



UNIVERSITY
OF TRENTO - Italy



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

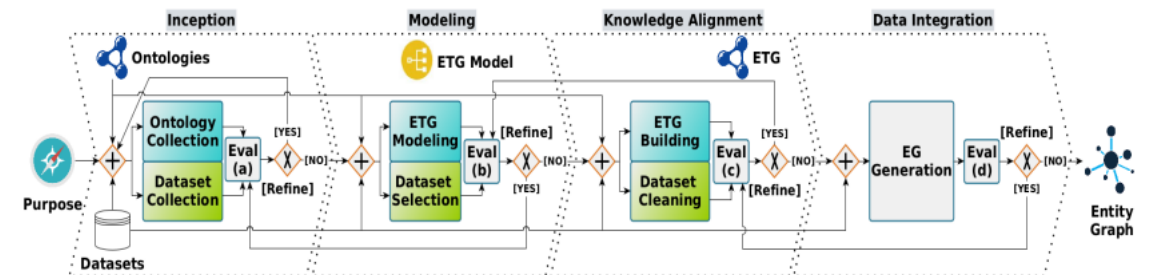
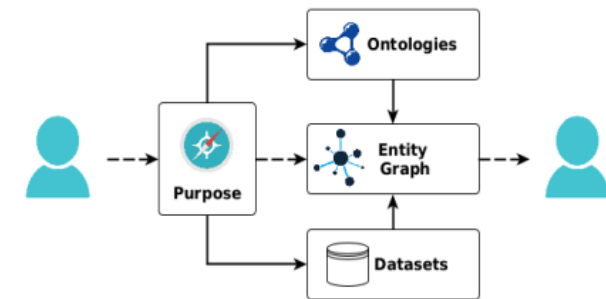
04. Inception

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

An abstract network diagram with a light gray background featuring a fine grid of dots. A complex web of thin gray lines connects various nodes. The nodes are represented by circles of different sizes and colors, including blue, green, pink, and purple. Some nodes have internal patterns like stripes or concentric circles. The overall composition suggests a global or interconnected network.

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

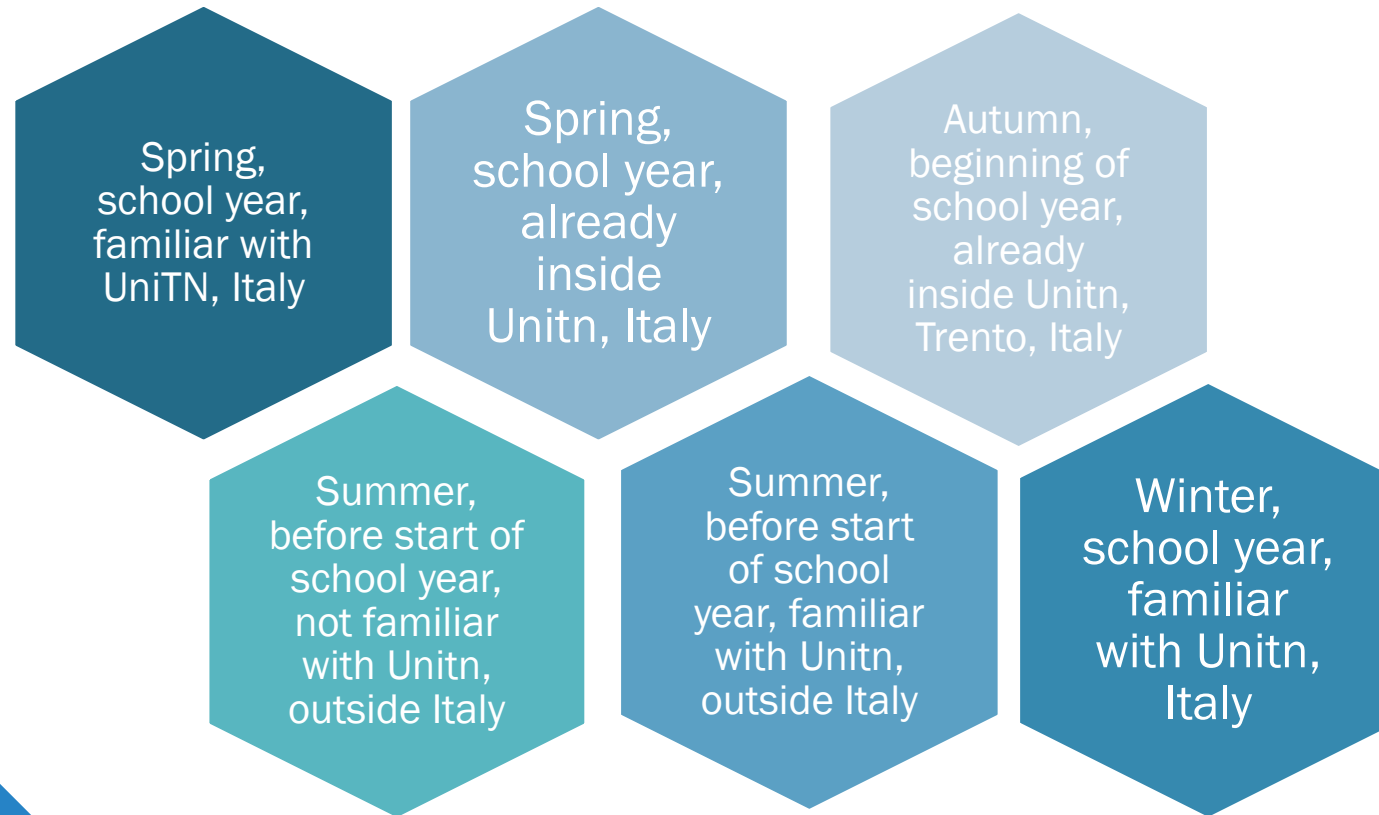
A person in a scenario asks a query to be fulfilled by the KG

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:

