

---

## 6 Evaluation

This section aims to describe the evaluation performed at the end of the whole process over the final outcome of the iTelos methodology. More in detail, this section reports:

- The final Knowledge Graph information statistics (like, number of etypes and properties, number of entities for each etype, and so on).
- Knowledge layer evaluation: the results of the application of the evaluation metrics applied over the knowledge layer of the final KG.
- Data layer evaluation: the results of the application of the evaluation metrics applied over the data layer of the final KG.
- Query execution: the description of the competency queries executed over the final KG to test the KG's suitability to satisfy the project purpose.

After completing all the phases of the construction of the knowledge graph and obtaining the final knowledge graph, an essential and important step is the evaluation. iTelos provides various methods to assess the execution. Specifically, both the reusability (secondary objective) and the ability of the knowledge graph to satisfy the competency queries (primary objective) are assessed.

One possible metric to evaluate the knowledge layer that can be applied is the Coverage, defined as the extent to which a portion of knowledge is represented in the knowledge graph.

### Teleontology vs CQs

At the EType level:

$$\text{Cov}_E(\text{CQ}_E) = \frac{|\text{CQ}_E \cap T_E|}{\text{CQ}_E} = \frac{5}{5} = 1$$

Where  $\text{CQ}_E$  represents the number of etypes extracted from the CQs and  $T_E$  is the number of etypes of the Teleontology.

At the Property level:

$$\text{Cov}_P(\text{CQ}_P) = \frac{|\text{CQ}_P \cap T_P|}{\text{CQ}_P} = \frac{18}{19} = 0.94$$

specifically, we have 14 data properties and 4 object properties

Where  $\text{CQ}_P$  represents the number of properties extracted from the CQs and  $T_P$  is the number of properties of the Teleontology.

---

## Teleontology vs Reference Ontologies (ROs)

The OHMI ontology consists of 1025 entities and 133 properties, whereas the DOID ontology consists of 18,839 entities and 45 properties. The shared entities between them are 165. Then the evaluation is of the following. At the EType level:

$$\text{Cov}_E(\text{RO}_E) = \frac{|\text{RO}_E \cap \text{T}_E|}{\text{RO}_E} = \frac{3}{19699} = 0.0001$$

Where  $\text{RO}_E$  represents the number of etype extracted from the ROs and  $T_E$  is the number of properties of the Teleontology.

At the Property level:

$$\text{Cov}_p(\text{RO}_p) = \frac{|\text{RO}_p \cap \text{T}_p|}{\text{RO}_p} = \frac{18}{178} = 0.1$$

Where  $\text{RO}_p$  represents the number of properties extracted from the ROs and  $T_p$  is the number of properties of the Teleontology.

## Connectivity

On the other hand, evaluating the Data layer involves measuring the connectivity of the knowledge graph (KG). This analysis is performed across two key dimensions, both quantified using the connectivity metric:

- **Entity connectivity:** This measures the degree of interconnection between entities within the graph, emphasizing the richness and density of relationships between them.
- **Property connectivity:** This evaluates how thoroughly entities are linked to their corresponding properties, reflecting the completeness and detail of the information represented in the graph.

	Person	Cancer	Risk Factors	Microbioma	Human-Microbiome Interactions
Person	4	258	253	403480	53611
Cancer		2			
Risk Factors			3		
Microbioma				2	53611
Human-Microbiome Interactions					1

Table 8: The KG's evaluation - Data layer

# SPARQL

To usefully exploit our knowledge graph, SPARQL was used to query and answer the competency questions. Useful information about relative abundances of species in different scenarios were retrieved. To answer CQ-6 and CQ-4.

```
5 SELECT (MAX(?mean) AS ?total_Mean)
6 WHERE {
7   {
8     SELECT ?person (AVG(xsd:float(?relative_value)) AS ?mean)
9     WHERE {
10       # Get species of interest
11       ?species rdf:type etype:Microbiome .
12       ?interaction rdf:type etype:Human-Microbiome-Interaction .
13       ?interaction etype:has_Relative_Species_Abundance ?relative_value .
14       ?species etype:has_Interaction ?interaction .
15       ?person rdf:type etype:Person_UKC-36.
16       ?person etype:has_Spiecies ?species .
17       ?person etype:has_Interaction ?interaction .
18       ?person etype:has_Study_Condition ?status.
19       FILTER(str(?status)='CRC').
20     }
21     GROUP BY ?person
22  }
```

