A Linked Data Representation for Summary Statistics and Grouping Criteria

RPI IDEA/Tetherless World Constellation

James P. McCusker, Michel Dumontier, Shruthi Chari, Joanne S. Luciano, and Deborah L. McGuinness
Summary statistics across groups can be formalized as linked data using $owl:\text{Class}$-based sets, expressing aggregate values as attributes of those classes.

Class: $G(\text{case:TCGA-BRCA})$

SubClassOf: $sio:\text{human}$ and $sio:\text{has role}$ some ($sio:\text{subject role}$ and $sio:\text{in relation to}$ value $\text{case:TCGA-BRCA}$)

- $G(\text{case:TCGA-BRCA})$ has attribute $\text{count}$ $1098$
- $\text{mean}$ $1098$ has value $\text{age}$ $1098$
- $\text{maximal value}$ $32872$ has value $\text{minimal value}$ $2009$ has unit $\text{day}$
Example Data Schema – Genomic Data Commons Clinical Annotations
Defining Grouping Criteria (starting with Calvanese et al. 2008)

Class: GDC_Subject

EquivalentTo: sio:human

and sio:'has role' some (sio:'subject role' and sio:'in relation to' some sio:investigation)

SPARQL

select ?GDC_Subject WHERE {
  ?GDC_Subject a sio:SIO_000485; # human
  sio:SIO_000228 [ # has role
    a sio:SIO_000883; # study subject
    sio:SIO_000668 [ # in relation to
      a sio:SIO_000747 # investigation
      ]
    ].
}
Defining Grouping Criteria (starting with Calvanese et al. 2008)

\[ q(\bar{x}, \alpha(\bar{y})) \leftarrow \phi \]

where

Class: \( \bar{x} \)
SubClassOf: \( \phi \)

We will reserve \( \alpha(\bar{y}) \) for later.
Class: $\bar{x}$
SubClassOf: $\phi$ 
$\bar{x} = G(g_1, \ldots, g_n)$

Class: $G(\ g_1, \ldots, g_n)$
SubClassOf: $\phi$

Class: $G(?x)$
SubClassOf: sio:human
  and sio:'has role' some (sio:'subject role' and sio:'in relation to' value ?x)
Grouping Criteria as a SPARQL query

select ?GDC_Subject ?x where {
    ?GDC_Subject a sio:SIO_000485; # human
    sio:SIO_000228 [ # has role
        a sio:SIO_000883; # study subject
        sio:SIO_000668 ?x # in relation to
    ].
    ?x a sio:SIO_000747 # investigation
}

Class: G(?x)
SubClassOf: sio:human
    and sio:'has role' some (sio:'subject role'
        and sio:'in relation to' value ?x)
Class: \( G(\text{case:FM-AD}) \)
SubClassOf: sio:human
and sio:'has role' some (sio:'subject role'
and sio:'in relation to' value case:FM-AD)

Class: \( G(\text{case:TARGET-NBL}) \)
SubClassOf: sio:human and
sio:'has role' some (sio:'subject role'
and sio:'in relation to' value case:TARGET-NBL)

Class: \( G(?x) \)
SubClassOf: sio:human
and sio:'has role' some (sio:'subject role'
and sio:'in relation to' value ?x)
owl:Classes with property restriction definitions can be assigned URIs automatically based on the graph digest of that property restriction using RGDA1 or similar graph digest algorithms.

```java
graph = IsomorphicGraph()

graph = source_graph.query("describe ?restr where {
}""), initBindings={"G":my.Class}

digest = graph.graph_digest()

source_graph.add((
  my.Class,
  owl:equivClass,
  digest_prefix[digest]
))
```
WARNING! We will be discussing the use of OWL 2 puns.
TL;DR for **OWL 2 Punning**:

Statements asserted about a resource as an OWL Class **cannot be used to draw inferences** about the resource as an OWL Individual or vice-versa.
Expressing aggregate values relies on the Semanticscience Integrated Ontology, or an expressive equivalent.
First, if needed we reify non-SIO statements as attributes.
\[ \forall G, \alpha(\bar{y}) \exists A \in \alpha, Y \in \bar{y} \]
\[ \text{attr}(G, Y) \land \text{attr}(Y, A) \land \text{val}(A, \alpha(\bar{y})) \]

Finally, here’s what we do with \( \alpha(\bar{y}) \).
Here's what it looks like in practice.
We can query summary statistics from an RDF graph and put the results into its own graph.

We query the statistics out and display them using Vega-Lite.
Many thanks to:

Coauthors: Deborah, Michel, Joanne, and Shruthi
Others whom I’ve bothered about this:
    John Erickson, Patrice Seyed, and James Michaelis.