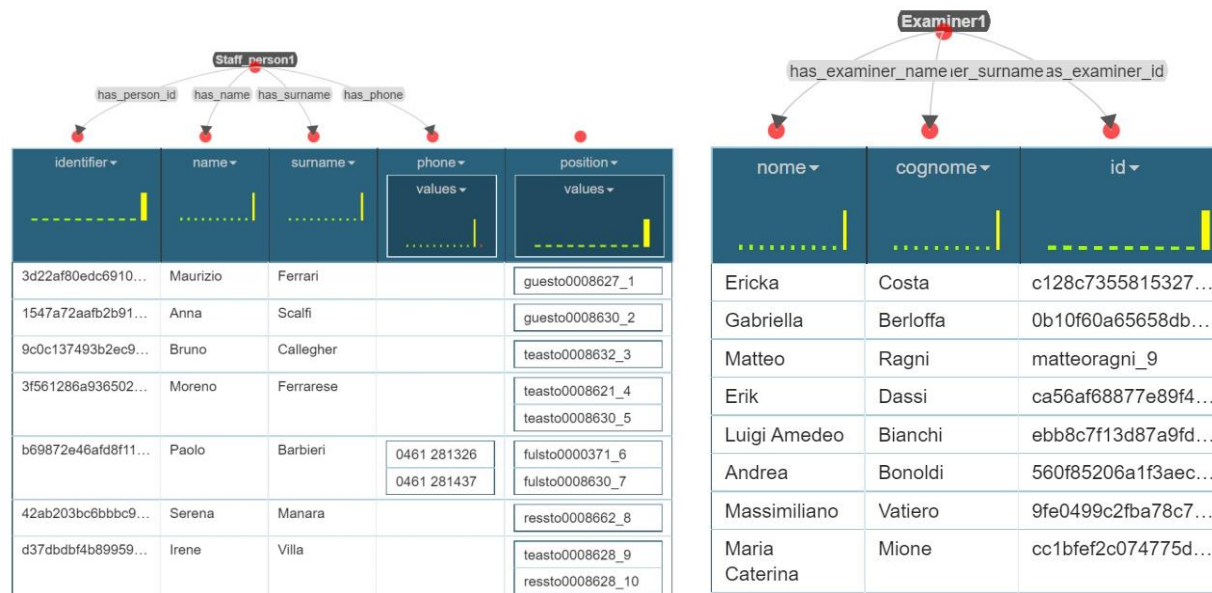
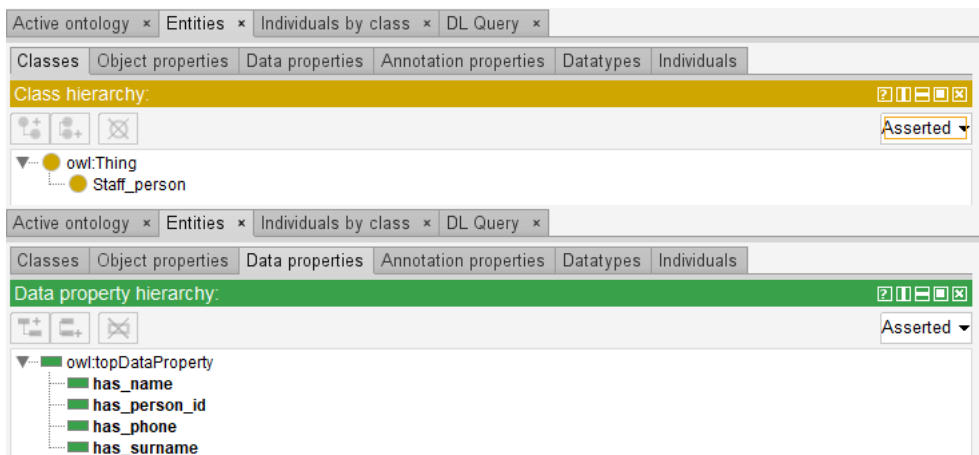
- Theses
- Students
- Supervisors
- Co-supervisors
- Examiners
- Publications
- Files
- Authors
- Staff persons
- Positions
- Degrees
- Departments

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.