Table

Raw response

Pivot Table

Google Chart

Filter query results

Compact view

Hide row numbers

	total_Mean
1	"3.4462788" <sup>^^xsd:float</sup>

Figure 12: Query 1: Retrieving the maximum average of a person’s relative abundance with all the species

```

select DISTINCT ?person ?alcohol ?cig_level ?species ?occurrence ?mean_of_species
where{
{
select ?species (COUNT(?species) AS ?occurrence) (AVG(xsd:float(?relative_value))AS ?mean_of_species)
where {
#get species of interest
?species rdf:type etype:Microbiome .
?species etype:has_Name ?name .
#get event of interest (relative abundance)
?interaction rdf:type etype:Human-Microbiome-Interaction .
?interaction etype:has_Relative_Species_Abundance ?relative_value .
FILTER(xsd:float(?relative_value) > 3.44).
?species etype:has_Interaction ?interaction .
#get person of interest
?person rdf:type etype:Person_UKC-36.
?person etype:has_Species ?species .
?person etype:has_Interaction ?interaction .
?person etype:has_Study_Condition ?status.

?risk rdf:type etype:Risk_factors_KGE-QCB1-1 .
?person etype:has_Age ?Age .
?risk etype:has_Alcohol_Level ?alcohol .
?risk etype:has_Brinkman_Index ?cig_level .
FILTER(xsd:float(?cig_level) > 600).
FILTER(xsd:float(?alcohol) > 300).
?person etype:has_Risk_Factors ?risk .
?cancer rdf:type etype:Cancer_UKC-67961.
?cancer etype:has_Stage ?stage.
FILTER(str(?stage)!='NA').
?person etype:has_Diagnosis ?cancer.
}
GROUP BY ?species #?cig_level ?alcohol
}
?person rdf:type etype:Person_UKC-36.
?person rdf:type etype:Person_UKC-36.
?person etype:has_Species ?species .
?person etype:has_Interaction ?interaction .
?person etype:has_Study_Condition ?status.

?risk rdf:type etype:Risk_factors_KGE-QCB1-1 .
?person etype:has_Age ?Age .
?risk etype:has_Alcohol_Level ?alcohol .
?risk etype:has_Brinkman_Index ?cig_level .
FILTER(xsd:float(?cig_level) > 600).
FILTER(xsd:float(?alcohol) > 300).
?person etype:has_Risk_Factors ?risk .

?cancer rdf:type etype:Cancer_UKC-67961.
?cancer etype:has_Stage ?stage.
FILTER(str(?stage)='iv').
?person etype:has_Diagnosis ?cancer.
}
ORDER BY DESC (?occurrence)

```

Figure 13: Query 2: Answers CQ-6, CQ-4 and CQ-1 by retrieving the species with high relative abundance (greater than the mean of relative abundances for cancer) for cancer patients having stage "iv" cancer, high alcohol level and high brinkman index level. The species are ordered by their connectivity number with the cancer patients. Additionally the mean for the species is returned.

	person	alcohol	cig_level	species	occurrence_of_species	mean_of_species	stage
1	http://localhost:8080/source/SAMD00114750	"638.786"	"820"	http://localhost:8080/source/Bacteroides%20uniformis	"12"xsd:integer	"11.230103"xsd:float	"iv"
2	http://localhost:8080/source/SAMD00114803	"1899"	"630"	http://localhost:8080/source/Bacteroides%20uniformis	"12"xsd:integer	"11.230103"xsd:float	"iv"
3	http://localhost:8080/source/SAMD00114810	"348"	"640"	http://localhost:8080/source/Bacteroides%20uniformis	"12"xsd:integer	"11.230103"xsd:float	"iv"
4	http://localhost:8080/source/SAMD00114817	"360"	"780"	http://localhost:8080/source/Bacteroides%20uniformis	"12"xsd:integer	"11.230103"xsd:float	"iv"
5	http://localhost:8080/source/SAMD00114750	"638.786"	"820"	http://localhost:8080/source/Eubacterium%20rectale	"11"xsd:integer	"6.4065185"xsd:float	"iv"
6	http://localhost:8080/source/SAMD00114810	"348"	"640"	http://localhost:8080/source/Eubacterium%20rectale	"11"xsd:integer	"6.4065185"xsd:float	"iv"
7	http://localhost:8080/source/SAMD00114817	"360"	"780"	http://localhost:8080/source/Eubacterium%20rectale	"11"xsd:integer	"6.4065185"xsd:float	"iv"
8	http://localhost:8080/source/SAMD00114750	"638.786"	"820"	http://localhost:8080/source/Parabacteroides%20distasonis	"10"xsd:integer	"8.460612"xsd:float	"iv"

Figure 14: Results of query 2; much of bacteriodes are found to be linked with cancer patients that have high alcohol and cigarettes levels.

