UNIVERSITY OF TRENTO - Italy

KNOW DIVE

Digital University Theses, Publications and Staff

Milena Atanasova

milena.atanasova@studenti.unitn.it

Knowledge Graph Engineering 2022/2023 UniTrento

December 2022

Project GitHub repository:
https://github.com/atanasova16/TrentinoDUniTPS

iTelos methodology

01. Purpose and Domain of Interest

02. Resources

03. Purpose Formalization

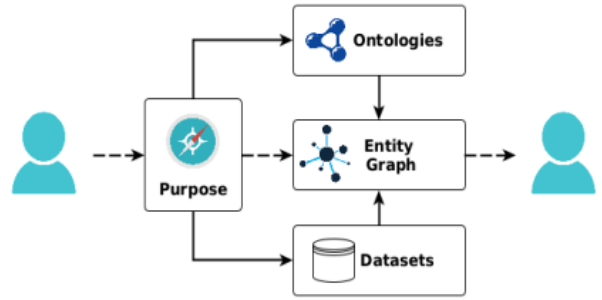
04. Inception

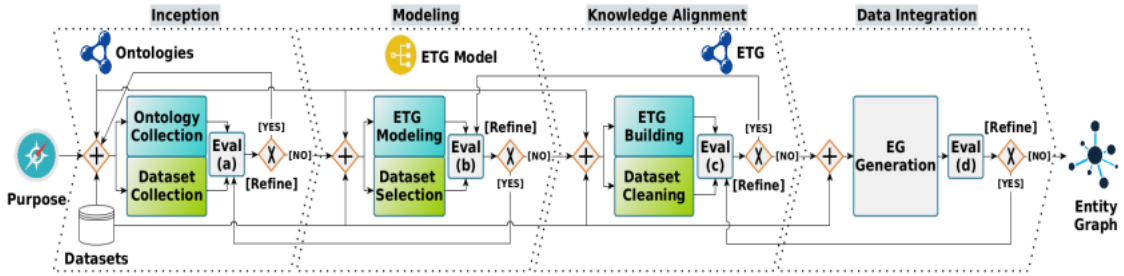
05. Informal modeling

06. Formal modeling

07. Outcome exploitation

08. Conclusions and open issues





Purpose and DoI

“A service which help the users to query and know about the different areas of academic and research interest being pursued at the University of Trento”

Academics at the University of Trento, including information mainly about research activities, between the years 2020 and 2022

Project Resources

Knowledge Sources	VIVO	FOAF	Schema.org	IAO	DCAT
Data Sources	Theses (Open Data Trentino)	Publications (Open Data Trentino)	Staff (Open Data Trentino)	Courses (Open Data Trentino)	Digital Portal of UniTn
Software Used	Python (Visual Studio Code)	Protégé	Karma data linker	KOS	Shapeness	GraphDB

Purpose Formalization

Purpose Formalization

Overview

Scenarios

- Identify a context in which the KG would be used
- Concern time (w.r.t. academic year), place

Personas

- Users who would benefit from the KG and would like to learn something from it
- 11 characters, different age, occupation, nationality

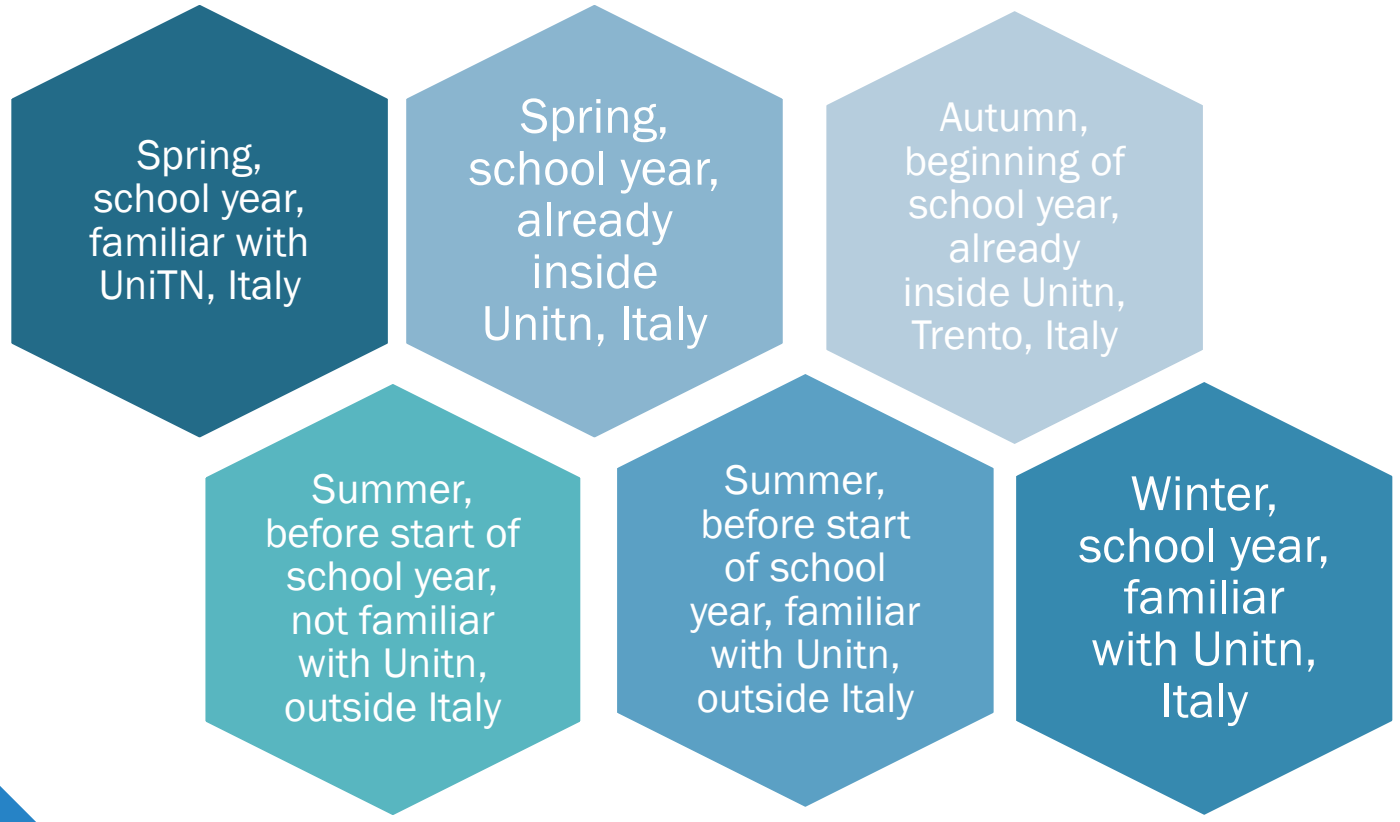
Competency questions

- A person in a scenario asks a query to be fulfilled by the KG
- For each person, 4 to 6 questions

Purpose Formalization

Scenarios

Aim to provide context for different uses of the knowledge graph.



Purpose Formalization

Personas

Marco Pierotti
Age: 19, Nationality: Italian
Occupation: high-school student
Description: Exploring study opportunities

Ivan Krumov
Age: 21, Nationality: Bulgarian
Occupation: Bachelor graduate
Description: Doing project in AI and is looking for information and help

Elena Ranzani
Age: 25, Nationality: Italian
Occupation: Assistant in biology UniTN
Description: Wants to do a PhD

Ginnie Anderson
Age: 32, Nationality: British
Occupation: editor in a Maths journal
Description: Looking for innovative maths-related articles to include and offer to people

Donatello Ferrari
Age: 63, Nationality: Italian
Occupation: Full-time professor in Psychology
Description: Looking for students/alumni to help him for his project

Paolo Lanza
Age: 44, Nationality: Italian
Occupation: Investor
Description: Passionate about Physics, wants to organize a conference

Purpose Formalization

Competency Questions

- Professors at Physics department?
- Courses of study at UniTN? Departments information?
- Authors of publications for AI?
- A way to contact such people?
- Articles only in English?
- Students who've written theses in Psychology? For which degrees?
- Publications to help a PhD research?
- Students and advisors for theses in Biology?
- Biology publications?
- Courses grouped by faculties and most prominent ones?
- Most prominent researchers in several fields?
- Theses from different degrees and their count?
- Access point to math publications' full text?