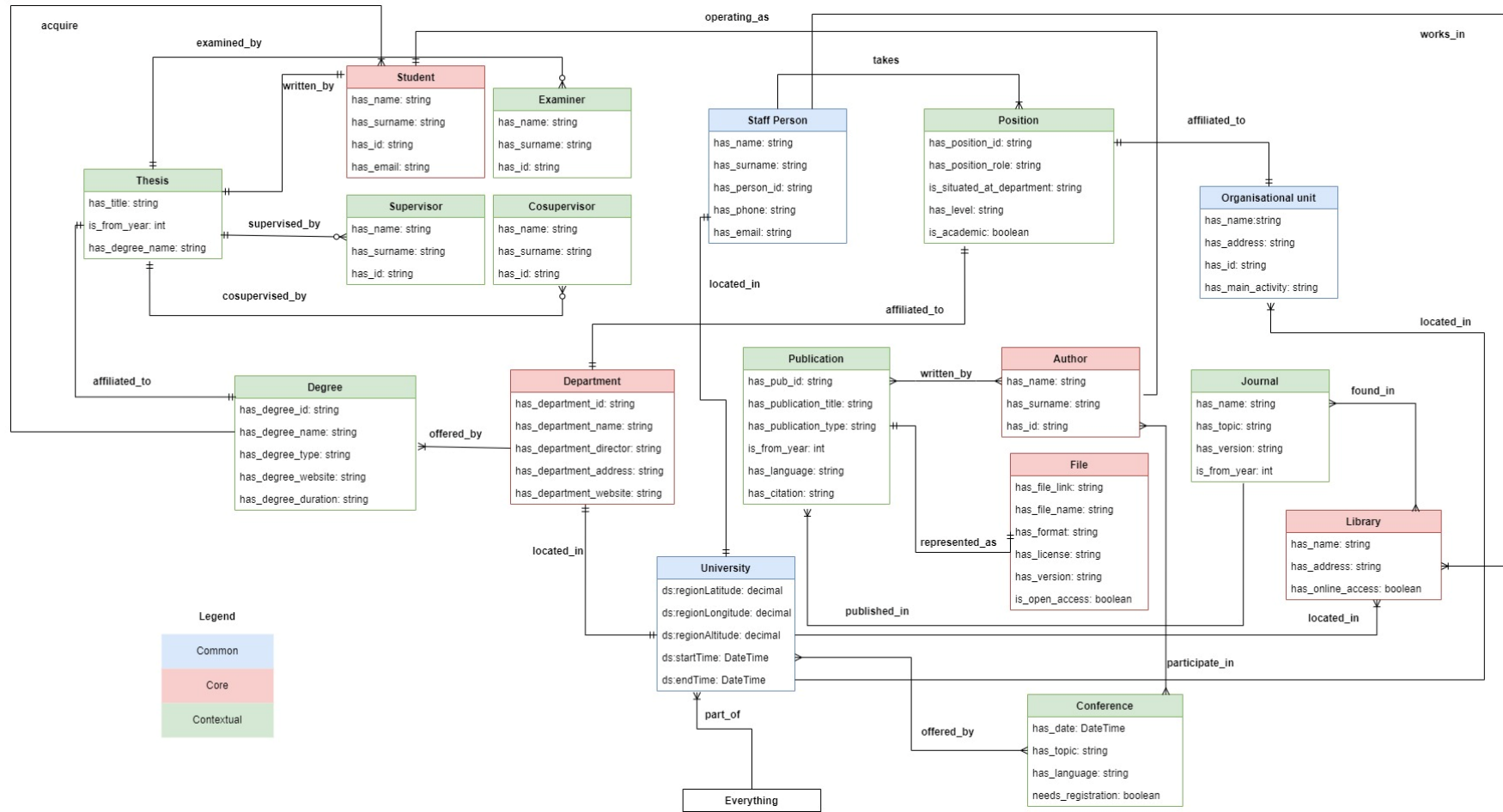


The background of the slide is a complex, abstract network diagram. It consists of numerous nodes of various sizes and colors (blue, green, pink, purple) connected by a dense web of thin, light gray lines. Some nodes are solid, while others have patterns like stripes or concentric circles. The overall effect is a sense of interconnectedness and data flow.

Informal Modeling

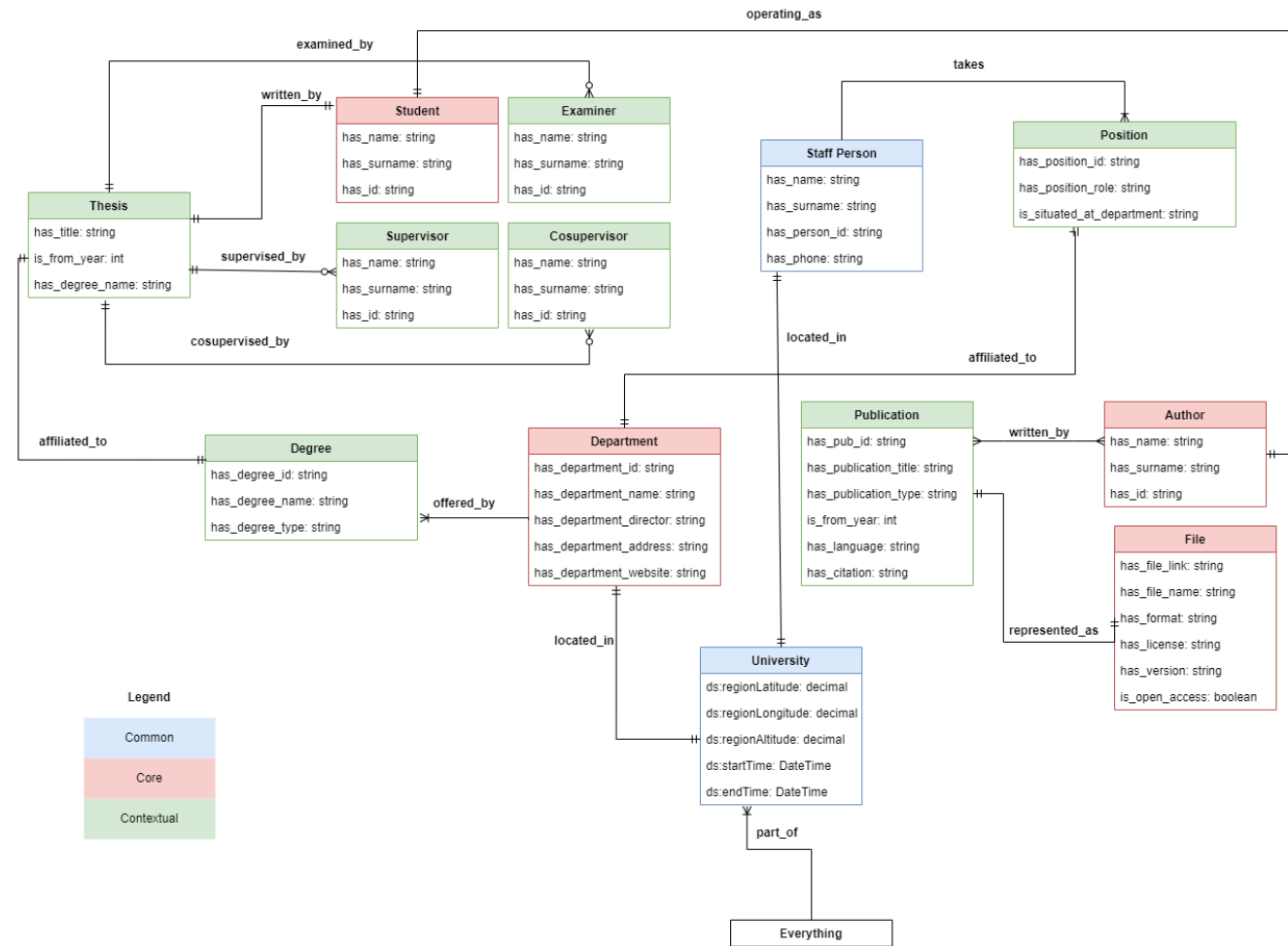
Informal Modeling phase

ER model General



Informal Modeling phase

ER model Specific



Informal Modeling phase

Protégé Implementation Teleology

The image displays three screenshots of the Protégé ontology editor, illustrating the implementation of teleology in an ontology.

Top Left Screenshot: Shows the "Class hierarchy: owl:Thing" view. The hierarchy is rooted at `owl:Thing` and includes the following classes: `Author`, `Cosupervisor`, `Degree`, `Department`, `ds:Everything`, `ds:University`, `Examiner`, `File`, `Position`, `Publication`, `Staff_person`, `Student`, `Supervisor`, and `Thesis`.