9	http://localhost:8080/source/SAMD00114803	"1899"	"630"	http://localhost:8080/source/Parabacteroides%20distasonis	"10"xsd:integer	"8.460612"xsd:float	"iv"
10	http://localhost:8080/source/SAMD00114810	"348"	"640"	http://localhost:8080/source/Parabacteroides%20distasonis	"10"xsd:integer	"8.460612"xsd:float	"iv"
11	http://localhost:8080/source/SAMD00114817	"360"	"780"	http://localhost:8080/source/Parabacteroides%20distasonis	"10"xsd:integer	"8.460612"xsd:float	"iv"
12	http://localhost:8080/source/SAMD00114750	"638.786"	"820"	http://localhost:8080/source/Prevotella%20copri	"10"xsd:integer	"37.641094"xsd:float	"iv"
13	http://localhost:8080/source/SAMD00114803	"1899"	"630"	http://localhost:8080/source/Prevotella%20copri	"10"xsd:integer	"37.641094"xsd:float	"iv"
14	http://localhost:8080/source/SAMD00114750	"638.786"	"820"	http://localhost:8080/source/Bacteroides%20stercoris	"9"xsd:integer	"9.614393"xsd:float	"iv"
15	http://localhost:8080/source/SAMD00114803	"1899"	"630"	http://localhost:8080/source/Bacteroides%20stercoris	"9"xsd:integer	"9.614393"xsd:float	"iv"

Figure 15: Extension of results for query 2



---

```

PREFIX etype: <http://knowdive.disi.unitn.it/etype#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

select DISTINCT ?person ?species ?relative_value ?stage#(AVG(xsd:float(?relative_value)) AS ?mean)
where {
  {select ?name
   where{
     #get species of interest
     ?species rdf:type etype:Microbiome .
     ?species etype:has_Name ?name .
     #get event of interest (relative abundance)
     ?interaction rdf:type etype:Human-Microbiome-Interaction .
     ?interaction etype:has_Relative_Species_Abundance ?relative_value .
     FILTER(xsd:float(?relative_value) > 1 && xsd:float(?relative_value) <5).
     ?species etype:has_Interaction ?interaction .
     #get person of interest
     ?person rdf:type etype:Person_UKC-36.
     ?person etype:has_Species ?species .
     ?person etype:has_Interaction ?interaction .
     ?person etype:has_Study_Condition ?status.
     FILTER(str(?status)!='CRC').
   }}
  ?species rdf:type etype:Microbiome .
  ?species etype:has_Name ?name .
  #FILTER(?name !='Prevotella copri' ).
  #get event of interest (relative abundance)
  ?interaction rdf:type etype:Human-Microbiome-Interaction .
  ?interaction etype:has_Relative_Species_Abundance ?relative_value .
  FILTER(xsd:float(?relative_value) > 10).
  ?species etype:has_Interaction ?interaction .
  #get person of interest
  ?person rdf:type etype:Person_UKC-36.
  ?person etype:has_Species ?species .
  ?person etype:has_Interaction ?interaction .
  ?person etype:has_Study_Condition ?status.
  FILTER(str(?status)='CRC').
  # ?cancer rdf:type etype:Cancer_UKC-67961.
  #?cancer etype:has_Stage ?stage.
  #?person etype:has_Diagnosis ?cancer
}
#GROUP BY (?person)

```

---

Figure 16: Query 3

Figure 16 answers general CQs and CQ-9 by identifying species associated with cancer. It highlights species that show a high relative abundance in individuals with cancer and a low abundance in healthy individuals.

	person		species		relative_value	
1	http://localhost:8080/source/SAMD00164889		http://localhost:8080/source/Bacteroides%20uniformis		"10.04378"	
2	http://localhost:8080/source/SAMD00114811		http://localhost:8080/source/Prevotella%20sp%20CAG5226		"10.12263"	
3	http://localhost:8080/source/SAMD00115010		http://localhost:8080/source/Prevotella%20sp%20CAG520		"10.12896"	
4	http://localhost:8080/source/SAMD00114775		http://localhost:8080/source/Faecalibacterium%20prausnitzii		"10.13096"	
5	http://localhost:8080/source/SAMD00164867		http://localhost:8080/source/Bacteroides%20uniformis		"10.13915"	

Figure 17: Results of Query 3: Highlights that species with low relative abundance in Healthy individuals are highly expressed in individuals diagnosed with cancer.

