Aligning Diverse Data for a Personalized Health Knowledge Graph

Matthew Johnson, Meenu Ravi, Sabbir Rashid, Paulo Pinheiro, and Deborah L. McGuinness

Rensselaer Polytechnic Institute, Troy, NY, 12180, USA
Personal Health Knowledge Graph

Patient

Generates

Personal Health Knowledge Graph

Reviews

Doctor

Reduce Calories

Reviews

Eat X for Dinner

Edits

Agent
Integration Problem

Provides

Step, Weight, Medication, ...

Reviews
<table>
<thead>
<tr>
<th>Column</th>
<th>Attribute</th>
<th>attributeOf</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>id</td>
<td>hasco:originalID</td>
<td>??mother</td>
<td></td>
</tr>
<tr>
<td>age</td>
<td>sio:SIO_001013</td>
<td>??mother</td>
<td>sio:SIO_000428</td>
</tr>
<tr>
<td>gender</td>
<td>sio:SIO_010029</td>
<td>??child</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Virtual Column</th>
<th>Entity</th>
<th>Role</th>
<th>inRelationTo</th>
</tr>
</thead>
<tbody>
<tr>
<td>??mother</td>
<td>sio:SIO_000485</td>
<td>hhear:00502</td>
<td>??child</td>
</tr>
<tr>
<td>??child</td>
<td>sio:SIO_000485</td>
<td>hhear:00492</td>
<td>??mother</td>
</tr>
</tbody>
</table>

**Data Dictionary**

- **id**: Participant ID number
- **age**: Age of mother at enrollment
- **gender**: Child gender
- **LNAS_C1**: Natural log of drinking water arsenic at enrollment

**Knowledge Graph**

**Semantic Data Dictionary**
SDD-Editor

Data Dictionary
Column Description

SDD-Gen Suggestions

Biocorporal Search Results
Age in months of the participant at the time of screening.

AMR entity-related Types
(:ARG0, :ARG1, :ARG2, :ARG3, :conj, :domain, :topic, :location)

[0.6,0.2,...0.9 | 0,...0 | 0.2,0.4,...0.7 | 0,0,...0 | 0,0,...0 | 0,0,...0 | 0,...0 | 0.5,0.7,...1 | 0,0,...0]

GloVe Vectors

Autoencoder

Lower Dimensional Representation

- Lightweight Python library
- Generates knowledge graph fragments as nanopublications
- Applied to a variety of biomedical datasets:
  - NHANES
  - TCGA
  - CIViC
  - SEER
  - MIMIC-III
Data Integration: HADatAc

- Framework to align data across studies
- Ingests files and message streams
- User Interface:
  - Explore ontologies
  - Ingest datasets
  - Investigate knowledge graphs
Conclusion

- Semantic Data Dictionaries can be used to harmonize diverse datasets

- Our SDD-Editor allows users to semi-automatically generate Semantic Data Dictionaries, thus reducing user burden and broadening the class of users

- HADatAc and sdd2rdf can populate a PHKG using Semantic Data Dictionaries

Acknowledgements: This work is partially supported through AFRL 88ABW-2020-0991 and NIH 2U2CES026555-02