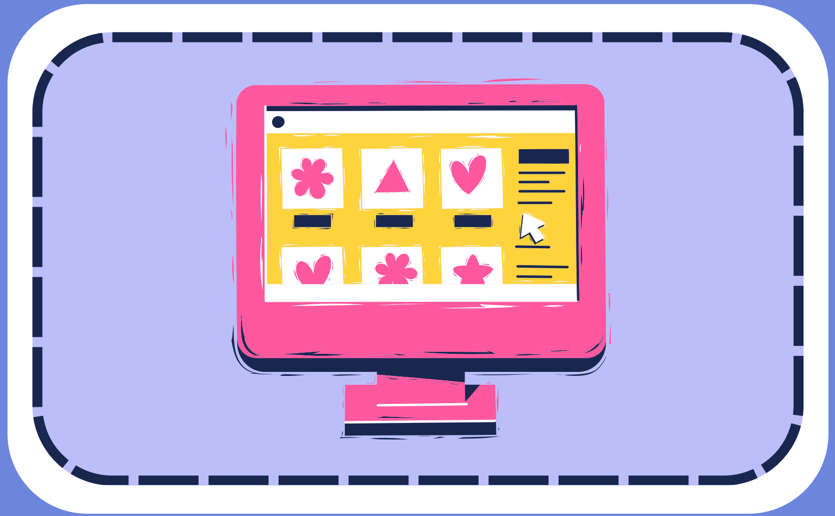
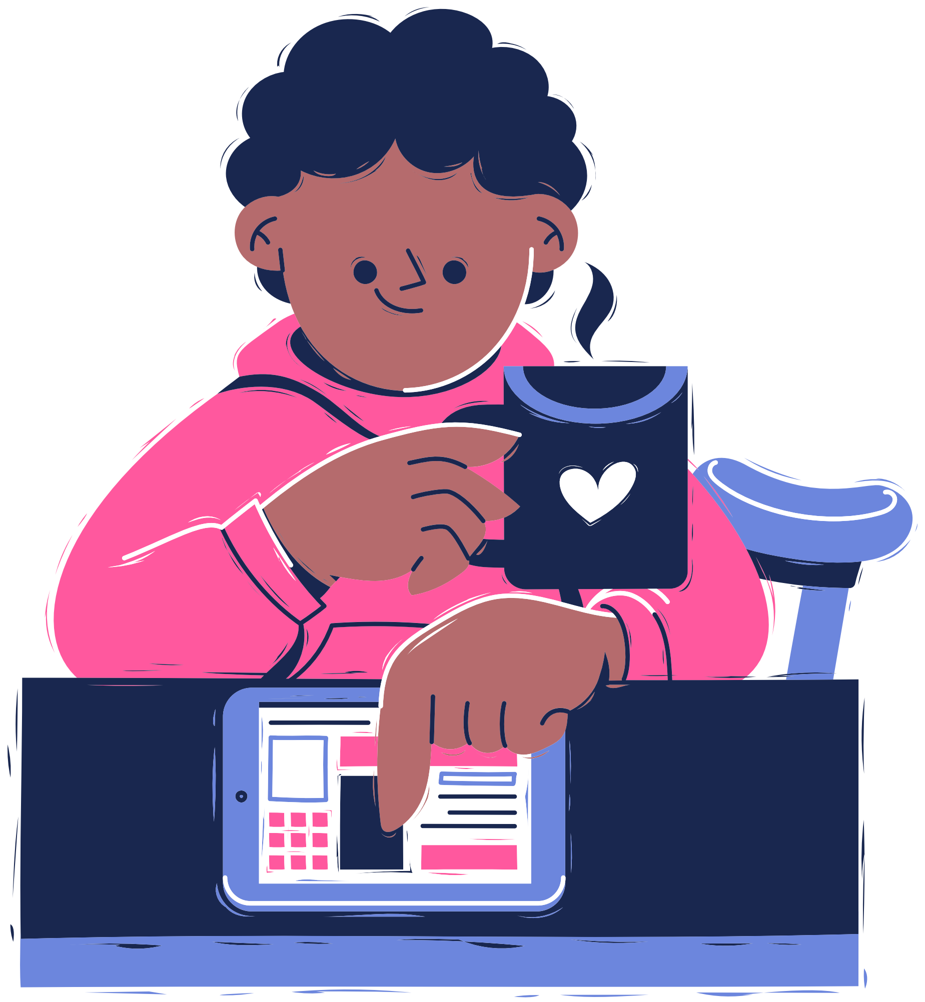


