



UNIVERSITY
OF TRENTO - Italy



Dipartimento di Ingegneria e Scienza dell'Informazione

– KnowDive Group –

KGE 2022 - Digital University UniTn. Theses, Publications and Staff

Document Data:

December 18, 2022

Reference Persons:

Milena Atanasova

© 2022 University of Trento

Trento, Italy

KnowDive (internal) reports are for internal only use within the KnowDive Group. They describe preliminary or instrumental work which should not be disclosed outside the group. KnowDive reports cannot be mentioned or cited by documents which are not KnowDive reports. KnowDive reports are the result of the collaborative work of members of the KnowDive group. The people whose names are in this page cannot be taken to be the authors of this report, but only the people who can better provide detailed information about its contents. Official, citable material produced by the KnowDive group may take any of the official Academic forms, for instance: Master and PhD theses, DISI technical reports, papers in conferences and journals, or books.



Index:

1	Introduction	1
2	Purpose and Domain of Interest (DoI)	1
3	Data Sources	2
4	Purpose Formalization	8
4.1	Phase Description	8
4.2	Scenarios, Personas, Competency Questions and Entities	8
4.3	Metadata Purpose Formalization Sheet	14
5	Inception	14
5.1	Phase Description	14
5.2	Data Management	15
5.3	Some exploration of the datasets defined	17
5.4	Creation of Ontologies Dataset by Dataset	21
5.5	Connecting data with ontology	22
6	Informal Modeling	27
6.1	Phase Description	27
6.2	Initial considerations	27
6.3	Schema Level E-type	28
6.4	Description of ER Model	31
6.5	Implementation of Informal Modeling in Protege	35
6.6	Metadata	37
7	Formal Modeling	38
7.1	Phase Description	38
7.2	Ontology	39
7.3	Ontology Implementation in Protege	42
7.4	Teleontology	43
7.5	Teleontology Implementation in Protege	46
7.6	Language Annotation	47
7.7	Language Annotation Implementation	48
8	KGC	51
8.1	Entity Identification. Semantic heterogeneity	51
8.2	Entity Matching	52
8.3	Implementation using the data mapping tool Karma	53
8.4	KG's evaluation	54
8.5	GraphDB Implementation	55

9 Outcome Exploitation	56
10 Conclusions & Open Issues	62

Revision History:

Revision	Date	Author	Description of Changes
0.1	10.11.2022	Milena Atanasova	Document created
0.2	11.11.2022	Milena Atanasova	Purpose and Dol
0.3	11.11.2022	Milena Atanasova	Data sources added
0.4	12.11.2022	Marco Zardo	Scenarios, personas, CQs from inception sheet
0.5	12.11.2022	Milena Atanasova	Inception initial part; entities refinement
0.6	14.11.2022	Milena Atanasova	Metadata tables, knowledge resources
0.7	14.11.2022	Milena Atanasova	Data sources added
0.8	14.11.2022	Milena Atanasova	Images, tables in Inception part; small refinement
0.9	20.11.2022	Milena Atanasova	ER diagram added; modified Inception; added metadata from Karma
0.10	20.11.2022	Marco Zardo	Description ER diagram; description of entities Informal modelling
0.11	23.11.2022	Marco Zardo	Modifying competency questions and scenarios
0.12	22.11.2022	Milena Atanasova	Added images from Protege teleology
0.13	23.11.2022	Milena Atanasova	More general ER diagram; expanded description of ER models
0.14	24.11.2022	Milena Atanasova	Metadata from Shapeness added
0.15	29.11.2022	Milena Atanasova	Graphs for Ontology and Teleontology for Formal Modeling
0.16	01.12.2022	Milena Atanasova	Added ontology, teleontology and metadata in Formal Modeling
0.17	03.12.2022	Marco Zardo	Description teleontology
0.18	04.12.2022	Milena Atanasova	More detailed overview of entities, schemas and hierarchies for teleontology
0.19	06.12.2022	Milena Atanasova	Datasets exploration; Language annotation
0.20	09.12.2022	Milena Atanasova	Annotation table
0.21	11.12.2022	Milena Atanasova	KGC
0.22	14.12.2022	Milena Atanasova	GraphDB examples, KGC
0.23	17.12.2022	Milena Atanasova	Conclusions, open issues
0.24	18.12.2022	Milena Atanasova	Final modifications

1 Introduction

This project aims to construct a Knowledge Graph about the Digital University of Trento, with a focus on research activities conducted inside the university. It is built following the iTelos methodology which is an innovative practice to follow and consider when designing projects of this kind.

