gap discovery: A case study of wikidata, in: L. Kaffee, S. Razniewski, K. Alghamdi, H. Arnaout (Eds.), Proceedings of the Wikidata Workshop 2023 co-located with 22nd International Semantic Web Conference (ISWC 2023), Athens, Greece, November 13, 2023, volume 3640 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023. URL: https://ceur-ws.org/Vol-3640/paper7.pdf.

[14] D. Abián, A. Meroño-Peñuela, E. Simperl, An analysis of content gaps versus user needs in the wikidata knowledge graph, in: U. Sattler, A. Hogan, C. M. Keet, V. Presutti, J. P. A. Almeida, H. Takeda, P. Monnin, G. Pirrò, C. d'Amato (Eds.), The Semantic Web - ISWC 2022 - 21st International Semantic Web Conference, Virtual Event, October 23-27, 2022, Proceedings, volume 13489 of *Lecture Notes in Computer Science*, Springer, 2022, pp. 354–374. URL: https://doi.org/10.1007/978-3-031-19433-7_21. doi:10.1007/978-3-031-19433-7_21.

[15] R. Cyganiak, D. Wood, M. Lanthaler, RDF 1.1 Concepts and Abstract Syntax, https://www.w3.org/TR/rdf11-concepts/, 2014. W3C Recommendation.

[16] J. Foster, A. Sean, On Economic Inequality, Oxford University Press, 1997.

[17] N. C. Kakwani, Applications of Lorenz Curves in Economic Analysis, Econometrica 45 (1977) 719–728. URL: https://doi.org/10.2307/1911684.