FAIR DATA ENGINEERING

FAIRification of Glassdoor employee data

OUTLINE



- FAIR Principles
- Dataset
- Reconciliation of Data
- RDF data triplification
- Semantic Data Model
- Metadata Schema
- FAIRness assessments

FAIR PRINCIPLES

FINDABLE

Ensure that the data is easily identifiable / discoverable using unique identifiers and metadata tags

ACCESSIBLE

Ensure that the data or metadata is easily accessible even if the original data is moved

INTEROPERABLE

Ensure that the data can be seamlessly used along with other data formats by using commonly used data types

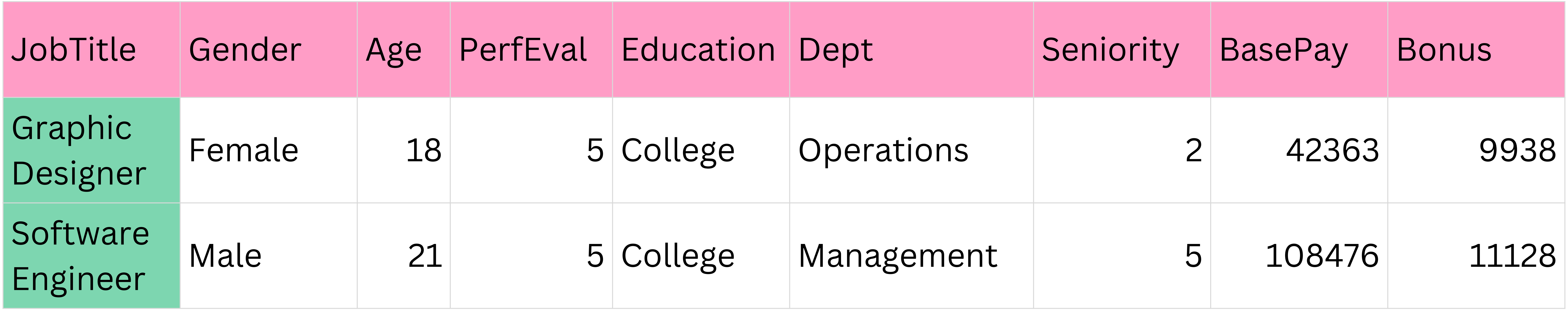
REUSABLE

Ensure that the data can be reused or updated by other data analysts and researchers

DATASET

The original dataset has been taken from Kaggle which consists of data generated from the Glassdoor website to observe the remuneration that employees receive with respect to various aspects.

sample data



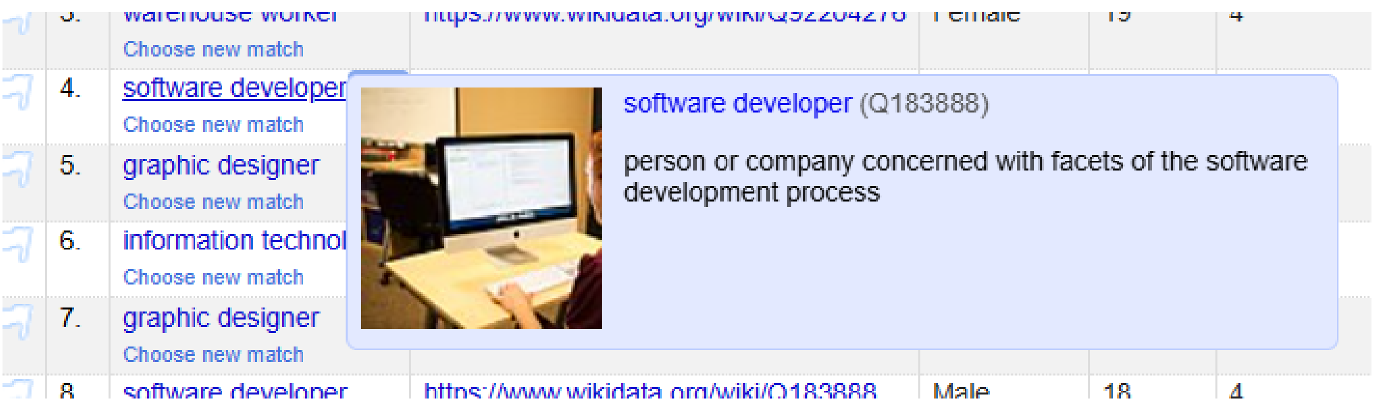
JobTitle	Gender	Age	PerfEval	Education	Dept	Seniority	BasePay	Bonus
Graphic Designer	Female	18	5	College	Operations	2	42363	9938
Software Engineer	Male	21	5	College	Management	5	108476	11128