Top Right Screenshot: Shows the "Data property hierarchy: owl:topDataProperty" view. The hierarchy is rooted at `owl:topDataProperty` and includes the following data properties: `ds:endTime`, `ds:regionAltitude`, `ds:regionLatitude`, `ds:regionLongitude`, `ds:startTime`, `has_author_id`, `has_author_name`, `has_author_surname`, `has_citation`, `has_cosupervisor_id`, `has_cosupervisor_name`, `has_cosupervisor_surname`, `has_degree_id`, `has_degree_name`, `has_degree_type`, `has_department_address`, `has_department_director`, `has_department_id`, `has_department_name`, `has_department_website`, `has_examiner_id`, `has_examiner_name`, `has_examiner_surname`, `has_file_link`, `has_file_name`, `has_format`, `has_language`, `has_license`, `has_name`, `has_person_id`, `has_phone`, `has_position_id`, `has_position_role`, `has_pub_id`, `has_publication_title`, `has_publication_type`, `has_student_id`, `has_student_name`, `has_student_surname`, `has_supervisor_id`, `has_supervisor_name`, `has_supervisor_surname`, `has_surname`, `has_thesis_id`, `has_title`, and `has_version`.

Bottom Left Screenshot: Shows the "Object property hierarchy: owl:topObjectProperty" view. The hierarchy is rooted at `owl:topObjectProperty` and includes the following object properties: `located_in`, `affiliated_to`, `written_by`, `operating_as`, `part_of`, `represented_as`, `takes`, `offered...`, `cosupervised...`, `supervised_by`, and `examined_by`.

Bottom Right Screenshot: Shows the "Restricted property" and "Restriction filler" views. The "Restricted property" view shows the hierarchy for `owl:topObjectProperty` with the following properties: `cosupervised_by`, `examined_by`, `offered_by`, `operating_as`, `part_of`, `represented_as`, `supervised_by`, `takes`, and `written_by`. The "Restriction filler" view shows the hierarchy for `owl:Thing` with the following fillers: `Author`, `Cosupervisor`, `Degree`, `Department`, `ds:Everything`, `ds:University`, `Examiner`, `File`, `Position`, `Publication`, `Staff_person`, `Student`, `Supervisor`, and `Thesis`.

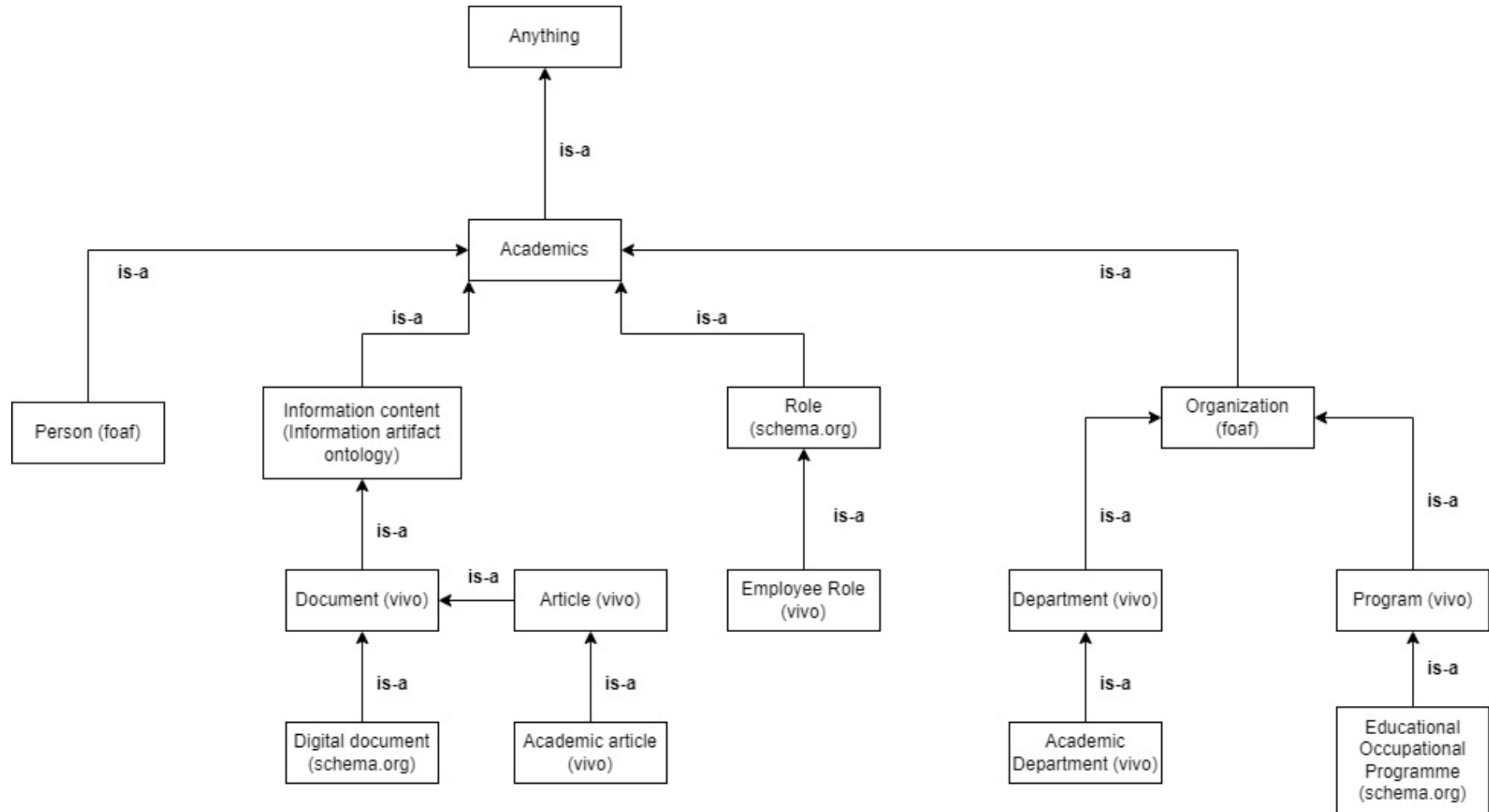
An abstract network diagram with a light gray background featuring a fine grid of dots. A complex web of thin gray lines connects various nodes. The nodes are represented by circles of different sizes and colors, including blue, green, pink, and purple. Some nodes have internal patterns like stripes or concentric circles. The overall composition suggests a complex, interconnected system.