PREFIX etype: <http://knowdive.disi.unitn.it/etype#>

PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

```

SELECT ?person ?species ?relative_value ?cig ?bmi ?status
WHERE {
    # Get species of interest
    ?species rdf:type etype:Microbiome .
    ?species etype:has_Name 'Helicobacter pylori'.
    ?interaction rdf:type etype:Human-Microbiome-Interaction .
    ?interaction etype:has_Relative_Species_Abundance ?relative_value .
    ?species etype:has_Interaction ?interaction .
    ?person rdf:type etype:Person_UKC-36.
    ?person etype:has_Species ?species .
    ?person etype:has_Interaction ?interaction .
    ?person etype:has_Study_Condition ?status.
    #FILTER(str(?status)='CRC').
    ?risk rdf:type etype:Risk_factors_KGE-QCB1-1 .
    ?person etype:has_Age ?Age .
    ?risk etype:has_BMI ?bmi .
    ?risk etype:has_Brinkman_Index ?cig .
    ?person etype:has_Risk_Factors ?risk .
}

```

Figure 18: Query 4

Figure 18 answers CQ-8 by retrieving the individuals whose microbiome includes *Helicobacter pylori*.

	person	species	relative_value	cig	bmi	status
1	http://localhost:8080/ source/SAMD00114899	http://localhost:8080/ source/ Helicobacter%20pylori	'0.00157'	'0'	'22.18934911'	'control'
2	http://localhost:8080/ source/SAMD00164772	http://localhost:8080/ source/ Helicobacter%20pylori	'0.00337'	'570'	'25.40281608'	'adenoma'
3	http://localhost:8080/ source/SAMD00164834	http://localhost:8080/ source/ Helicobacter%20pylori	'0.01394'	'360'	'22.14532872'	'CRC'
4	http://localhost:8080/ source/SAMD00164893	http://localhost:8080/ source/ Helicobacter%20pylori	'0.00399'	'0'	'18.7961895'	'adenoma'

Figure 19: Shows that out of the four persons that contain the Helicobacter pylori one person with cancer has the most relative abundance among the others.

```

SELECT ?person ?status ?cig ?bmi
  (SUM(?ecoli_abundance) AS ?Escherichia_coli_abundance)
  (SUM(?gnavus_abundance) AS ?Ruminococcus_gnavus_abundance)
  (ABS(SUM(?ecoli_abundance) - SUM(?gnavus_abundance)) AS ?Abs_Diff_Ecoli_Gnavus)
  (ABS(SUM(?ecoli_abundance) - ?mean_gnavus_crc) AS ?Abs_Diff_CRC)
  (ABS(SUM(?ecoli_abundance) - ?mean_gnavus_noncrc) AS ?Abs_Diff_NonCRC)
WHERE {
  # Get species of interest
  ?species rdf:type etype:Microbiome .
  ?species etype:has_Name ?name .
  FILTER(?name IN ('Ruminococcus gnavus', 'Escherichia coli'))
  ?interaction rdf:type etype:Human-Microbiome-Interaction .
  ?interaction etype:has_Relative_Species_Abundance ?relative_value .
  ?species etype:has_Interaction ?interaction .
  # Ensure valid number format
  BIND(xsd:float(?relative_value) AS ?rel_value)
  # Assign values conditionally
  BIND(IF(?name = 'Escherichia coli', ?rel_value, 0) AS ?ecoli_abundance)
  BIND(IF(?name = 'Ruminococcus gnavus', ?rel_value, 0) AS ?gnavus_abundance)

  # Person details
  ?person rdf:type etype:Person_UKC-36.
  ?person etype:has_Species ?species .
  ?person etype:has_Interaction ?interaction .
  ?person etype:has_Study_Condition ?status.
  ?person etype:has_Gender 'female'.

  # Risk factors
  ?risk rdf:type etype:Risk_factors_KGE-QCBI-1 .
  ?person etype:has_Risk_Factors ?risk .
  ?risk etype:has_BMI ?bmi .
  ?risk etype:has_Brinkman_Index ?cig .
  # Compute mean Ruminococcus gnavus abundance per group
  {
    SELECT ?status (AVG(xsd:float(?relative_value)) AS ?mean_gnavus)
    WHERE {
      ?species rdf:type etype:Microbiome .
      ?species etype:has_Name 'Ruminococcus gnavus' .
      ?interaction rdf:type etype:Human-Microbiome-Interaction .
      ?interaction etype:has_Relative_Species_Abundance ?relative_value .
      ?species etype:has_Interaction ?interaction .
      ?person rdf:type etype:Person_UKC-36.
      ?person etype:has_Species ?species .
      ?person etype:has_Interaction ?interaction .
      ?person etype:has_Study_Condition ?status.
    }
    GROUP BY ?status
  }
  # Bind mean values separately for CRC and non-CRC
  BIND(IF(?status = 'CRC', ?mean_gnavus, 0) AS ?mean_gnavus_crc)
  BIND(IF(?status != 'CRC', ?mean_gnavus, 0) AS ?mean_gnavus_noncrc)
}
GROUP BY ?person ?status ?cig ?bmi ?mean_gnavus_crc ?mean_gnavus_noncrc