Inception

Inception phase

Data management

1. Separating nested structured datasets

- Extracting students, supervisors, co-supervisors, examiners from Theses
- Extracting files from publications
- Extracting positions from staff

2. Dealing with IDs

- Creation when not available
- Refined at Data Integration phase regarding the concept of URIs and Identification sets

3. Departments and Degrees

- Departments data was scraped from the Digital University Portal
- Degrees were extracted from the Courses distribution and then linked to theses and departments

4. Datasets as output informal resources:

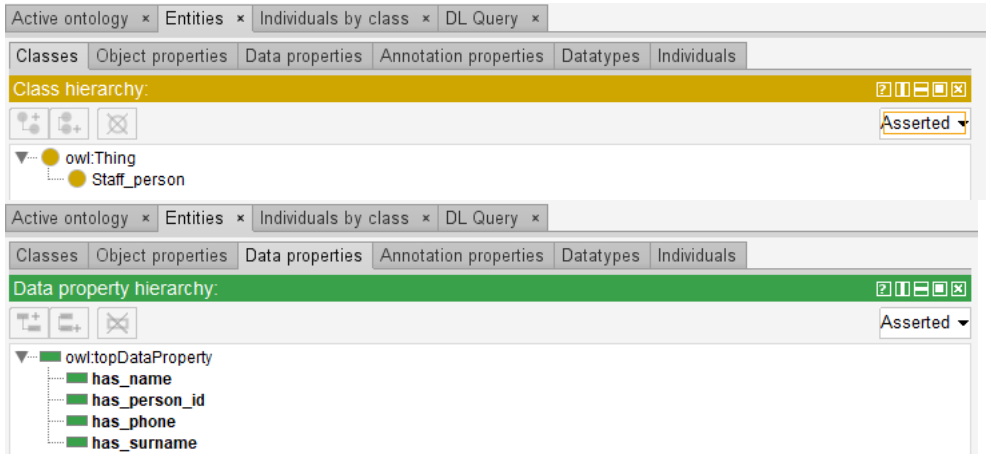
- Theses
- Students
- Supervisors
- Co-supervisors
- Examiners
- Publications
- Files
- Authors
- Staff persons
- Positions
- Degrees
- Departments

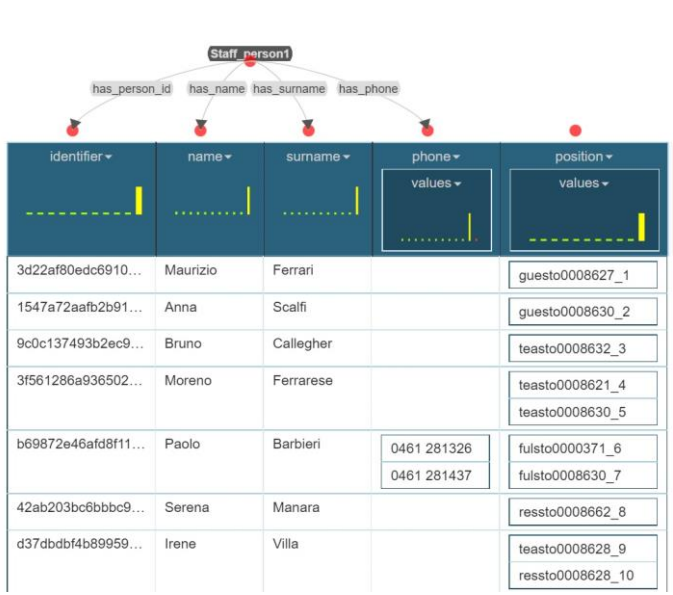
12 datasets as semi-formal resources to be utilized as input in the next phase. Format: CSV and JSON

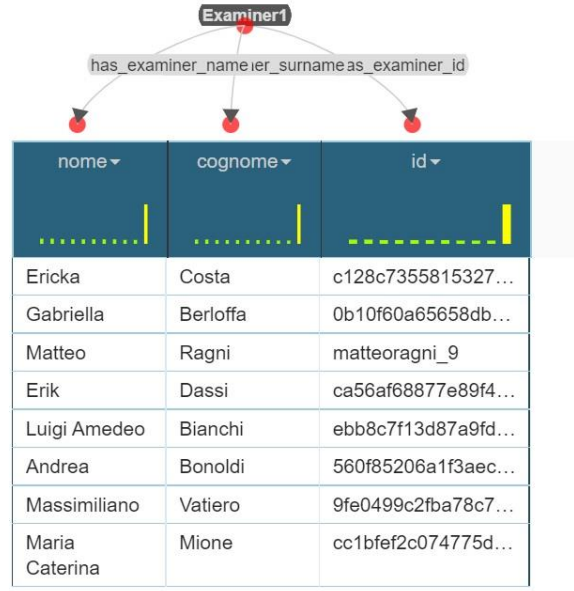
Inception phase

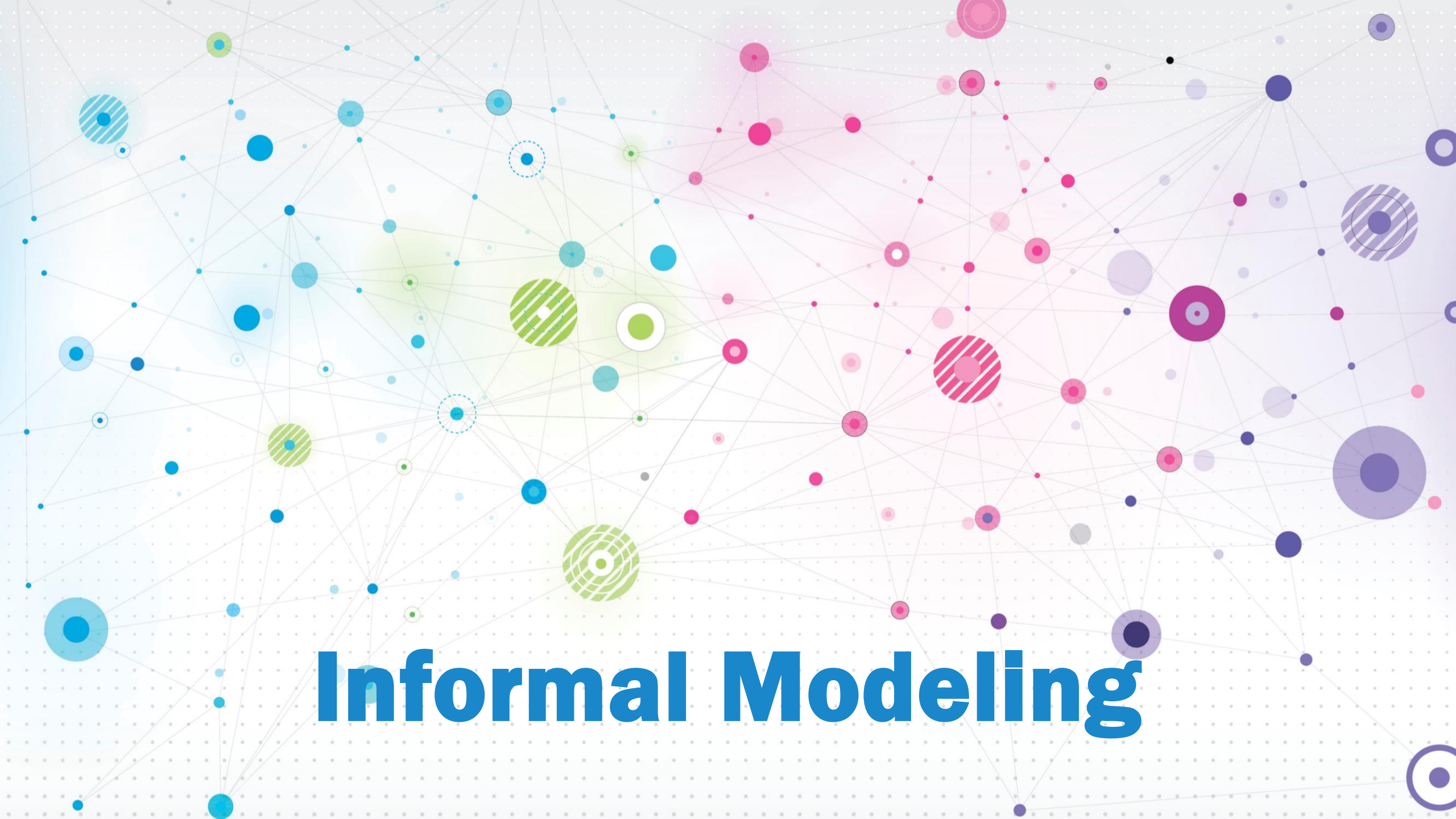
Single datasets ontologies

- Using Protégé, small ontologies were defined for each dataset. Each one represents an entity, with the data properties present as fields from the dataset files, keeping those fields which are useful for the purpose.
- Later, those were imported in Protégé to link with object properties and create the teleology.
- Using Karma Data Linker, these small ontologies were mapped to the datasets.





