# SoCK Evaluation

See in depth the results of the completeness evaluation on **Wikidata** and **DBpedia**

# Schema Completeness

The schema completeness pattern checks the existence of properties from entities of a specific class that should exist. In the experiment, we validated 480,891 entities with 1,106 instances of schema completeness pattern. We used three approaches to create an instance of the completeness schema pattern: automatic, ontology, and statistics. The validation process in the experiment involved entities on Wikidata and DBpedia.

We validated over 469,891 Wikidata entities through an automatic approach using 1,095 SHACL shapes as instances of schema completeness patterns code "SC1". Entities checked are from classes that have the "properties for this type" (P1963) property on Wikidata. Figure 1 shows the validation result of the schema completeness pattern with an automatic approach on Wikidata aggregately. Based on the result, the completeness of Wikidata entities in schema completeness varies. However, the most completeness values are in the range of 0 to 0.05. The result shows that many entities still do not have the necessary properties according to the properties defined in "properties for this type". In addition, the completeness value with a range of 0.95 to 1 became the second highest. So, in general, the majority of entities on Wikidata are either complete or incomplete.

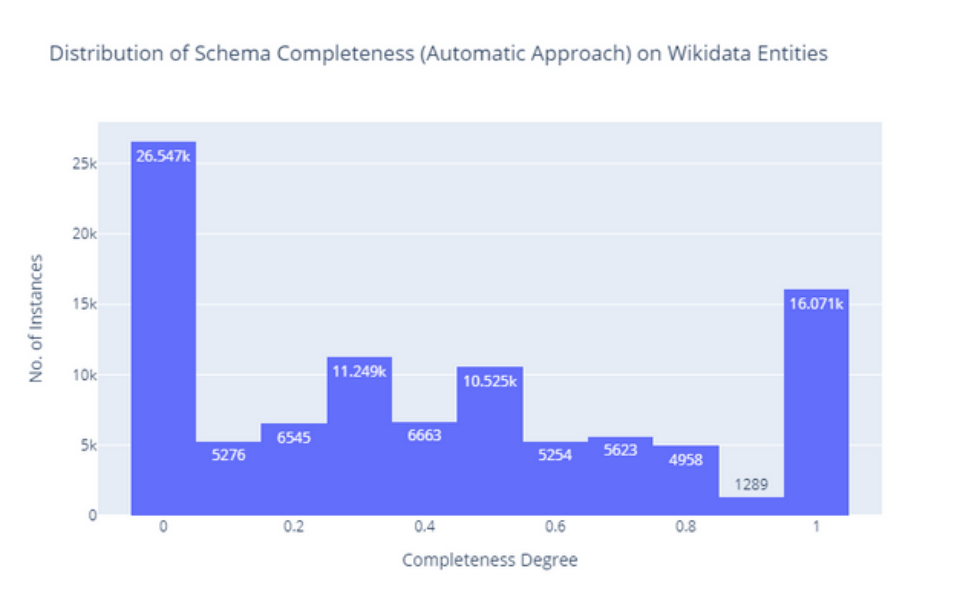


**Figure 1** The validation result of the completeness of entities on Wikidata with an instance of the schema completeness pattern and an automatic approach

With a statistical approach, we validated 6,000 entities using six SHACL shapes as instances of the completeness schema pattern code "SC1". The completeness context checked in the validation experiment is entities from the country, person, activity, film, museum, and university classes. Figure 2 shows the results of the schema completeness pattern validation with a statistical approach on Wikidata. The results indicate that entities on DBpedia with property selection using a statistical approach have a good average completeness value. 69.4% of the total validation entities have a value above 0.65.