Formal Modeling

Formal modeling phase

Ontology

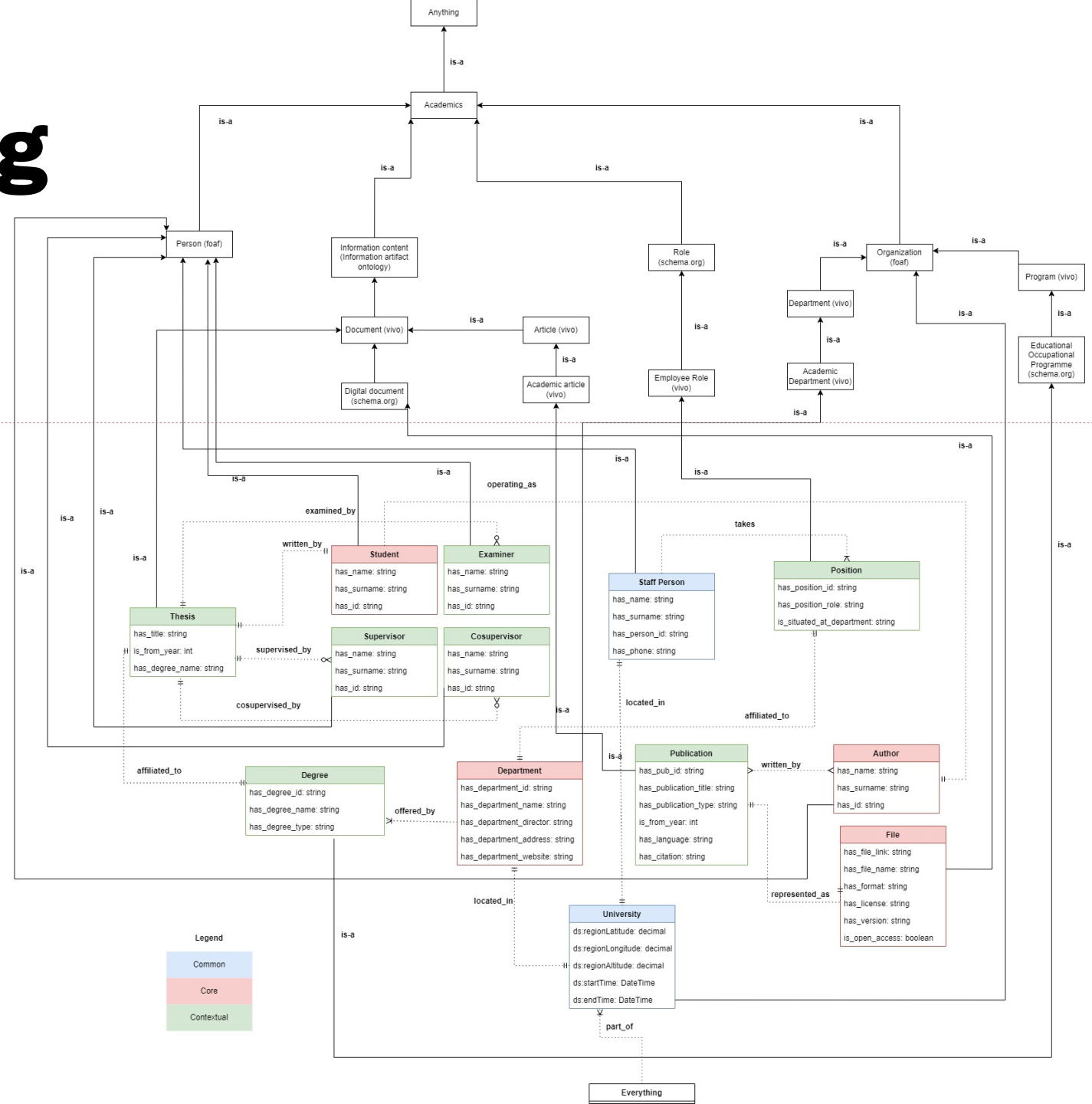
- Showing hierarchical relations
- The higher-level of abstraction knowledge layer
- E-types taken from Reference ontology for the purpose of reusability