RECONCILIATION OF DATA

Wikibase

The terms in the data were mapped to the closest definition available in wikibase

example



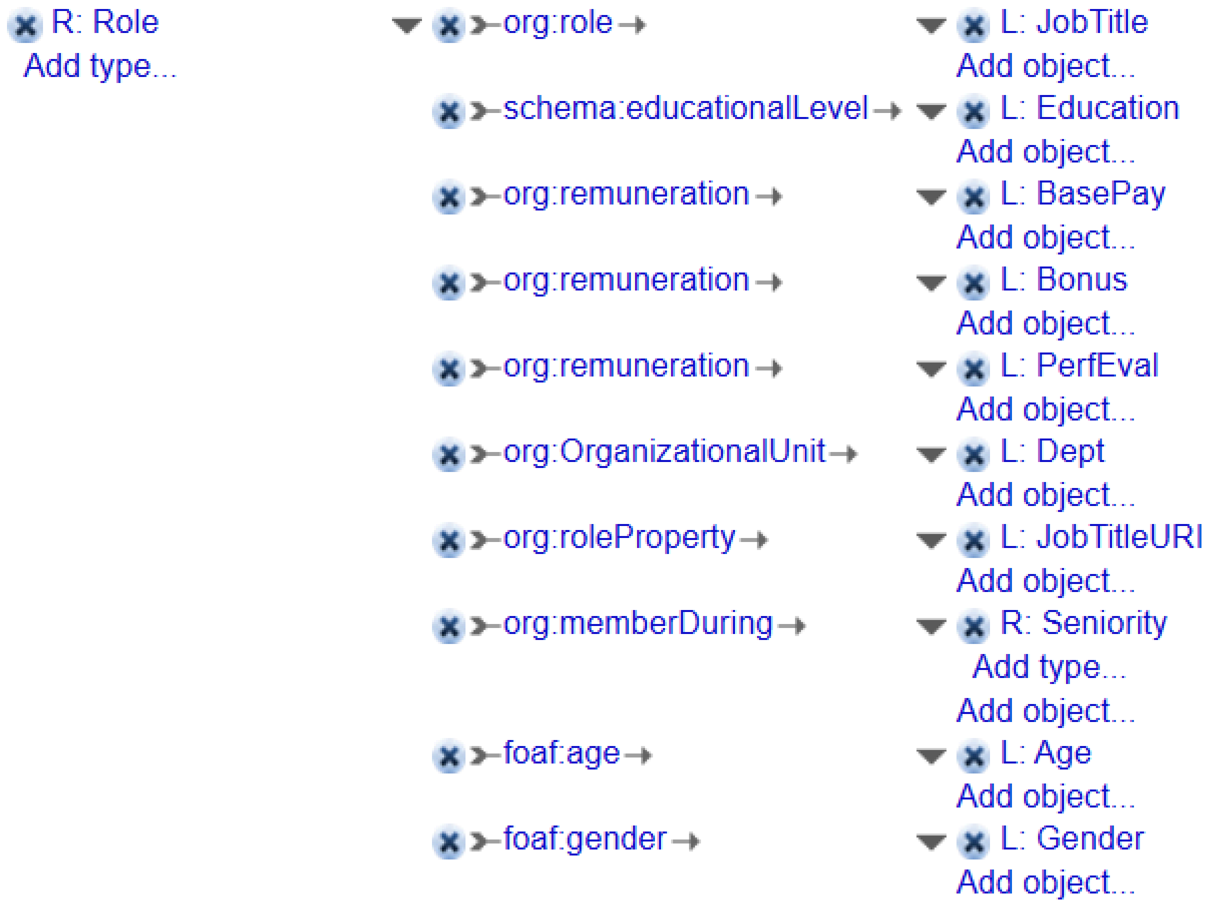
3.	warehouse worker Choose new match	https://www.wikidata.org/wiki/Q92204276	Female	19	4
4.	software developer Choose new match				
5.	graphic designer Choose new match				
6.	information technol Choose new match				
7.	graphic designer Choose new match				
8.	software developer	https://www.wikidata.org/wiki/Q183888	Male	18	4

software developer (Q183888)

person or company concerned with facets of the software development process

Ontology

The relationships between the entities were mapped to the closest relationship found in ontologies from w3.org



R: Role
Add type...