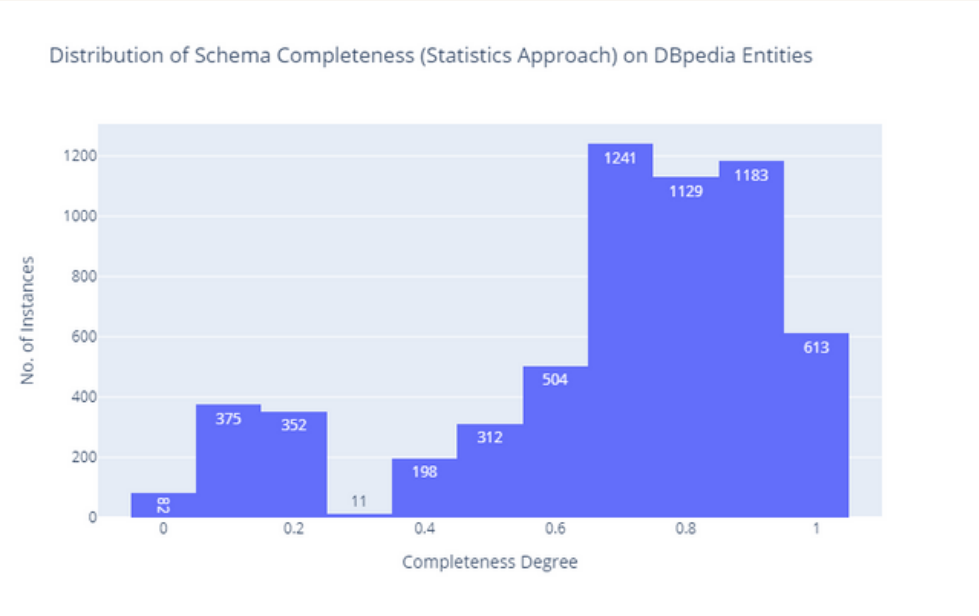


**Figure 2** The validation result of the completeness of entities on DBpedia with instances of the schema completeness pattern and a statistical approach

# Property Completeness

The property completeness pattern checks the specific properties of an entity. We experimented with this completeness pattern with validation on Wikidata. Experiments used instances of two approaches to creating an instance of the property completeness pattern, namely the automatic and spreadsheet approach. In this experiment, we validated the entity as many as the number of shapes created. This way followed the description of the property completeness that shape checks individual entities with sh:targetNode. The experiment involved 357,892 instances of the property completeness pattern to validate 357,892 knowledge graph entities.

In the automatic approach, we validated 357,749 entities using 357,749 SHACL shapes as the property completeness pattern instance with code "PC1". The completeness context that is checked in the validation experiment is the number of authors from the entities of the class "Scholarly Article" (Q13442814), the number of children from the entities of the class "Human" (Q5), the number of episodes from the entities of the class "Television Series Season" (Q3464665), the number of seasons from the entities of the class "Television Series" (Q5398426), and the number of participants from the entities of the class "Sport Season" (Q27020041). Figure 3 shows the validation result of the property completeness pattern with an automatic approach on Wikidata.

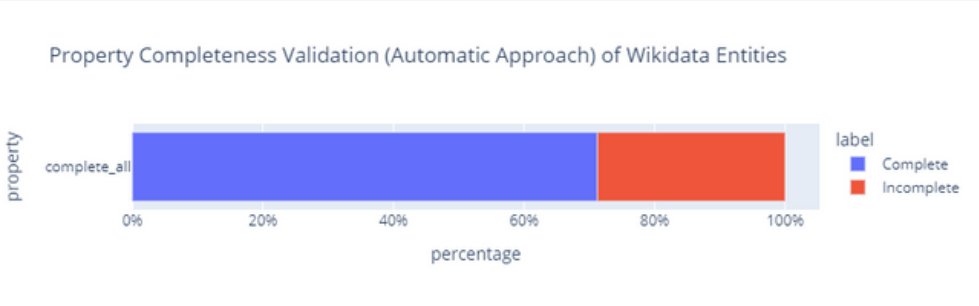


**Figure 3** The validation result of the completeness of entities on Wikidata with instances of the property completeness pattern and an automatic approach

In the spreadsheet approach, we validated 143 entities using 143 SHACL shapes as an instance of the property completeness pattern with code "PC1". The context of completeness examined in the validation experiment is the number of children of Indonesian actors, the number of cities and regencies from each province in Indonesia, the number of authors of scientific articles, the number of districts from each city and regency in Indonesia, and the number of seasons of TV series. In the context of the number of districts, the cities and districts examined are limited to East Java and Jambi. This decision aims to simplify the data collection process. Figure 4 shows the validation result of the property completeness pattern with the spreadsheet approach on Wikidata.