Formal modeling

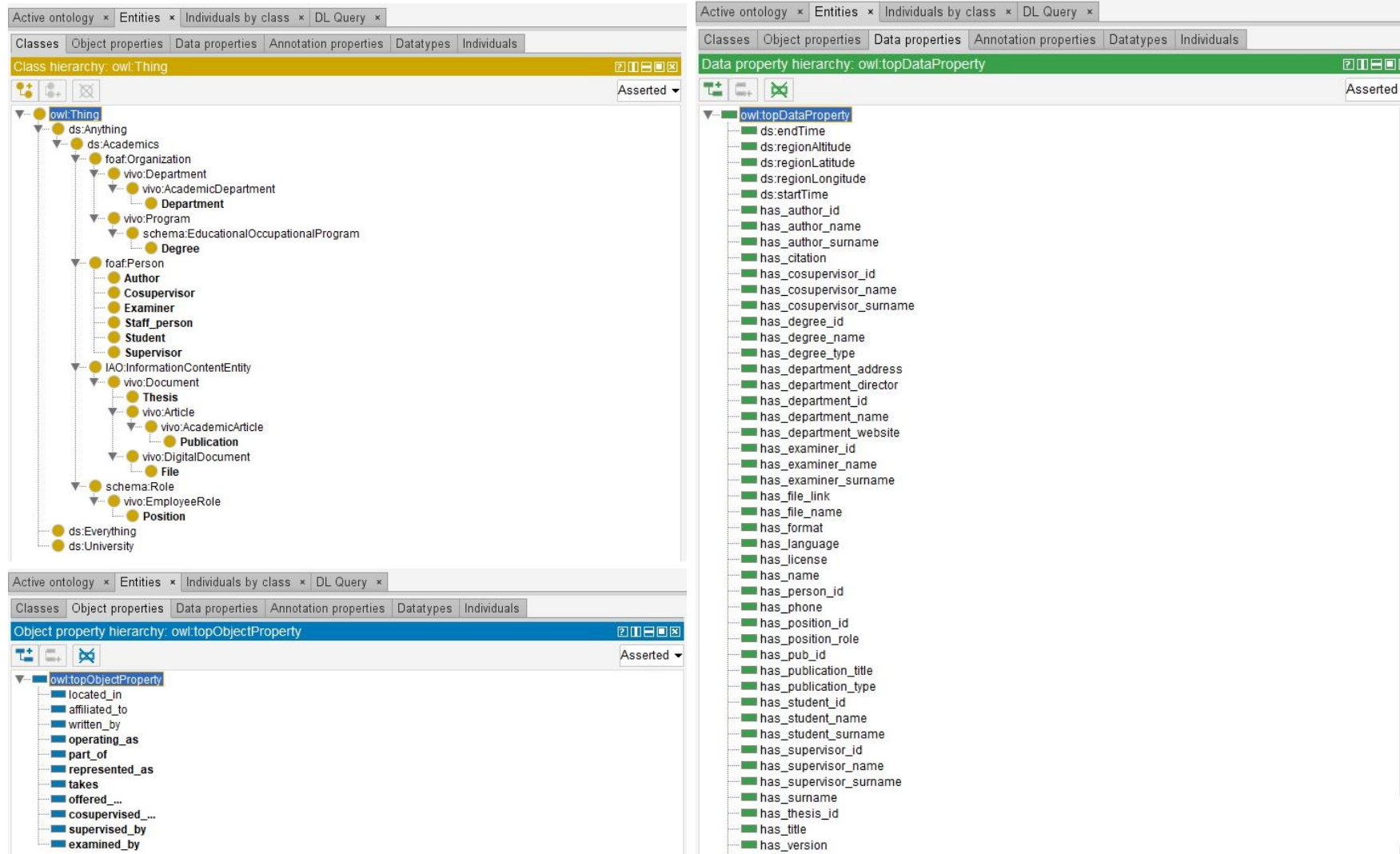
Teleontology – combining ontology and teleology

- 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

Has_department_director noun

- Someone who is the main responsible person for a department

Publication noun

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

Has_license noun

- Kind of permission a certain document has

Cosupervisor noun

- Someone who is a secondary supervisor

Has_format noun

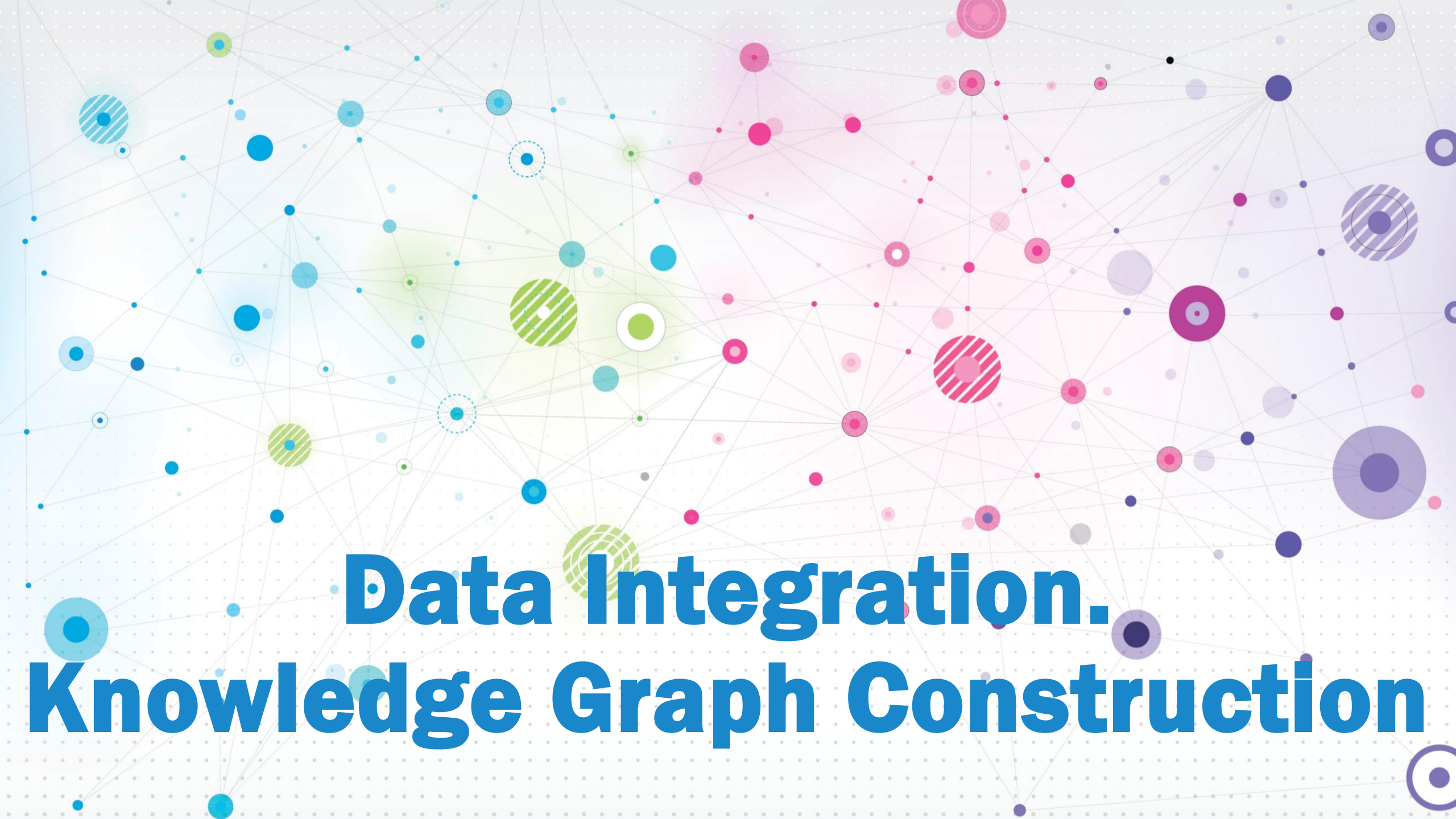
- The electronic format of a file

File noun

- An electronic representation of a document

Has_file_link noun

- A reference to access a website



Data Integration. Knowledge Graph Construction

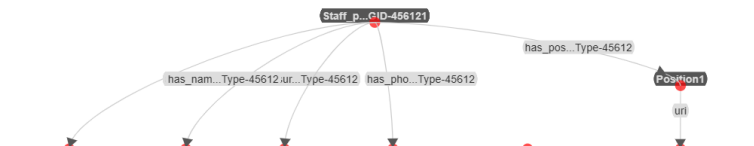
Data Integration.