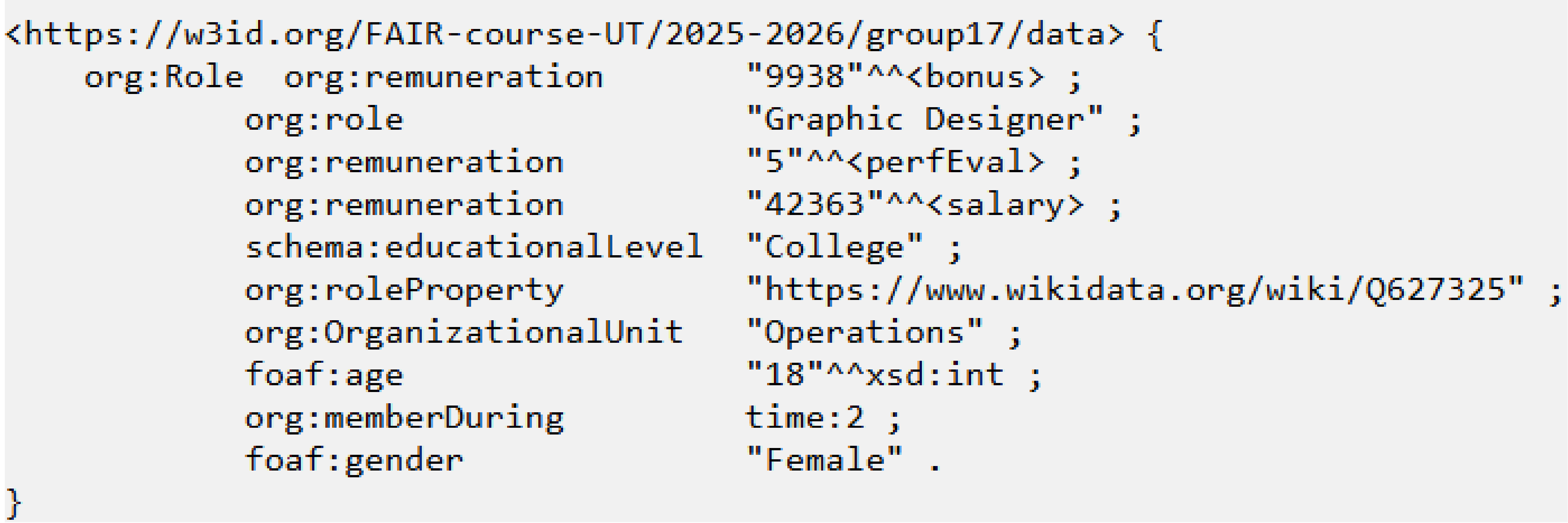
org:role→ L: JobTitle
Add object...
schema:educationalLevel→ L: Education
Add object...
org:remuneration→ L: BasePay
Add object...
org:remuneration→ L: Bonus
Add object...
org:remuneration→ L: PerfEval
Add object...
org:OrganizationalUnit→ L: Dept
Add object...
org:roleProperty→ L: JobTitleURI
Add object...
org:memberDuring→ R: Seniority
Add type...
Add object...
foaf:age→ L: Age
Add object...
foaf:gender→ L: Gender
Add object...

RDF DATA TRIPLIFICATION

Data was triplified and defined with a subject, predicate and object.