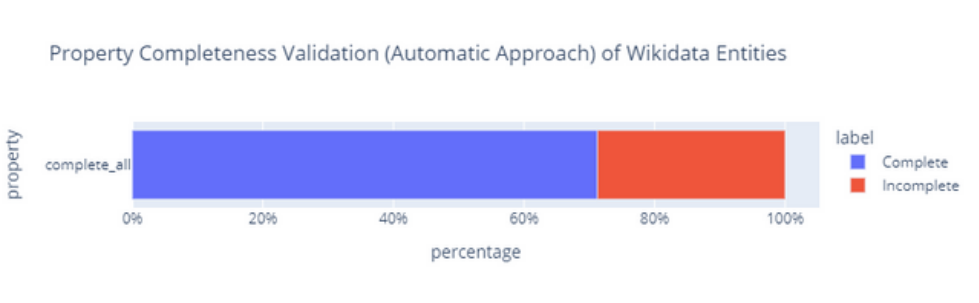


**Figure 4** The validation result of the completeness of entities on Wikidata with instances of the property completeness pattern and a spreadsheet approach

Based on Figures 3 and 4, the two approaches obtained completeness scores that were not much different, ranging from 60-70%. The completeness value taken is the average value for each validation experiment aggregately. This value is good enough, but it should be noted that there are still incomplete entities because they do not have as many property values as they should in reality.

# No-Value Completeness

The no-value completeness pattern checks the completeness of the data based on the absence of a property value in the knowledge graph entities. We conducted experiments with patterns made with an automatic approach using the information that existed in Wikidata. These patterns follow the no-value completeness pattern with the code "NVC1". The context of completeness checked in the validation experiment is the same as the one we have used for the completeness property. The check involved

- the number of authors from the entity class "Scholarly Article" (Q13442814),
- the number of children from the entity class "Human" (Q5),
- the number of seasons from the entity class "Television Series" (Q5398426), and
- the number of participants from the entity class "Sport Season" (Q27020041).

The number of validated entities is the same as the number of shapes used as checks, that is, 2,230 entities and 2,230 SHACL shapes. The shape will check each entity as a target with sh:targetNode.

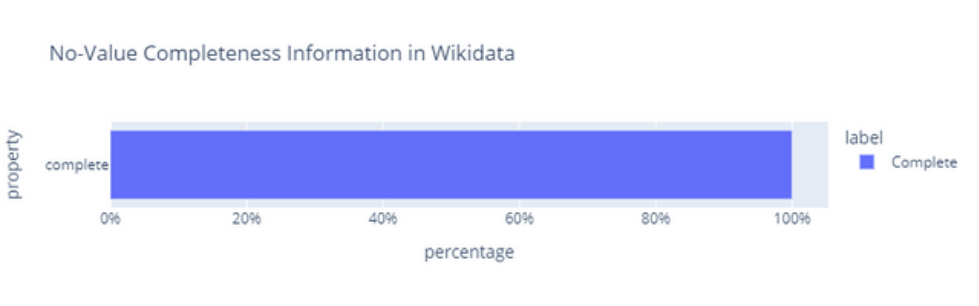


**Figure 5** The validation result of the completeness of entities on Wikidata with instances of the no-value completeness pattern

Figure 5 shows the validation result of the no-value completeness pattern on Wikidata entities. Based on Figure 5, the Wikidata entities are 100% complete with no-value completeness. It is because shape only checks for the absence of property values, so each entity is automatically said to be complete.