Knowledge Graph Construction

Entity Identification and Semantic Heterogeneity.

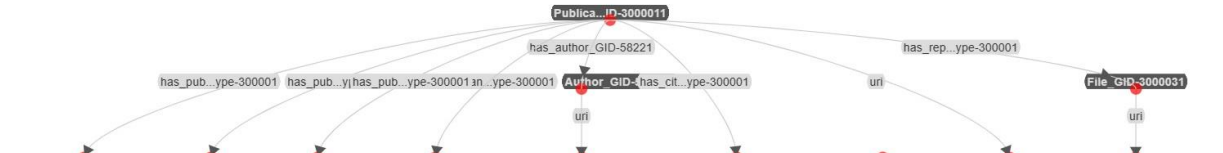
Entity matching

- 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



A knowledge graph diagram showing a central node 'Staff p...GID-456121' connected to several other nodes: 'has_nam...Type-45612:ur...Type-45612', 'has_pos...Type-45612', 'has_pho...Type-45612', and 'Position1'. The 'Position1' node is further connected to a 'uri' node.

identifier	name	surname	phone	position	positionIS
b69872e46afdf11...	Paolo	Barbieri	0461 281326 0461 281437	fulsto000371_6 fulsto0008630_7	FullprofessorSTO...
42ab203bc8bbbc9...	Serena	Manara		resto0008662_8	Researchcollabor...
d37dbdbf4b89959...	Irene	Villa		teasto0008628_9 resto0008628_10	Teachingassistant... Researchfellowshi...
e90f3d5f81ed448...	Jing	Zhang		phdsto0013634_11	PhDstudentSTO0...
2d64e2fa7cf97b3c...	Paola	Giacopelli	0461 282837	stasto0000863_12	StaffSTO0000863
b041408b841fcf8c...	Leonardo	Franceschini		stasto0008911_13	StaffSTO0008911
2f28495f31b1f3f...	Giulia	Citroni		stasto0013543_14	StaffSTO0013543
b3510f7ca84dcba...	Andrea	DallapE		phdsto0002101_15	PhDstudentSTO0...
fd526339790f5c9...	Rudj	Gorian	0461 283185	stasto0001641_16	StaffSTO0001641

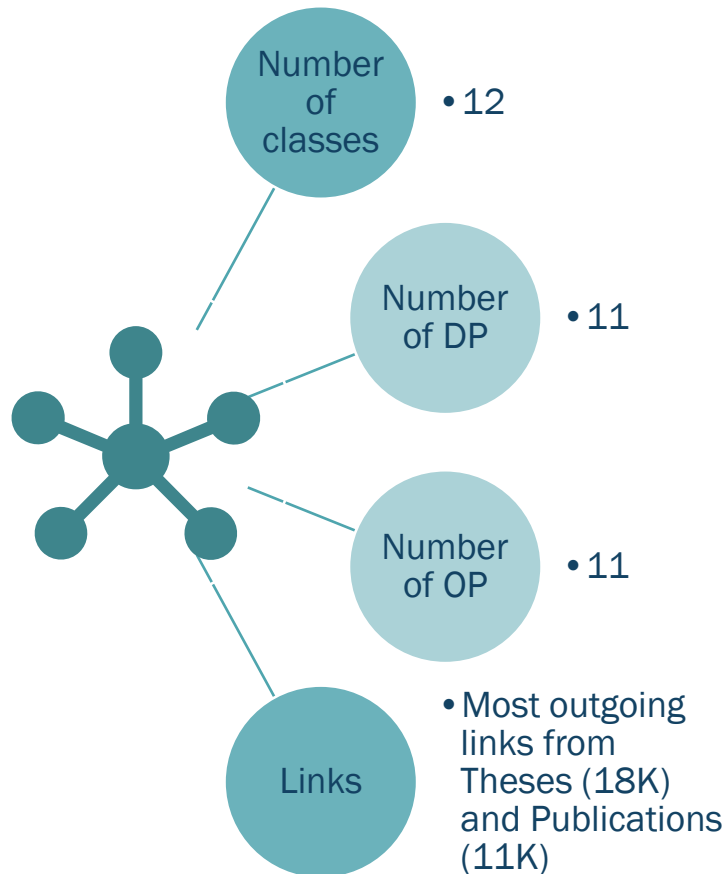


A knowledge graph diagram showing a central node 'Publica...ID-3000011' connected to several other nodes: 'has_pub...ype-300001', 'has_author_GID-58221', 'has_rep...ype-300001', 'uri', and 'File_GID-3000031'. The 'uri' node is further connected to a 'uri' node.

titolo	tipo	anno	lingua	autori	citazioni	file	pub_id	filematch
Competition, land prices and city size	articolo su rivista	2020	inglese	9bb1a60bad1db5...	Kichko, S., "Competition, land prices and city size" in	https://iris.unitn.it/r...	pub1	KichkoJoEGpdf
Multiscale simulation of the focused electron beam induced deposition	articolo su rivista	2020	inglese	martinaazzolini_1 6258df5bc7ef98...	de Vera, P.; Azzolini, M.; Sushko, G.; Abril, I.; Garcia-Molina, R.; Dapor,	NOT_PRESENT	pub2	NOT_PRESENT
L'altra pedagogia di Rosmini: dilemmi, occultamenti, traduzioni	articolo su rivista	2020	italiano	b6d16c8768f069c...	Marangon, Paolo. Recensione a: Bonafede, P., "L'altra	https://iris.unitn.it/r...	pub3	RecBonafedeAnn...
Beyond Born-Oppenheimer approximation in ultracold atomic	articolo su rivista	2020	inglese	7a7f7fa87ff9d387...	Tiemann, E.; Gersema, P.; Voges, K. K.; Hartmann, T.; Zenesini, A.,	NOT_PRESENT	pub4	NOT_PRESENT
The sex factor: How women made the West rich Victoria Bateman Medford,	articolo su rivista	2020	inglese	bad30fbc3dff95f97...	Cuel, Roberta. Recensione a: Victoria Bateman, "The sex factor:	https://iris.unitn.it/r...	pub5	gwao12446pdf