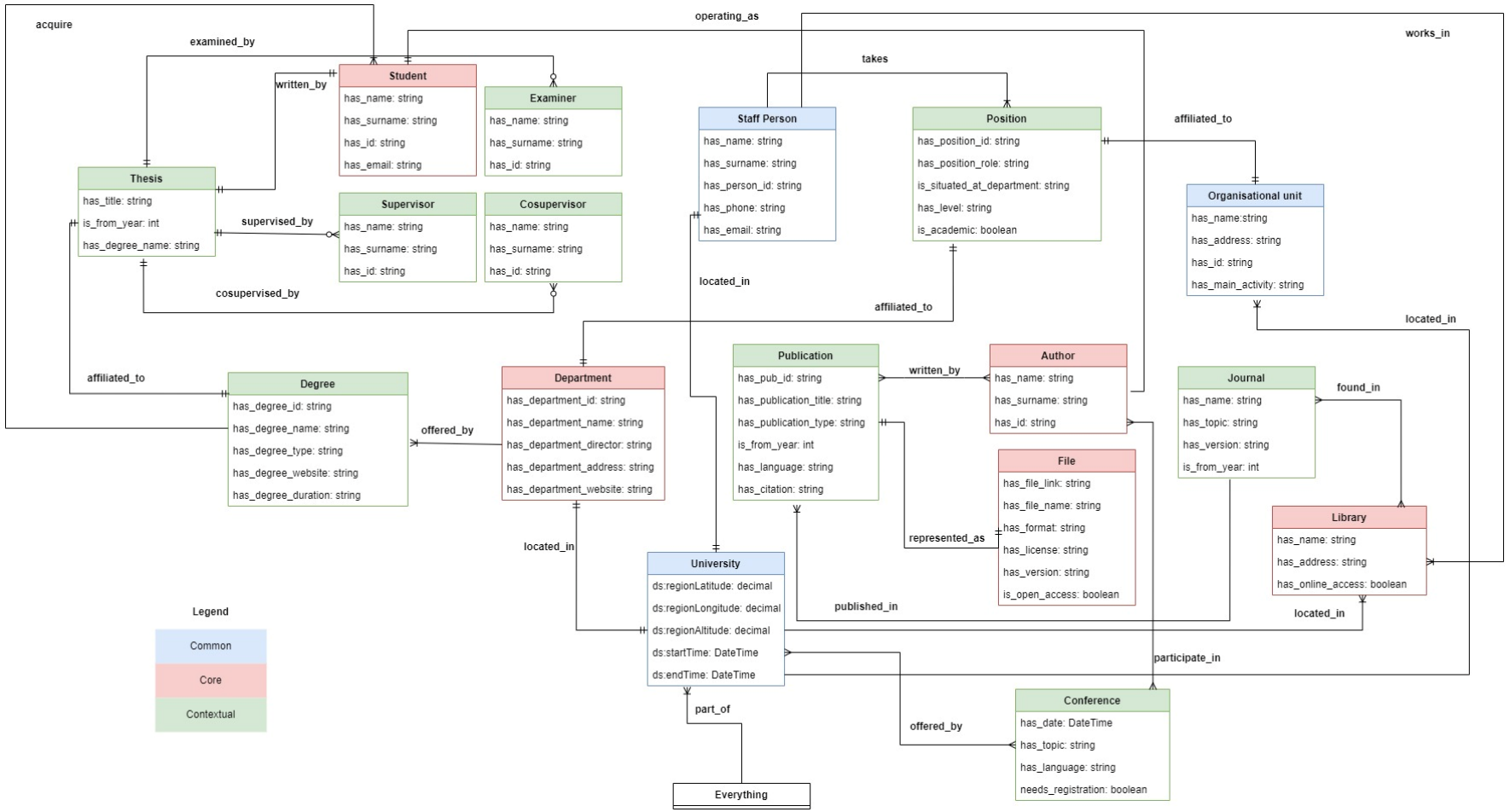




Informal Modeling

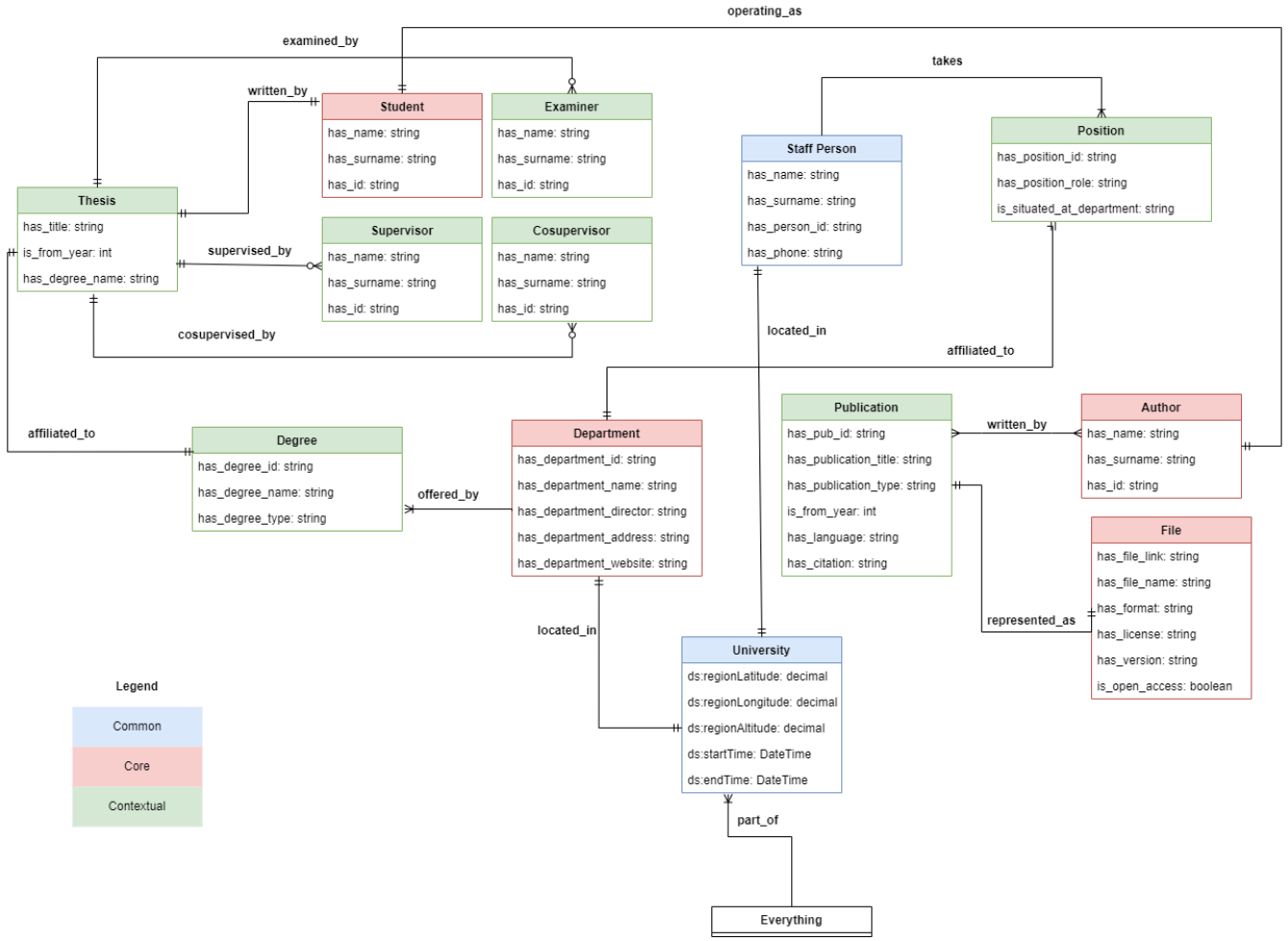
Informal Modeling phase

ER model General



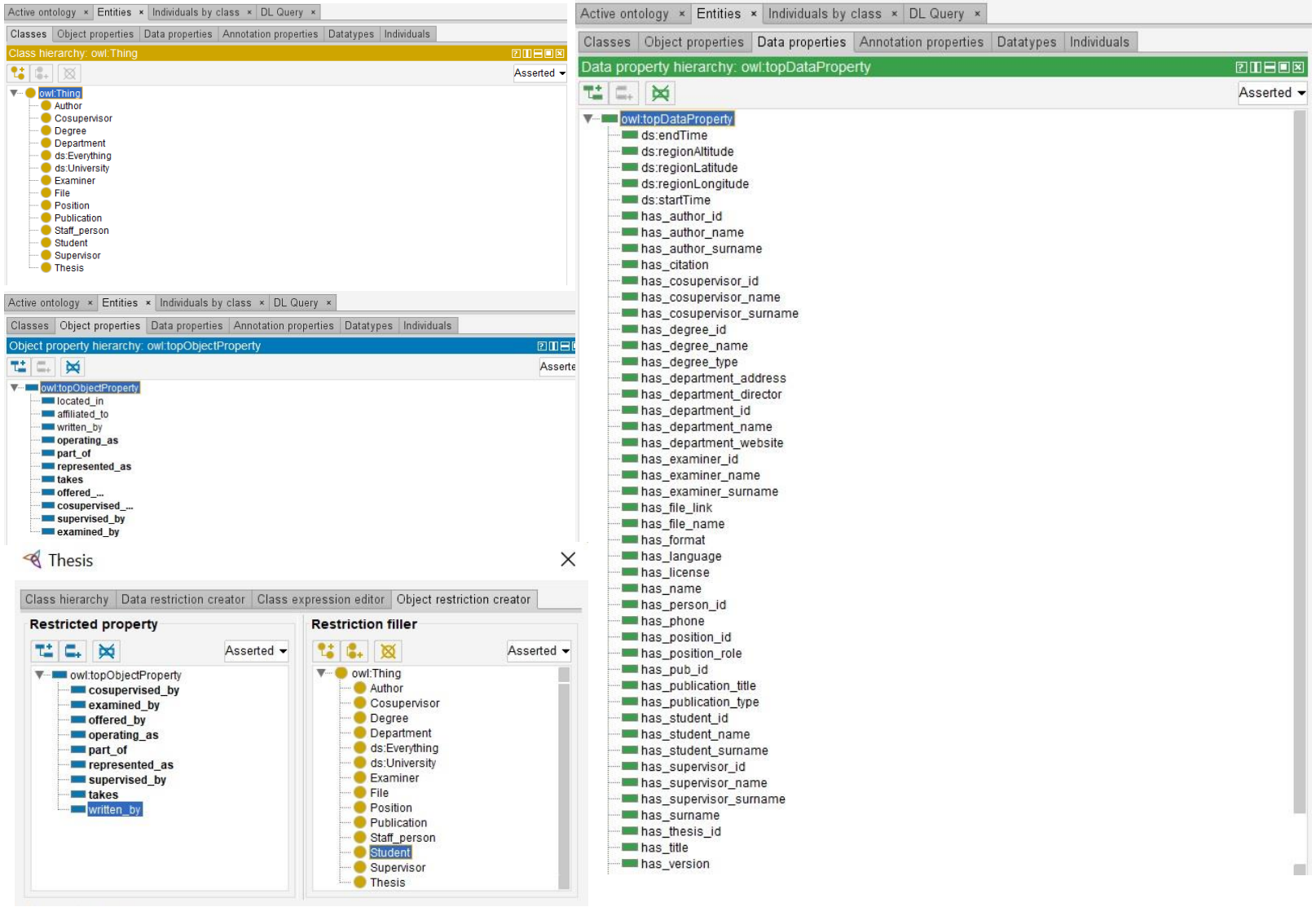
Informal Modeling phase

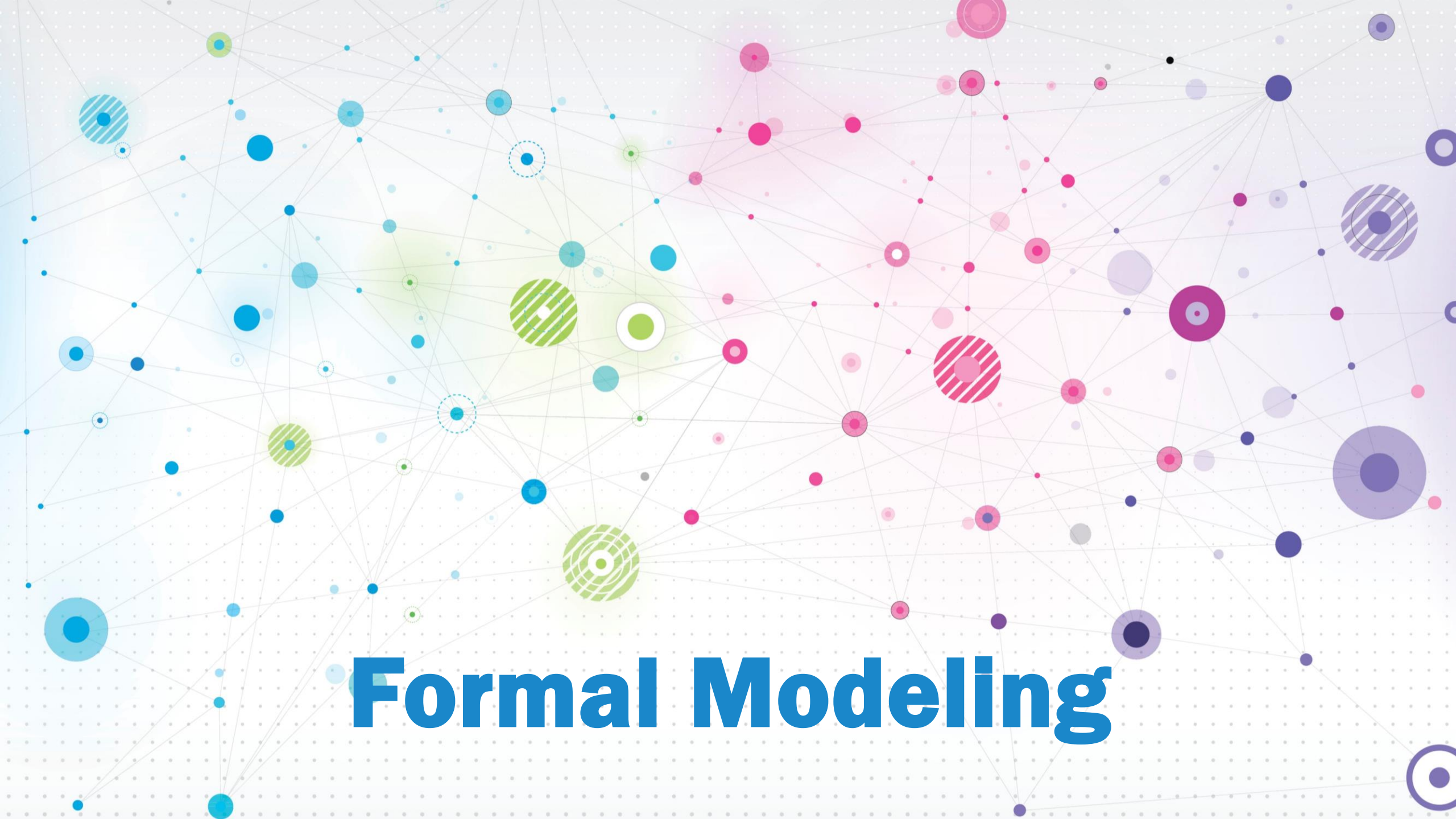