Reusability is one of the main principles in the Knowledge Graph Engineering (KGE) process defined by iTelos. The KGE project documentation plays an important role in order to enhance the reusability of the resources handled and produced during the process. A clear description of the resources and the process developed, provides a clear understanding of the KGE project, thus serving such an information to external readers in order to exploit that in new projects.

The current document aims to provide a detailed report of the KGE project developed following the iTelos methodology. The report is structured, to describe:

- Section 2: The project's purpose and the domain of interest and the resources involved (both schema and data resources) in the integration process.
- Section 2: The input resources considered by the KGE project.
- Section 4, 5, 6, 7: The integration process along the different iTelos phases, respectively.
- Section 8: How the result of the KGE process (the KG) can be exploited.
- Section 9: Conclusions and open issues summary.

All the files created and artifacts collected during the execution of the project can be found at the *github* repository for the project¹.

2 Purpose and Domain of Interest (DoI)

This project concerns the area of the Digital University of the University of Trento, which is a virtual place where information about various areas supporting students, professors, and others interested in retrieving data about theses, publications and staff. These resources should be integrated in such an environment in order to provide a fluid user experience in querying data about the resources available and receiving the right response in an effective and efficient way.

Therefore, this Knowledge Graph should be able to satisfy the needs of all stakeholders involved by collecting data from different sources and integrating them. In such a way, this service built would prove helpful to users in finding information about academics and research at the University of Trento, along with some potential data about personas involved with this educational institution. The main purpose of this project as expressed by the final user is the following:

"A service which helps the users to query and know about the different areas of academic and research interest being pursued at the University of Trento."

¹<https://github.com/atanasova16/TrentinoDUniTPS>

The main keywords identified here are: different areas, academic, research, UniTn.
Having in mind this informal purpose, the specific Domain of Interest (DoI) was defined as :
Academics at the University of Trento, including information mainly about research activities, between the years 2020 and 2022.

3 Data Sources

The input resources for building a Knowledge Graph are divided in two main groups: Knowledge and Data. The ones reviewed for this specific project considering the purpose of the final user are the following:

- **Knowledge sources:** The sources for reference schemas and ontologies initially collected to satisfy the purpose along the KGE process.
 - VIVO (for modelling academics and research). In VIVO there are several classes that correspond to the information at hand: Research Project, Research Opportunity, Student Role, Employee Role, Staff Role, Faculty Role, Licensing, Document, Journal, Academic Degree, Academic Department, Postdoctoral Position, Student
 - AIISO (for modelling academic institution). Classes connected to this purpose: Course, Department, Faculty, Organizational Unit
 - FOAF (for modelling persons and their relations). Classes: Document, Person
 - schema.org (for modelling common concepts). Includes: Thesis, Article, Academic degree, Academic department
 - DCAT2 (for modelling metadata)

Along the process of building the knowledge graph, some of these Knowledge sources were filtered and dropped, and those different ontologies were combined with the aim of becoming of a greater use for the specific project. More on this issue is done in section 7 - Formal modeling.

- **Data sources:** The sources for datasets initially collected to satisfy the purpose along the KGE process.

Data sources come from two main providers: Trentino Open Data Portal and UniTrento Digital University portal. Following below are the distributions:

- Theses of the University of Trento (Tesi di laurea dell'Università di Trento 2020-2022)
- Publications of the University of Trento (Pubblicazioni dell'Università di Trento 2020, 2021, 2022)
- Staff at the University of Trento
- Courses, degree programs and departments at the University of Trento (Insegnamenti)
- Digital University Unitn, page about the departments and some basic information about them

- **Metadata**

The datasets gathered through the resource collection phase have the attributes showed below.

As we distinguish between dataset (where the resources are located), and distribution (the specific resource), below are listed the metadata for both.

First, the datasets exhibit the following metadata.