# Population Completeness

The population completeness pattern checks the completeness of the knowledge graph data from the existence of population entities in the real world. We experimented with this completeness pattern by validating DBpedia entities. The experiment involved eleven completeness pattern instances for the validation of eleven populations. The population completeness pattern we used is a pattern code "POC1". The pattern checks the total sum of the dct:subject property inversely. During the validation process, we examined the populations of cantons of Switzerland, continents, countries in Africa continent, countries in Europe continent, G20 nations, Indonesian active volcanoes, Indonesian legislative election events, NATO nations, oceans, and Summer Olympics events. Figure 6 shows validating the population completeness pattern on DBpedia entities.

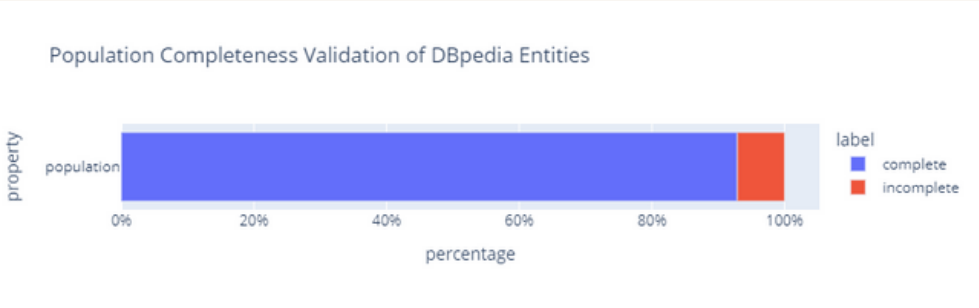


**Figure 6** The validation result of the completeness of entities on DBpedia with instances of the population completeness pattern

Based on Figure 6, DBpedia includes 92.8% of entities from the population used in the experiment aggregately. The result shows that DBpedia has covered almost all the population entities tested. In other words, checking the properties used, such as dct:subject, rdf:type, dbp:type, and dbo:type, can be considered appropriate as values that accommodate entities from the population in the form of DBpedia Category entities.

# Label and Description Completeness

The label and description completeness pattern checks the completeness of the label and description properties that people can read on each entity. We experimented with this completeness pattern by validating it on Wikidata and DBpedia entities. The experiment involved nine instances of the label and description completeness pattern for validation on a total of 69,045 entities.

On DBpedia, we validated 5,000 entities with five instances of the label and description completeness pattern code "LDC1". The pattern only checks for the existence of properties that provide labels and descriptions of an entity, such as rdfs:label, rdfs:comment, and dbo:abstract. During the validation process, we checked the entities from the mountain, politician, country, disease, and musical artists classes. Figure 7 shows the validation result of the label and description completeness on DBpedia.

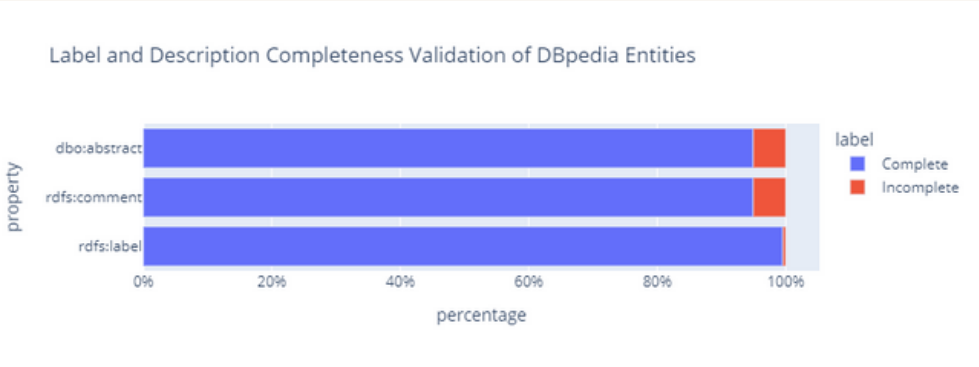


**Figure 7** The validation result of the completeness of entities on DBpedia with instances of the label and description completeness pattern

Based on Figure 7, the completeness value of the rdfs:label, rdfs:comment, and dbo:abstract properties are 99,96%, 94,96%, and 94,96%, respectively. Moreover, the completeness value for the rdfs:label property is almost 100%. The result shows that the entities on DBpedia have provided properties in the form of labels and descriptions so that humans can more easily understand these entities.