ER model Specific



Informal Modeling phase

Protégé Implementation Teleology



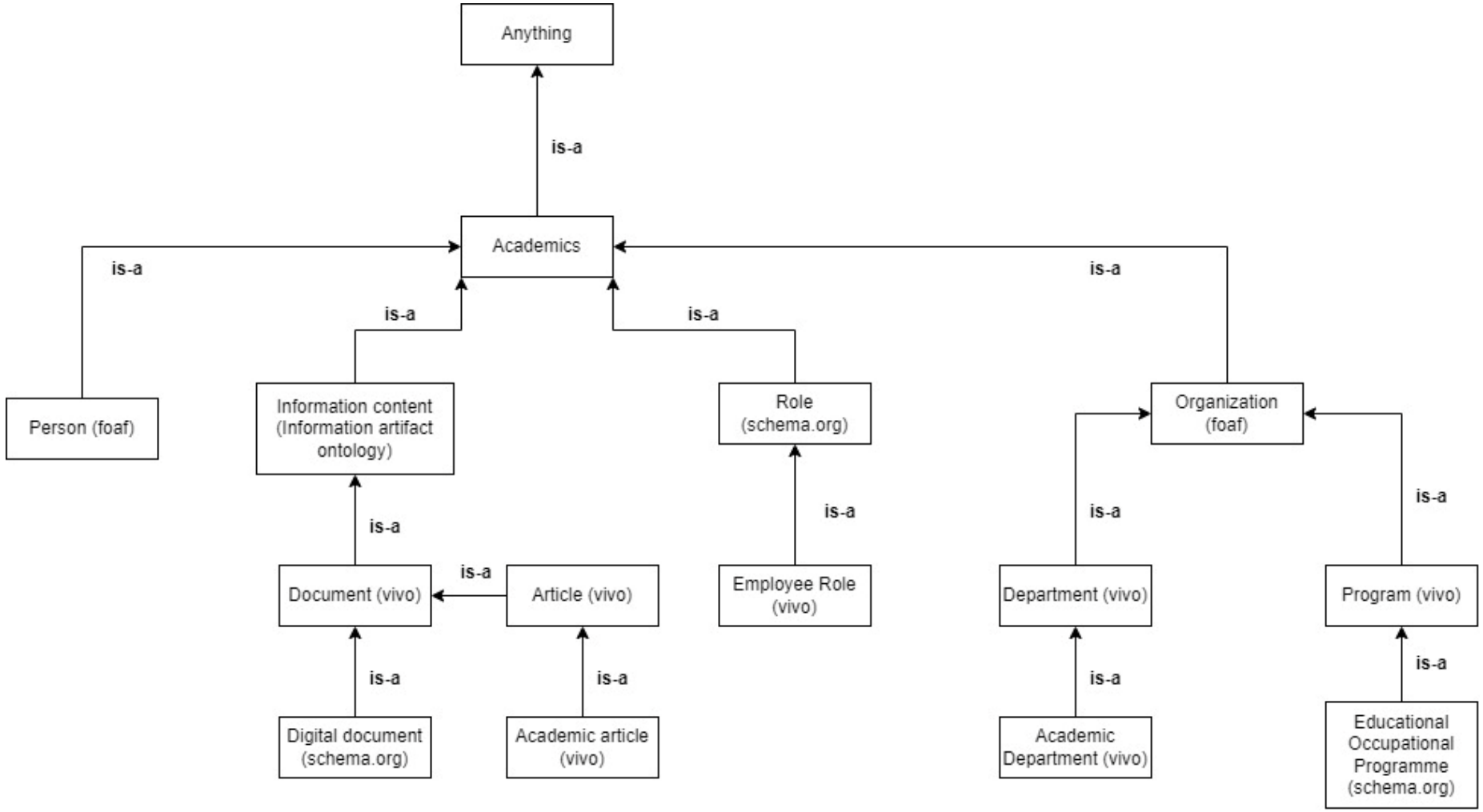


Formal Modeling

Formal modeling phase

Ontology

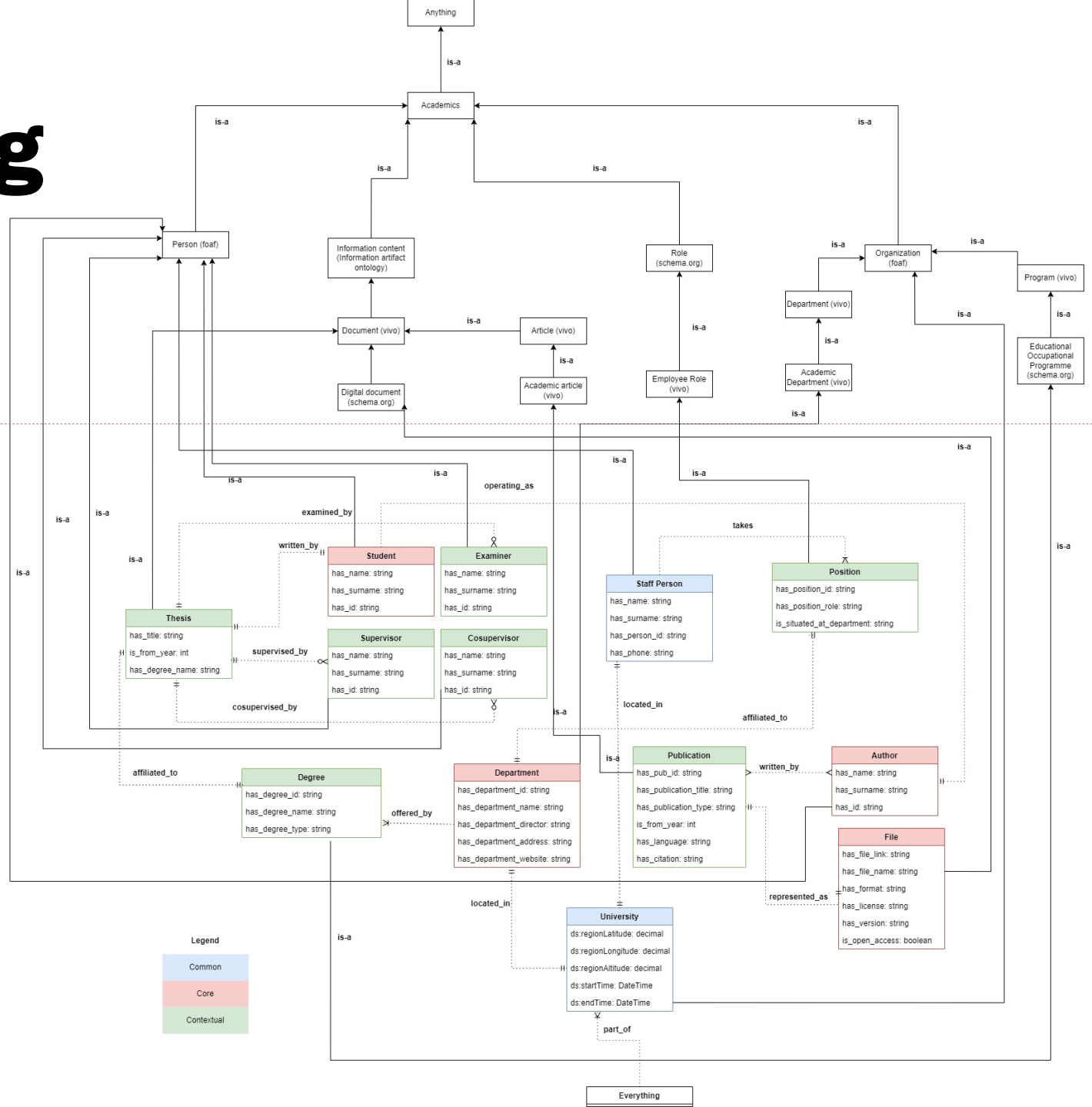
- Showing hierarchical relations