Data Integration. Knowledge Graph Construction

Evaluation. Some statistics



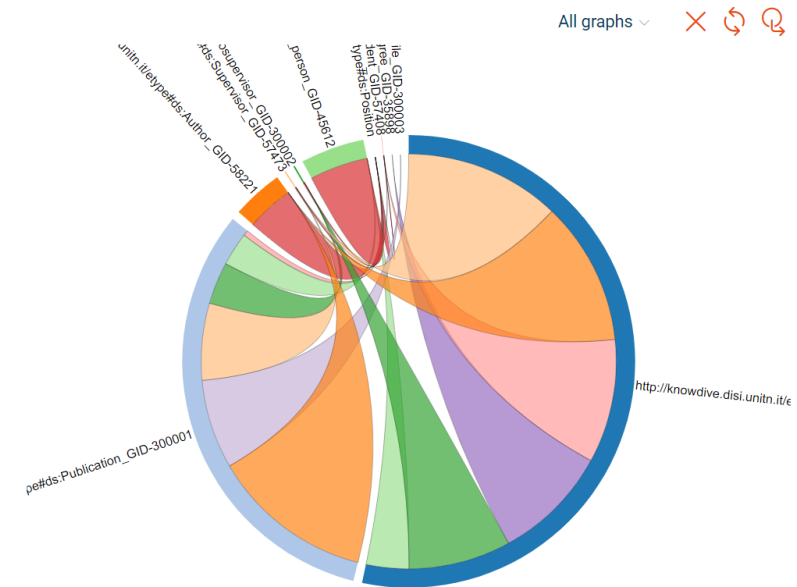
Class relationships ⓘ

Showing the dependencies between 10 classes

Filter classes

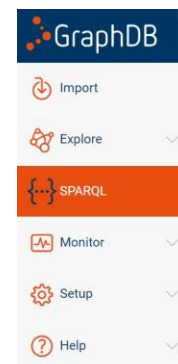
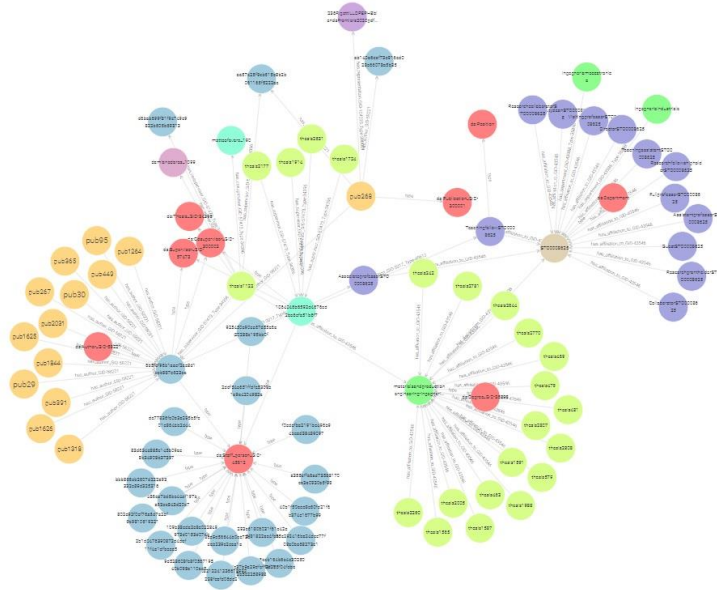
☒ All ☐ Incoming ☐ Outgoing

Class	Links	
http://knowdiver.disi.unitn.it/etypeds:Thesis_GID-34396	33K	→
http://knowdiver.disi.unitn.it/etypeds:Publication_GID-300001	19K	→
http://knowdiver.disi.unitn.it/etypeds:Author_GID-58221	16K	=
http://knowdiver.disi.unitn.it/etypeds:Supervisor_GID-57473	11K	=
http://knowdiver.disi.unitn.it/etypeds:Cosupervisor_GID-300002	7K	=
http://knowdiver.disi.unitn.it/	6K	=



GraphDB

Visual graph ⓘ



```
1 PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
2 PREFIX ds: <http://knowdiver.disi.unitn.it/etyp#>
3 PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
4 select ?deg (count(?thesis) as ?counts) where {
5   ?thesis rdf:type <http://knowdiver.disi.unitn.it/etyp#ds:Thesis_GID-34396> ;
6   ds:has_thesis_degree_GID-35789_Type-34396 ?deg.
7 } group by ?deg order by desc (?counts)
8
```

Table Raw Response Pivot Table Google Chart Download as

Filter query results Showing results from 1 to 67 of 67. Query took 0.1s, moments ago.

	deg	counts
1	"Laurea Magistrale Ciclo Unico 5 anni - Giurisprudenza"	"587""xsd:integer
2	"Corso di Laurea Magistrale - Management"	"160""xsd:integer
3	"Corso di Laurea Magistrale - Psicologia"	"121""xsd:integer
4	"Corso di Laurea Magistrale - INFORMATICA"	"120""xsd:integer
5	"Corso di Laurea Magistrale - Ingegneria Meccatronica"	"105""xsd:integer
6	"Corso di Laurea Magistrale - MATEMATICA"	"104""xsd:integer

Let's see a demo of SPARQL queries!

The background is a complex network diagram. It features a dense web of thin, light gray lines connecting various nodes. The nodes are represented by circles of different sizes and colors, including blue, green, pink, and purple. Some nodes have internal patterns like stripes or concentric circles. The overall layout is organic and interconnected, suggesting a global or digital network.

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