On Wikidata, we validated 64,045 entities with four instances of the label and description completeness pattern code "LDC2". The pattern checks label and description properties along with the language tags that match the context of the entity. The properties are rdfs:label, schema:description, and skos:altLabel. During the validation process, we examined the entities from the national heroes of Indonesia, South Korean music groups, American film, and Japanese Manga classes. Figure 8 shows the validation result of the label pattern and description completeness on Wikidata.

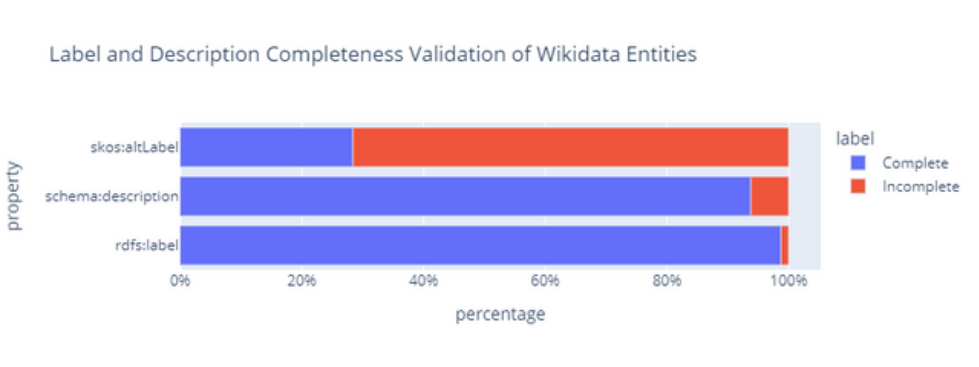


**Figure 8** The validation result of the completeness of entities on Wikidata with instances of the label and description completeness pattern

Based on Figure 8, the completeness value of the rdfs:label and schema:description properties are 98,72% dan 93,8%, respectively. These results show that the entity on Wikidata has fulfilled the intent of the label and description completeness pattern in providing a property with the value of a label and description of the entity. However, the completeness of the skos:altLabel as a property that provides the entity's alias is only 28.41% of the validation entity. This result indicates that the skos:altLabel property is incomplete because it is rarely used or there is no alias for the entity.

In addition, we also conducted second experiment validating DBpedia entities which are equivalent to Wikidata entities. The search for these entities uses information from the owl:sameAs property on each DBpedia entity. The validation process checks for the existence of specific entity labels and descriptions in English with pattern code "LDC2". It involved 5,000 entities from the country, mountain, film, hotel, and song classes. Figure 9 and 10 shows the label pattern and description completeness validation results on DBpedia and Wikidata, respectively.

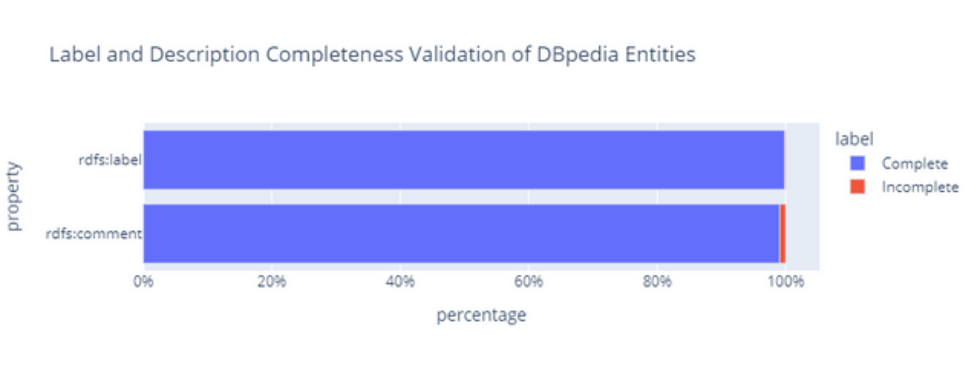


**Figure 9** The validation result of the completeness of entities on DBpedia with instances of the label and description completeness pattern on second experiment