- The higher-level of abstraction knowledge layer
- E-types taken from Reference ontology for the purpose of reusability



Formal modeling

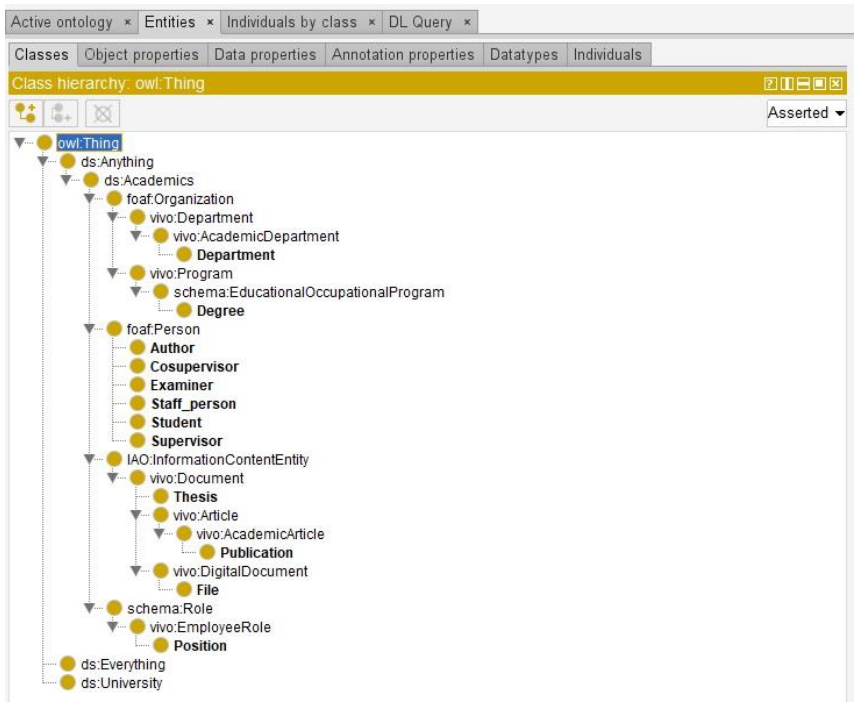
Teleontology – combining ontology and teleology