Theses dataset	
Dataset identifier	UNSTTREN:b6fa0ce8-4cd0-4c7c-875f-6b9a93a82e3c
Themes	Education, culture and sport
Dataset editor	University of Trento
Release date	20.2.2020
Geographic scope	Trento
URI of GeoNames	https://www.geonames.org/3165241
Language of dataset	English

Publications dataset	
Dataset identifier	UNSTTREN:e1b609a2-072e-4bd5-96f6-449bd340ae2a
Themes	Education, culture and sport
Dataset editor	University of Trento
Release date	20.2.2020
Geographic scope	Trento
URI of GeoNames	https://www.geonames.org/3165241
Language of dataset	English

Staff dataset	
Dataset identifier	UNSTTREN:1a9b06d9-b56c-499a-be5b-f22fa9bcceaf
Themes	Education, culture and sport
Dataset editor	University of Trento
Release date	20.2.2020
Geographic scope	Trento
URI of GeoNames	https://www.geonames.org/3165241
Language of dataset	English

Courses and degrees dataset	
Dataset identifier	UNSTTREN:138a0046-5a87-4180-bc7a-9f0d75d0547b
Themes	Education, culture and sport
Dataset editor	University of Trento
Release date	15.4.2019
Geographic scope	Trento
URI of GeoNames	https://www.geonames.org/3165241
Language of dataset	Italian

Next are the metadata concerning the concrete distributions collected from Open data Trentino portal that are considered for the specific purpose.

Theses distribution	
Format	JSON
License	Creative Commons Attribuzione 4.0 Internazionale (CC BY 4.0) LINK
Created	2021-05-11T14:45:12.248045
package_id	1504318d-7d56-4098-a93f-5061a34322cc
id	6e41ca07-c5c2-45a4-b427-379ae6977248
Language	italiano [it]
Description	Graduation theses from University of Trento, of which at least one of the supervisors still works at the university. Considered are theses for master programs. Fields: title; student(name, surname, id); course of study; year; supervisors; cosupervisors; examiners
url	https://dati.unitn.it/du/Thesis/2020/2022
license_type	https://w3id.org/italia/controlled-vocabulary/licences/A21_CCBY40

Publications distribution	
Format	JSON
License	Creative Commons Attribuzione 4.0 Internazionale (CC BY 4.0) LINK
Created	2022-04-28T13:28:51.384241
package_id	fa446d1f-7aa2-42c0-aca9-7a98ce17d461
id	eb6a441e-6981-4608-ba5d-53f3116a5904
Language	italiano [it]
Description	Publications from UniTn, that are preserved in the archive of the university. Fields: type; title; year; authors; file (name, link, format, version, license, openAccess)
url	https://dati.unitn.it/du/MindProduct/2021
license_type	https://w3id.org/italia/controlled-vocabulary/licences/A21_CCBY40

Staff distribution	
Format	JSON
License	Creative Commons Attribuzione 4.0 Internazionale (CC BY 4.0) LINK
Created	2020-02-20T15:21:29.305084
package_id	c14f7615-9d8a-43b2-abd9-3c90f3b62de0
id	57720d62-a7d1-47a4-b6d6-afb76fd27e88
Description	Staff of the University of Trento. Fields: name; surname; identifier; phone; positions (role, unitName, unitId)
url	https://dati.unitn.it/du/Person/en
license_type	https://w3id.org/italia/controlled-vocabulary/licences/A21_CCBY40

Courses and degrees distribution	
Format	JSON
License	Creative Commons Attribuzione 4.0 Internazionale (CC BY 4.0) LINK
package_id	af70e1fd-5ac6-4c4f-9625-dbfefc5d93a8
id	3b3bc3c8-de0a-4f00-9a40-c73dba8574b4
language	italiano [it]
Description	Courses offered at UniTn from the Digital University portal. Fields: name; description; department, website; professors; assistents
url	https://dati.trentino.it/dataset/af70e1fd-5ac6-4c4f-9625-dbfefc5d93a8/resource/3b3bc3c8-de0a-4f00-9a40-c73dba8574b4/download/2022-04-28insegnamentiit.json
license_type	https://w3id.org/italia/controlled-vocabulary/licences/A21_CCBY40