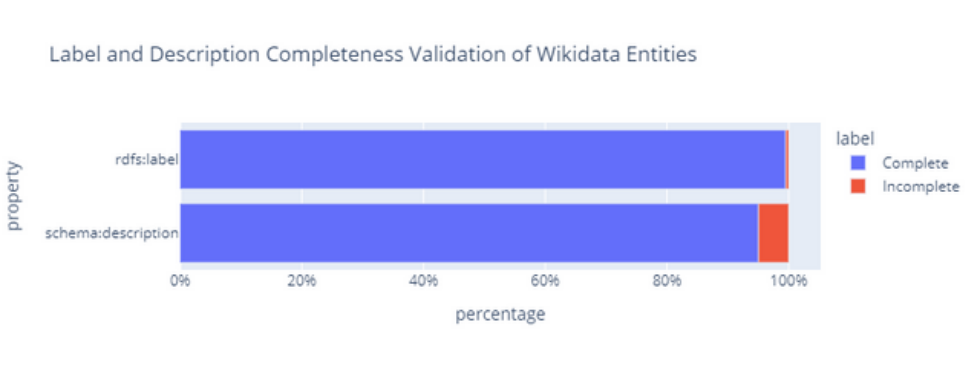


**Figure 10** The validation result of the completeness of entities on Wikidata with instances of the label and description completeness pattern on second experiment

Figures 9 and 10 show that DBpedia has higher completeness values for label and description properties. In detail, DBpedia has a completeness value of 99.86% for the label property (rdfs:label) and 99.06% for the description property (rdfs:comment). Meanwhile, Wikidata has a completeness value of 99.5% for the label property (rdfs:label) and 95.04% for the description property (schema:description).

We tried to find out more about the difference in the completeness value, especially for the description property, which has a difference of 4.02%. Then, we found out that the Wikidata entity used is not proper. The use of the Wikidata entity is taken from the property of owl:sameAs value in the DBpedia entity. In one DBpedia entity, it is probable to have more than one property value owl:sameAs in the form of a Wikidata entity. The selection of Wikidata entities is made simply using random sampling. However, only one Wikidata entity is proper for these values. For example, the entity “Leizhou Peninsula” on DBpedia has two values for property owl:sameAs, namely entities Q875712 (correct) and Q49333343 (incorrect).

# Interlinking Completeness

The interlinking completeness pattern checks information on entities' connectedness in a knowledge graph with equivalent entities in other knowledge graphs. We experimented this pattern with the validation process on DBpedia and Wikidata. The experiment used ten instances of the completeness pattern with a total validation of 9,194 entities.

In DBpedia, we validated 5,000 entities with five instances of the interlinking completeness pattern code "IC1". The pattern checked for the existence of the owl:sameAs, schema:sameAs, and skos:exactMatch properties. In particular, the check of owl:sameAs property more stringently looks for entity values from Wikidata and YAGO using completeness pattern code IC2. We checked the entities from the country, actor, island, museum, and hotel classes during the validation process. Figure 11 shows the validation result of the interlinking completeness pattern on DBpedia.