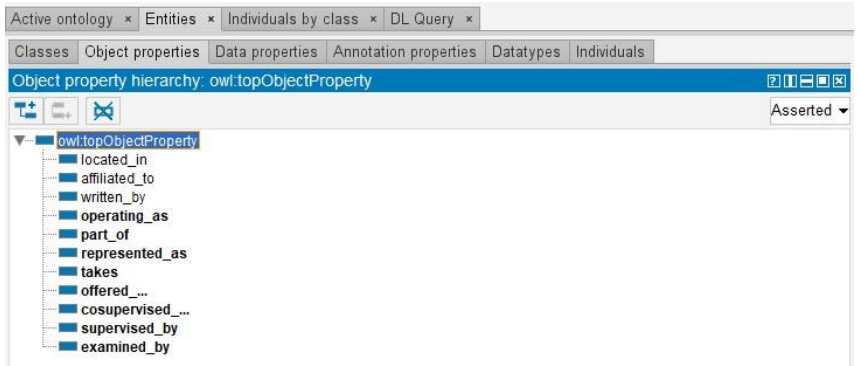
- How the E-types with their Data and Object properties are mapped to the schema
- A complete view of the structure of the KG

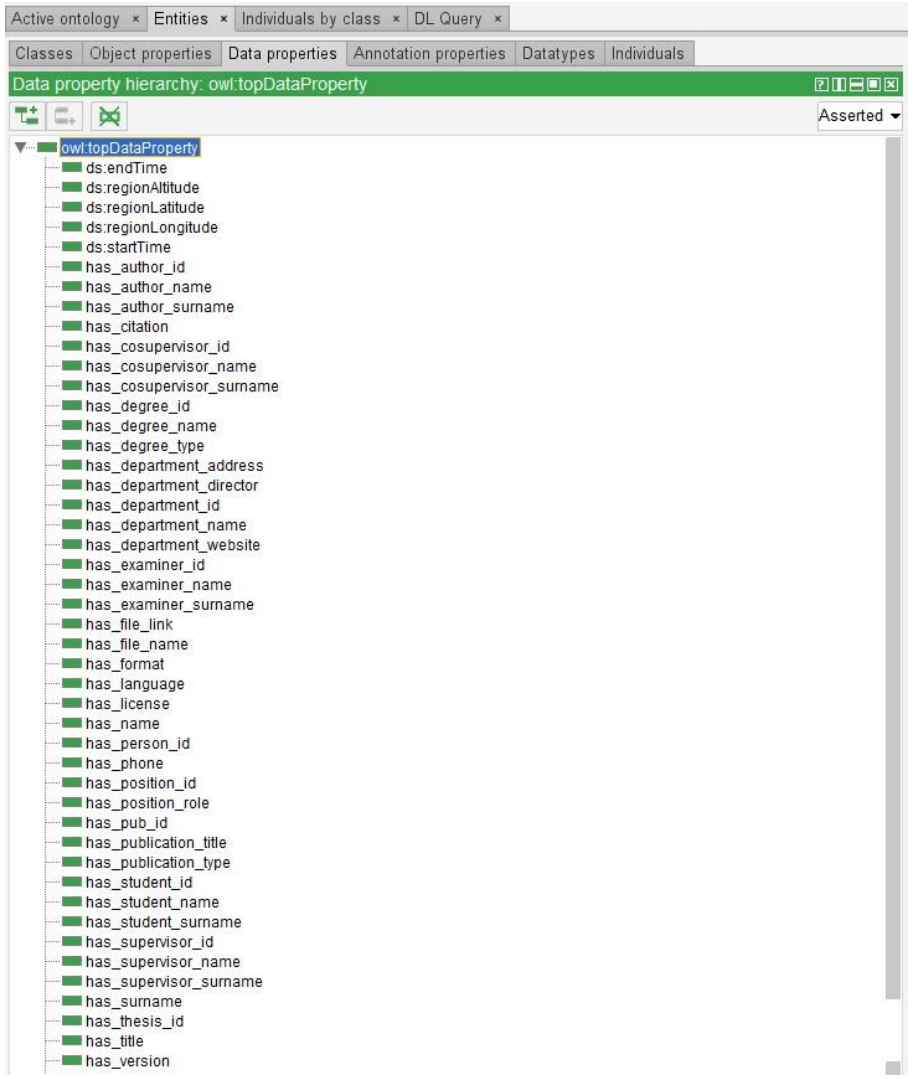


Formal Modeling phase

Protégé Implementation







Formal modeling phase

Language annotation

Concepts defined for the use of this KG, which enlarge the dictionary of meaning of words

Offered_by verb

- provide possibility to subscribe to a program/course/event

Publication noun

- An abstract, article or paper in a journal or electronic repository

Cosupervisor noun

- Someone who is a secondary supervisor

File noun

- An electronic representation of a document

Has_department_director noun

- Someone who is the main responsible person for a department

Has_license noun

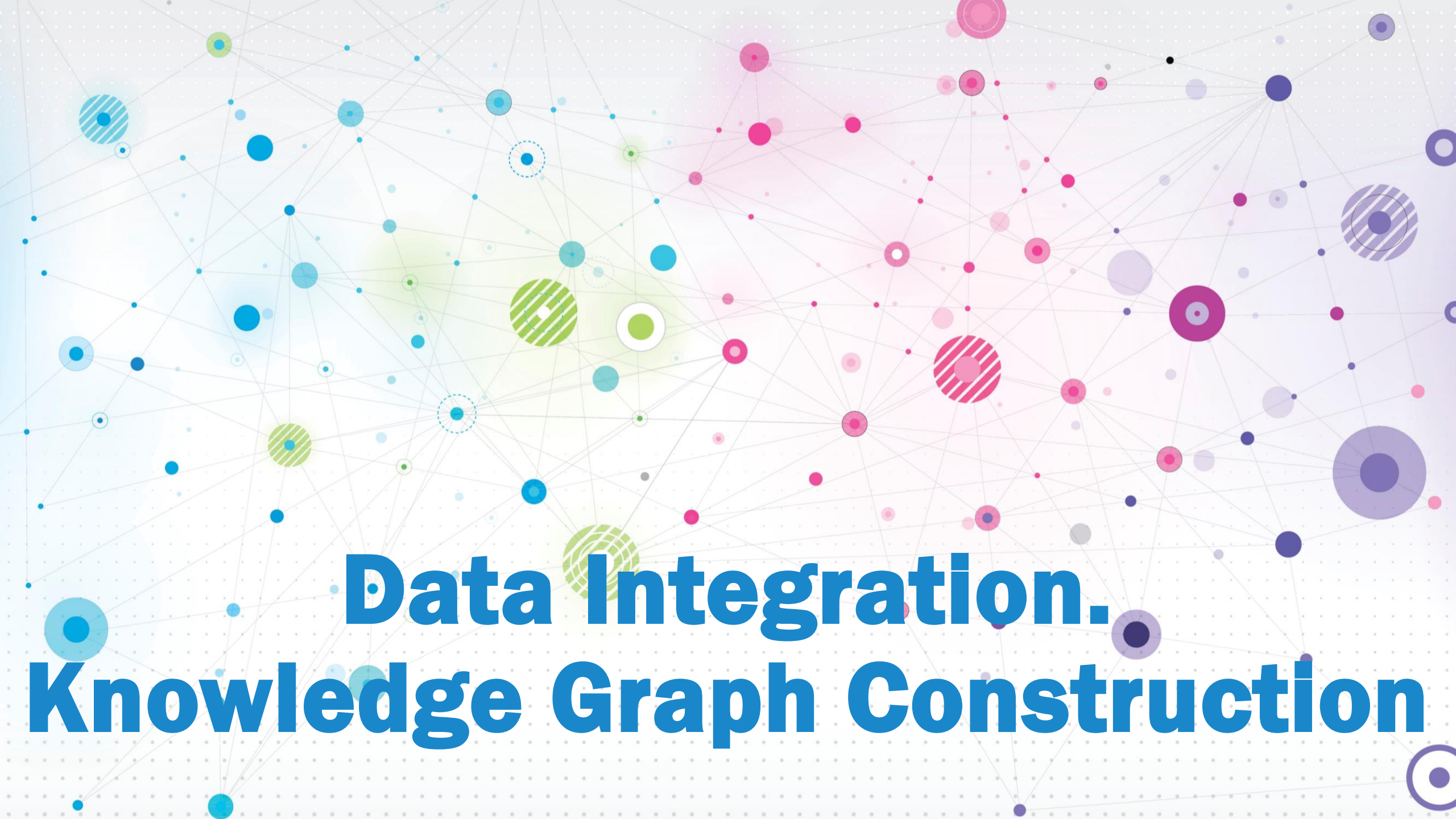
- Kind of permission a certain document has