```

Figure 20: Query 5



	person	status	cig	gender	Escherichia_coll_...	Ruminococcus_g...	Abs_Diff_Ecoli_Gn...	Abs_Diff_CRC	Abs_Diff_NonCRC
1	http://localhost:8080/source/SAMD00114718	"control"	"1290"	"male"	"1.31309" <sup>xsd:float</sup>	"2.26875" <sup>xsd:float</sup>	"0.9556599855422974" <sup>xsd:float</sup>	"1.3130899667739868" <sup>xsd:float</sup>	"1.0476807355880737" <sup>xsd:float</sup>
2	http://localhost:8080/source/SAMD00114719	"control"	"100"	"male"	"9.1E-4" <sup>xsd:float</sup>	"0.07518" <sup>xsd:float</sup>	"0.07427000254392624" <sup>xsd:float</sup>	"9.10000002477318E-4" <sup>xsd:float</sup>	"2.35986065864563" <sup>xsd:float</sup>
3	http://localhost:8080/source/SAMD00114720	"control"	"1800"	"male"	"0.00525" <sup>xsd:float</sup>	"4.14455" <sup>xsd:float</sup>	"4.1392998695373535" <sup>xsd:float</sup>	"0.005249999929219484" <sup>xsd:float</sup>	"2.355520725250244" <sup>xsd:float</sup>
4	http://localhost:8080/source/SAMD00114721	"control"	"300"	"male"	"16.3262" <sup>xsd:float</sup>	"0.09351" <sup>xsd:float</sup>	"16.232690811157227" <sup>xsd:float</sup>	"16.326200485229492" <sup>xsd:float</sup>	"13.965429306030273" <sup>xsd:float</sup>
5	http://localhost:8080/source/SAMD00114730	"control"	"0"	"female"	"0" <sup>xsd:integer</sup>	"0.19638" <sup>xsd:float</sup>	"0.19638000428676605" <sup>xsd:float</sup>	"0" <sup>xsd:integer</sup>	"2.3607707023620605" <sup>xsd:float</sup>
6	http://localhost:8080/source/SAMD00114734	"control"	"900"	"male"	"0.02027" <sup>xsd:float</sup>	"0" <sup>xsd:integer</sup>	"0.020269999280571938" <sup>xsd:float</sup>	"0.020269999280571938" <sup>xsd:float</sup>	"2.340500593185425" <sup>xsd:float</sup>
7	http://localhost:8080/source/SAMD00114736	"control"	"0"	"female"	"0.6486" <sup>xsd:float</sup>	"9.47762" <sup>xsd:float</sup>	"8.829020500183105" <sup>xsd:float</sup>	"0.6485999822616577" <sup>xsd:float</sup>	"1.7121707201004028" <sup>xsd:float</sup>

Figure 21: Results of Query 5

Figure 20 illustrates the process of answering Query 5, which addresses CQ-10 through the following steps:

- Mean Relative Abundance Calculation:** Compute the mean relative abundance of one of the most abundant species, *Ruminococcus gnavus*, specifically for the 'CRC' group in both the 'Healthy' and 'CRC' populations.
- Individual Distance Measurement:** for each person calculate the distance between her *E. coli* relative abundance and their own relative abundance of *Ruminococcus gnavus* computed in step (A).
- Group Distance Comparison:** Measure the distance between each person's *E. coli* relative abundance and the mean relative abundances computed in step (A) for the 'Healthy' and 'CRC' groups.

This stepwise approach provides a structured method for addressing the query and comparing microbial profiles across the two groups.