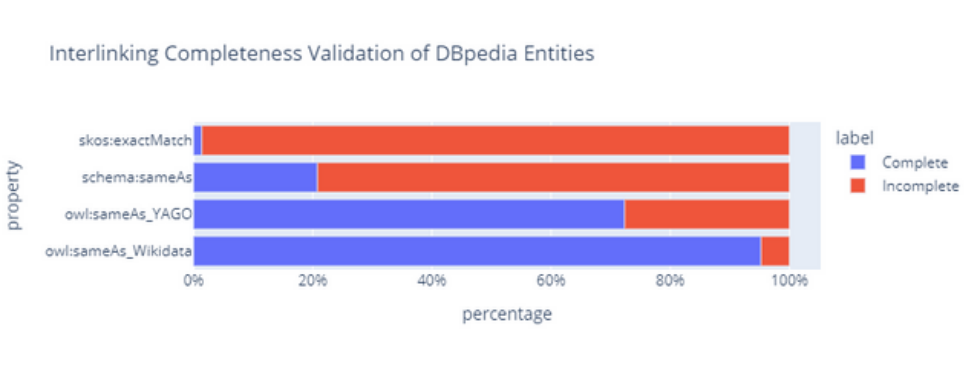


**Figure 11** The validation result of the completeness of entities on DBpedia with instances of the interlinking completeness pattern

Based on Figure 11, the completeness value of the owl:sameAs property is the highest, especially those with a value of Wikidata entities reaching 95.24%. The completeness value of the owl:sameAs property in the form of the YAGO entity is relatively high, around 72.36%. Meanwhile, the completeness value of the schema:sameAs property is about 20,80%. The completeness value of the skos:exactMatch property is only 1,40%. If we look at several validation entities, the value of the schema:sameAs property is only an entity from the Virtual International Authority File (VIAF). The result shows that the entities in DBpedia are not yet complete in covering external entities with schema:sameAs and skos:exactMatch properties. Another indication is the possibility that most validation entities do not have VIAF equivalent entities.

On Wikidata, we validated 4,194 entities with five instances of interlinking completeness pattern code "IC1". The pattern checked the existence of the property correlated to an external ID of another database or knowledge graph. During the validation, we examined the entities from the country (Q6256), hotel (Q27686), film (Q11424), singer-songwriter (Q488205), and university (Q3918) classes. Figure 12 shows the validation result of the interlinking completeness pattern on Wikidata aggregately.

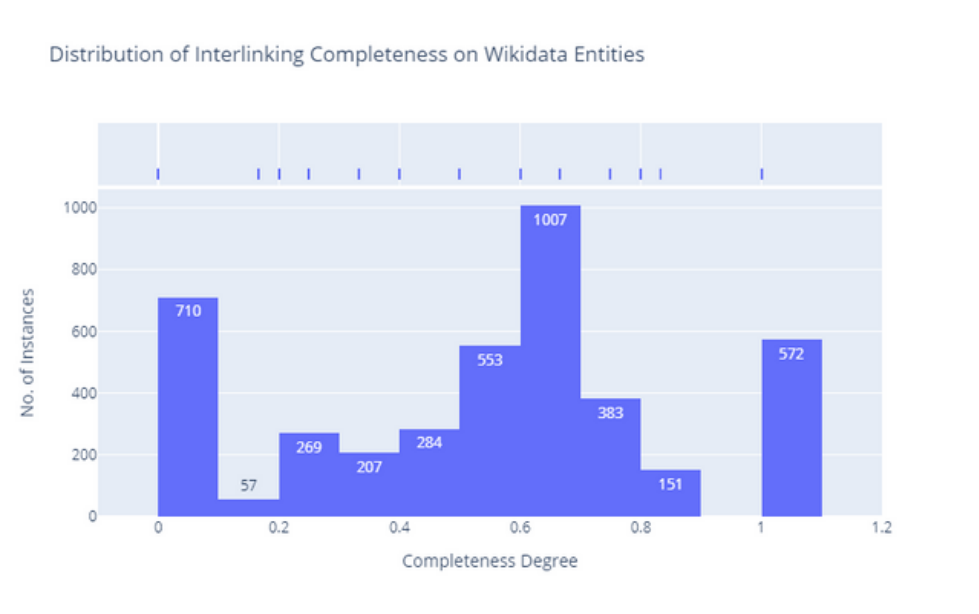


**Figure 12** The validation result of the completeness of entities on Wikidata with instances of the interlinking completeness pattern

Based on Figure 12, the validation result shows that the mean completeness values vary from 0 to 1. They are broadly divided into three groups: incomplete entities (<10%), semi-complete entities (20-90%), and complete entities (100%).