Has_format noun

- The electronic format of a file

Has_file_link noun

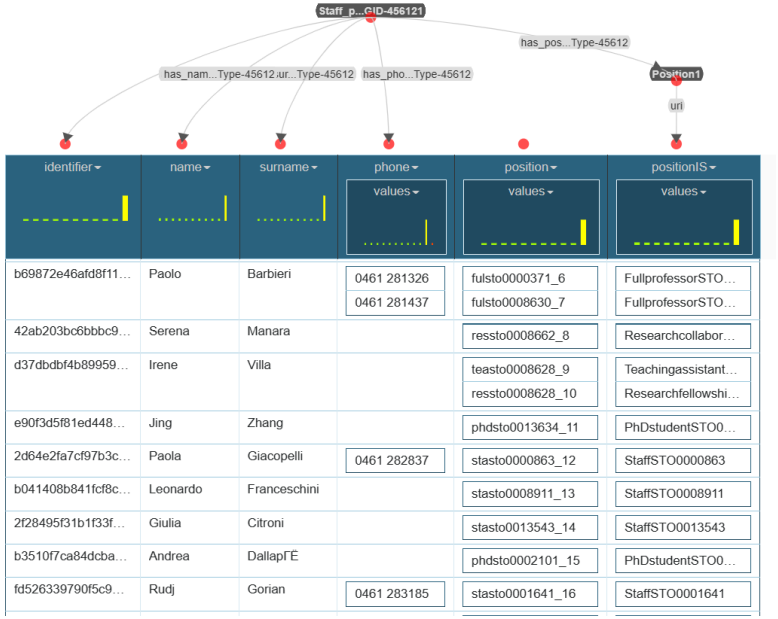
- A reference to access a website

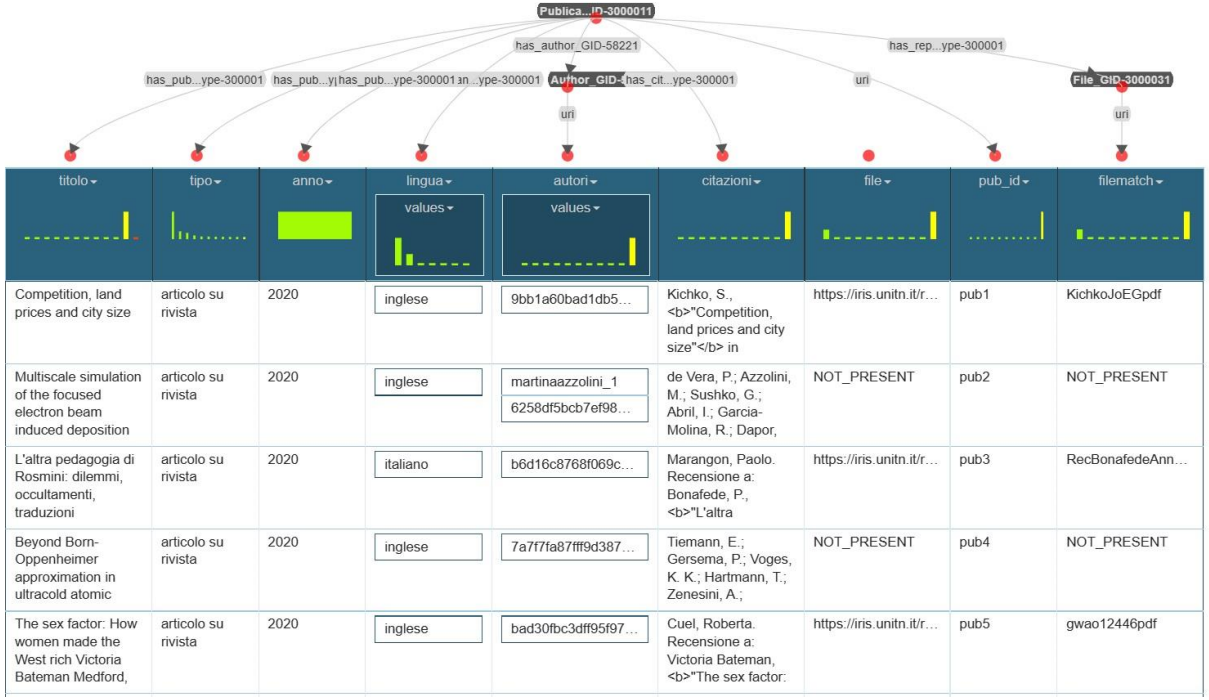

Data Integration.
Knowledge Graph Construction

Data Integration. Knowledge Graph Construction

Entity Identification and Semantic Heterogeneity. Entity matching

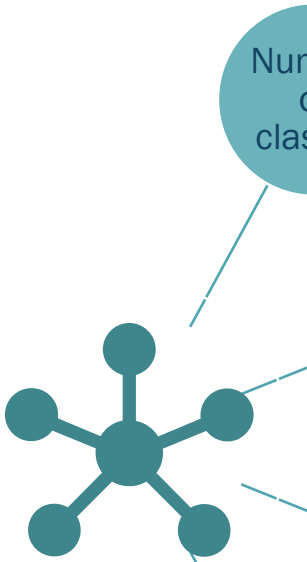
- Some datasets already had IDs (persons, departments)
- For others Identification Sets had to be identified from the data (e.g. for Position: *role* + *department*)
- Since datasets were divided at the beginning, no nested structures
- Karma Data Linker efficiently manages these issues



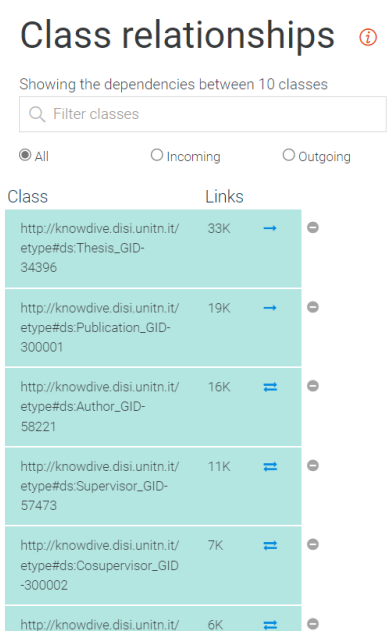


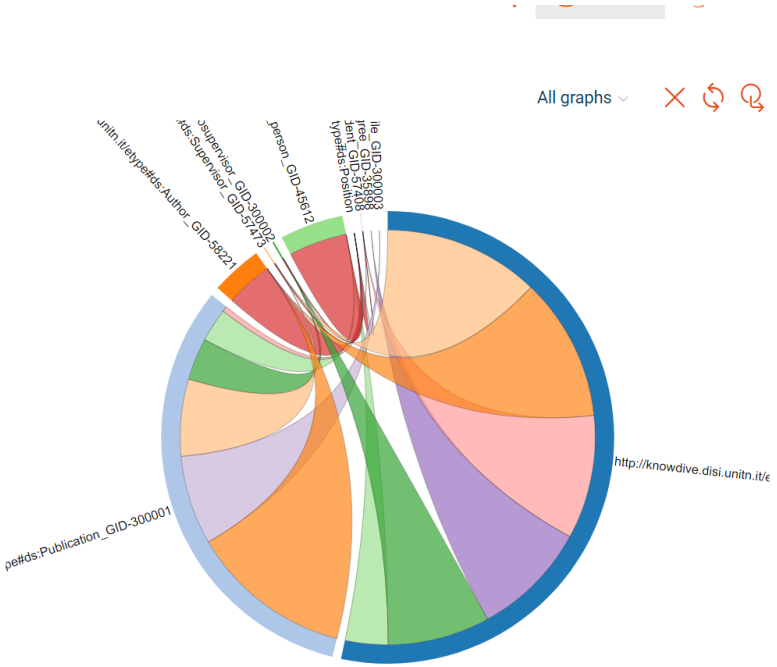
Data Integration. Knowledge Graph Construction