Metadata was also defined using the SHAPEness software. The file regarding it can be accessed at this link.

```
@prefix : <https://www.epos-eu.org/epos-dcat-ap#> .
@prefix schema: <http://schema.org/> .
@prefix spdx: <http://spdx.org/rdf/terms#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix gsp: <http://www.opengis.net/ont/geosparql#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix dqv: <http://www.w3.org/ns/dqv#> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix hydra: <http://www.w3.org/ns/hydra/core#> .
@prefix geo: <http://www.w3.org/2003/01/geo/wgs84_pos#> .
@prefix oa: <http://www.w3.org/ns/oa#> .
@prefix dct: <http://purl.org/dc/terms/> .
@prefix sh: <http://www.w3.org/ns/shacl#> .
@prefix dcat: <http://www.w3.org/ns/dcat#> .
@prefix locn: <http://www.w3.org/ns/locn#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix epos: <https://www.epos-eu.org/epos-dcat-ap#> .
@prefix adms: <http://www.w3.org/ns/adms#> .
@prefix org: <http://www.w3.org/ns/org#> .
@prefix cnt: <http://www.w3.org/2011/content#> .
@prefix vcard: <http://www.w3.org/2006/vcard/ns#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix http: <http://www.w3.org/2006/http#> .
@prefix dash: <http://datashapes.org/dash#> .
@prefix dc: <http://purl.org/dc/elements/1.1/> .

<https://www.epos-eu.org/epos-dcat-ap#Distribution/2928200a-4fad-4e47-88d1-50cac6934eb0>
    rdf:type      dcat:Distribution ;
    dct:description "Courses offered at University of Trento with professors and degree programs" ;
    dct:format    "JSON" ;
    dct:identifier "91c54d6f-ddce-480e-9214-216f23939c51" ;
    dct:issued    "2022-07-04T22:05:12Z^^xsd:dateTime" ;
    dct:language   "Italian" ;
    dct:license    "Creative Commons Attribuzione 4.0 Internazionale (CC BY 4.0) LINK" ;
    dct:title     "Insegnamenti" ;
    dcat:accessURL "https://dati.trentino.it/dataset/insegnamenti-universita-di-trento/resource/91c54d6f-ddce-480e-9214-216f23939c51" ;
    dcat:downloadURL "https://dati.unitn.it/du/Course"^^xsd:anyURI .

<https://www.epos-eu.org/epos-dcat-ap#Dataset/ec1fa3b0-e24e-4eb6-a3a5-e344c780fee1>
    rdf:type      dcat:Dataset ;
    dct:created    "2020-02-20T12:00:00Z"^^xsd:dateTime ;
    dct:description "Theses from the university of Trento" ;
    dct:identifier "UNSTTREN:b6fa0ce8-4cd0-4c7c-875f-6b9a93a82e3c"^^xsd:anyURI ;
    dct:publisher  <https://www.epos-eu.org/epos-dcat-ap#Organization/c1f39d45-f7bb-45c7-8b02-8b6a0cff0402> ;
    dct:title      "Theses of the University of Trento" ;
    dcat:distribution <https://www.epos-eu.org/epos-dcat-ap#Distribution/4bd1c8f9-d70d-4239-8da0-f9ed8f6bd406> ;
```

dcat:theme <<https://www.epos-eu.org/epos-dcat-ap#Concept/8e43526c-938a-4e4d-bea7-8f8d63673d7c>> .

<<https://www.epos-eu.org/epos-dcat-ap#Organization/c1f39d45-f7bb-45c7-8b02-8b6a0cff0402>>

rdf:type schema:Organization ;
schema:identifier "UNSTREN"^^xsd:anyURI .

<<https://www.epos-eu.org/epos-dcat-ap#Dataset/e1c79fbf-6c2f-425b-aba2-c9b9970eb228>>

rdf:type dcat:Dataset ;
dct:description "Publications of the University of [Trento](#), where at least one author currently works in the University (at the date of modification of the dataset). Publications stored in the IRIS archive of the University research database are reported" ;
dct:identifier "UNSTREN:e1b609a2-072e-4bd5-96f6-449bd340ae2a" ;
dct:issued "2020-02-20T21:40:59Z"^^xsd:dateTime ;
dct:publisher <<https://www.epos-eu.org/epos-dcat-ap#Organization/c1f39d45-f7bb-45c7-8b02-8b6a0cff0402>> ;
dct:title "Publications of the University of [Trento](#)" ;
dcat:distribution <<https://www.epos-eu.org/epos-dcat-ap#Distribution/cb21ae7a-f32e-4ba1-a763-6d3a4e9dcfdb>> ;
dcat:theme <<https://www.epos-eu.org/epos-dcat-ap#Concept/8e43526c-938a-4e4d-bea7-8f8d63673d7c>> .