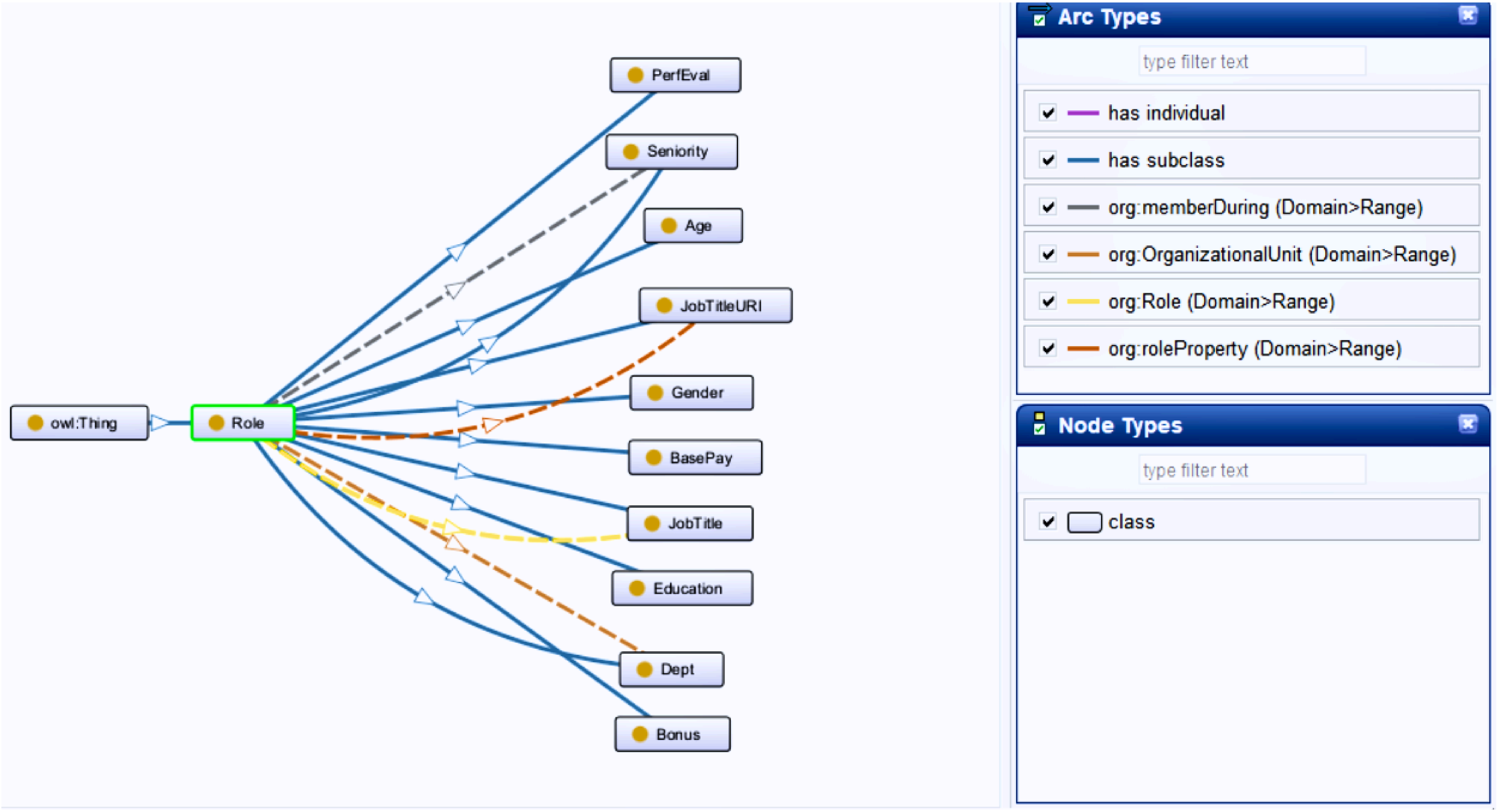
The prefix imports include :

- w3 organizational schema - org : <http://www.w3.org/ns/org#>
- schema : <http://schema.org/>



```
<https://w3id.org/FAIR-course-UT/2025-2026/group17/data> {
    org:Role  org:remuneration       "9938"^^<bonus> ;
            org:role                 "Graphic Designer" ;
            org:remuneration         "5"^^<perfEval> ;
            org:remuneration         "42363"^^<salary> ;
            schema:educationalLevel  "College" ;
            org:roleProperty         "https://www.wikidata.org/wiki/Q627325" ;
            org:OrganizationalUnit   "Operations" ;
            foaf:age                 "18"^^xsd:int ;
            org:memberDuring         time:2 ;
            foaf:gender              "Female" .
}
```

SEMANTIC DATA MODEL



Arc Types

type filter text

- ☑ has individual
- ☑ has subclass
- ☑ org:memberDuring (Domain>Range)
- ☑ org:OrganizationalUnit (Domain>Range)
- ☑ org:Role (Domain>Range)
- ☑ org:roleProperty (Domain>Range)

Node Types

type filter text

- ☑ class

METADATA SCHEMA

FAIR Data Point

Metadata for machines

Catalog

Default Fields :

- Title
- Description
- Date modified
- Date issued
- Theme Taxonomy
- Version
- Language

Dataset

Default Fields :

- Title
- Description
- Theme
- Keyword, etc.

Customised fields :

- Creator
- Last Updated
- Generated from
- Date created

Distribution

Default Fields :

- Title
- Description
- Date modified
- Date issued
- Download Link
- Language
- Media type

FAIRNESS ASSESSMENT

F1	The dataset and the entities have unique identifiers like URIs.
F2	Metadata are described for the data using FDP (titles, description, themes, keywords)
F3	Metadata files include the identifier of the data it describes.
F4	Data and Metadata are hosted in Github.
A1	Data and Metadata are hosted in a publicly available Github repository.

A2	Metadata defined in FDP persists even if the data moves.
I1	All metadata files are in .ttl format.
I2	The data and metadata use shared vocabularies (schema, foaf etc.)
I3	The data and metadata are linked to external vocabularies.
R1	Metadata have default FDP attributes and customised fileds for an enriched description

THANK YOU!