Evaluation. Some statistics



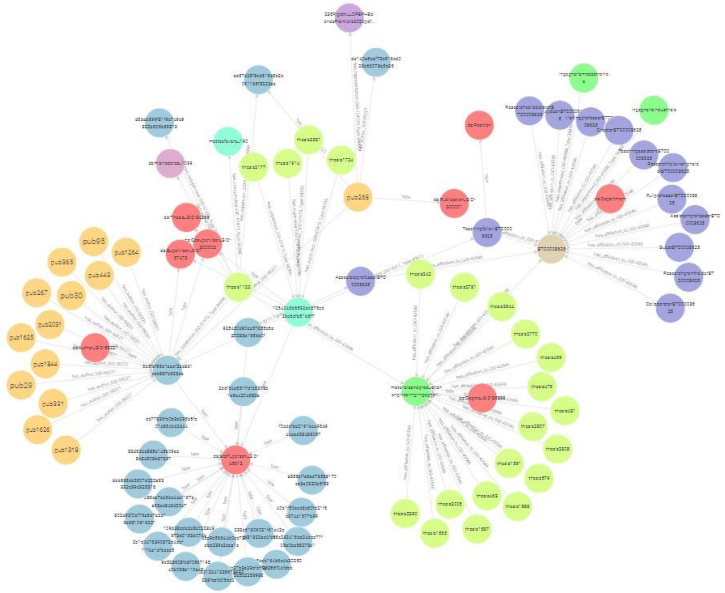
- Number of classes
  - 12
- Number of DP
  - 2 to 6 per class
- Number of OP
  - 11
- Links
  - Most outgoing links from Theses (18K) and Publications (11K)

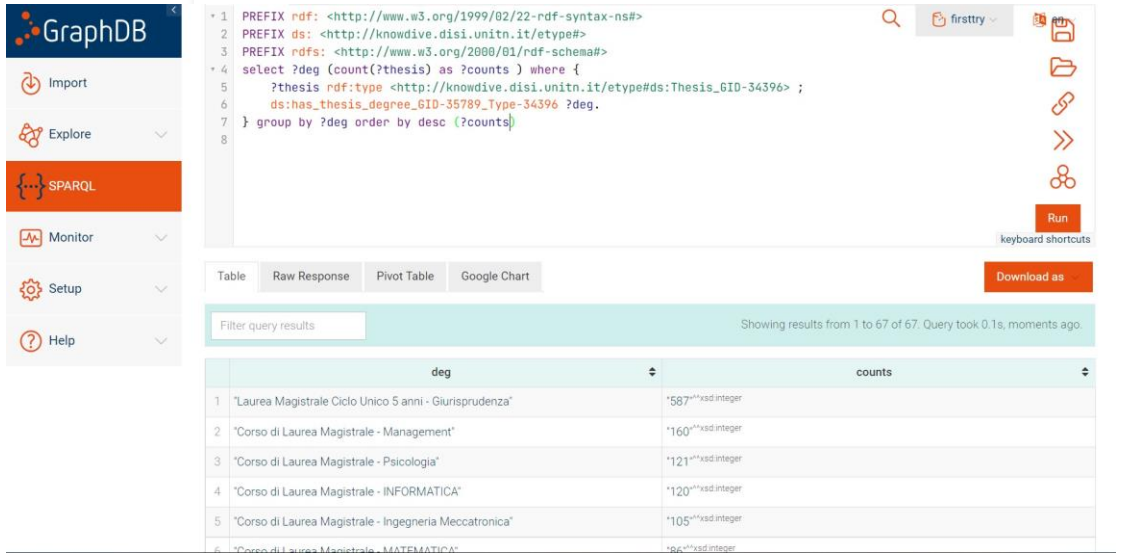




GraphDB

Visual graph





Let's see a demo of SPARQL queries!

Conclusions

Conclusions and Open Issues

Result

Fairly well connected graph.

Two entity types with high number of outgoing links.

Not so much information about persons.

Artifacts as outcomes of the phases

Cleaned and organized datasets
Metadata
Teleology, Ontology, and Teleontology models
Inception sheet
Table with language annotation

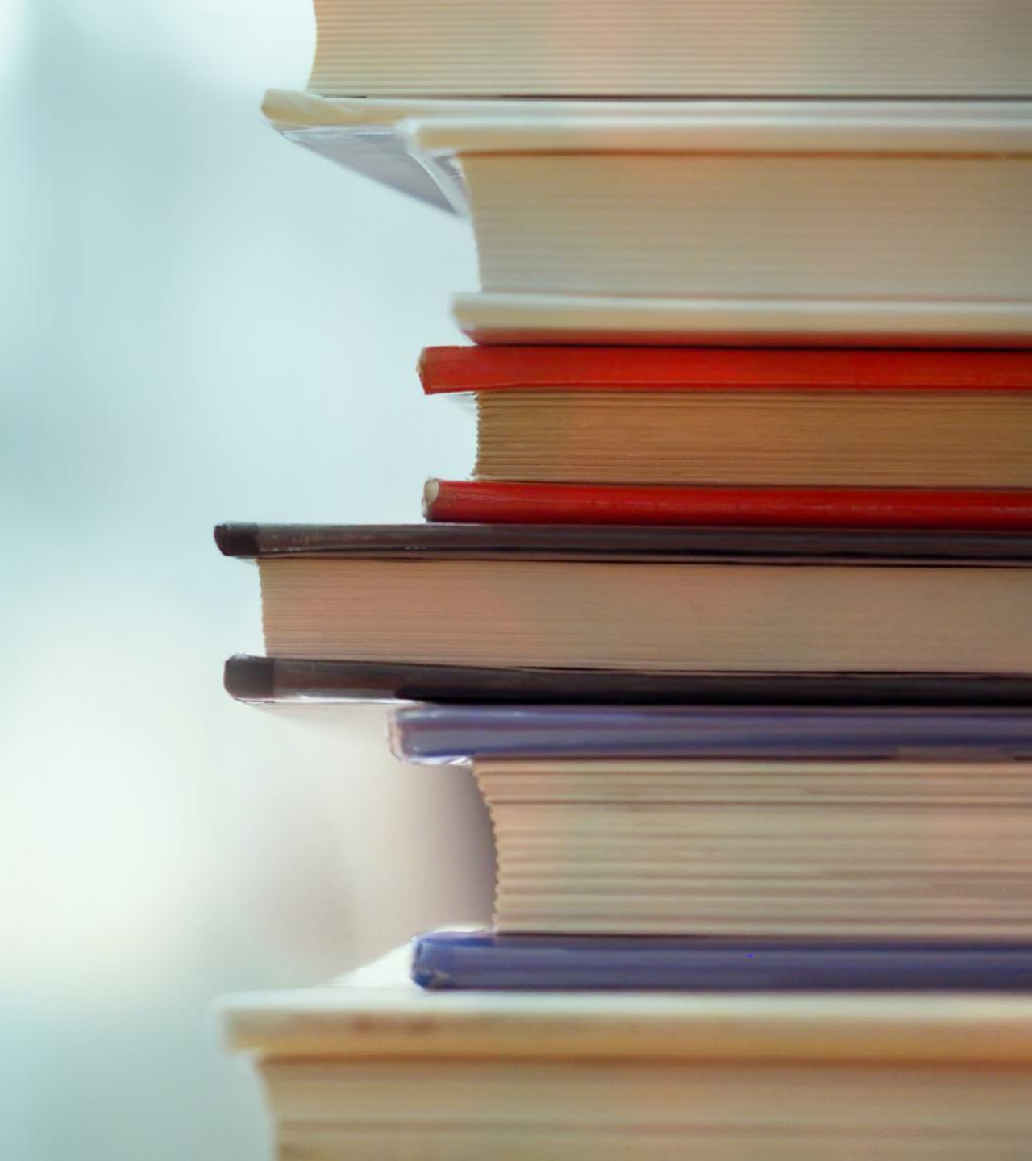
Open Issues

Should Journals and/or Conferences be included in the KG and if yes, how should they be linked?

Is there a way to collect more information about people? This would provide more information about the issue of entity recognition, and if two people are the same.

Each entity person who belongs to different classes has DP coming from all of them and they overlap.

Is there a way to get some kind of indicator for quality of department/degree program?



Thank you!

Q&A