<<https://www.epos-eu.org/epos-dcat-ap#Dataset/62029e4f-d558-46fd-93df-b030a7435088>>

rdf:type dcat:Dataset ;
dct:description "Staff of the University of [Trento](#) (at the date of modification of the dataset). Also include guests and temporary staff. For each person, positions held in each unit and contact information are indicated." ;
dct:identifier "UNSTREN:1a9b06d9-b56c-499a-be5b-f22fa9bcceaf" ;
dct:issued "2020-02-20T21:53:02Z"^^xsd:dateTime ;
dct:publisher <<https://www.epos-eu.org/epos-dcat-ap#Organization/c1f39d45-f7bb-45c7-8b02-8b6a0cff0402>> ;
dct:title "Staff of the University of [Trento](#)" ;
dcat:distribution <<https://www.epos-eu.org/epos-dcat-ap#Distribution/fb7437d7-8ffd-4b4a-85e5-c0bb6a165dfd>> ;
dcat:theme <<https://www.epos-eu.org/epos-dcat-ap#Concept/8e43526c-938a-4e4d-bea7-8f8d63673d7c>> .

<<https://www.epos-eu.org/epos-dcat-ap#Distribution/4bd1c8f9-d70d-4239-8da0-f9ed8f6bd406>>

rdf:type dcat:Distribution ;
dct:description "Graduation theses from University of [Trento](#) 2020 to 2022" ;
dct:identifier "6e41ca07-c5c2-45a4-b427-379ae6977248" ;
dct:issued "2021-05-11T14:45:12Z"^^xsd:dateTime ;
dct:language "Italian" ;
dct:license "Creative Commons [Attribuzione 4.0 Internazionale \(CC BY 4.0\) LINK](#)" ;
dct:title "Tesi di laurea (2020-2022)" ;
dcat:accessURL "<https://dati.trentino.it/dataset/tesi-di-laurea-universita-di-trento/resource/6e41ca07-c5c2-45a4-b427-379ae6977248>" ;
dcat:downloadURL "<https://dati.unitn.it/du/Thesis/2020/2022>"^^xsd:anyURI .

<<https://www.epos-eu.org/epos-dcat-ap#Dataset/a9947264-9880-4528-8123-ae0f2e533ef8>>

```

        rdf:type      dcat:Dataset ;
        dct:description "Courses offered by the University of Trento. The dataset is updated annually, providing the complete list of courses for each academic year (at the date of modification of the dataset).";
        dct:identifier "UNSTTREN:f2b30f8d-5d87-47a9-b3e9-60e412928518" ;
        dct:issued    "2020-02-20T22:00:15Z"^^xsd:dateTime ;
        dct:language   "English" ;
        dct:publisher <https://www.epos-eu.org/epos-dcat-ap#Organization/c1f39d45-f7bb-45c7-8b02-8b6a0cff0402> ;
        dct:title     "Courses and degree programs offered by the University of Trento" ;
        dcat:distribution <https://www.epos-eu.org/epos-dcat-ap#Distribution/2928200a-4fad-4e47-88d1-50cac6934eb0> ;
        dcat:theme    <https://www.epos-eu.org/epos-dcat-ap#Concept/8e43526c-938a-4e4d-bea7-8f8d63673d7c> .

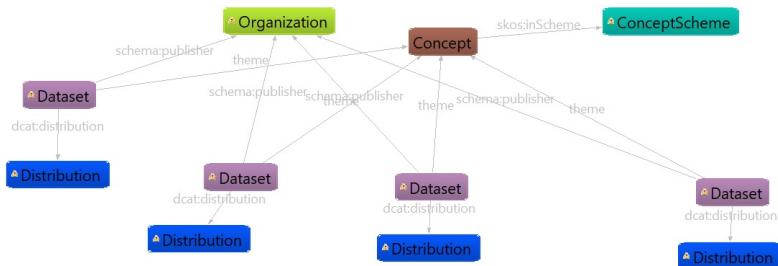
<https://www.epos-eu.org/epos-dcat-ap#Concept/8e43526c-938a-4e4d-bea7-8f8d63673d7c>
    rdf:type      skos:Concept ;
    skos:inScheme <https://www.epos-eu.org/epos-dcat-ap#ConceptScheme/81adcd41-5389-477c-b2b6-2527c374231f> .

<https://www.epos-eu.org/epos-dcat-ap#ConceptScheme/81adcd41-5389-477c-b2b6-2527c374231f>
    rdf:type      skos:ConceptScheme ;
    dct:title     "Education, culture and sport" .

<https://www.epos-eu.org/epos-dcat-ap#Distribution/cb21ae7a-f32e-4ba1-a763-6d3a4e9dcfdb>
    rdf:type      dcat:Distribution ;
    dct:description "Publications of the University of Trento 2022" ;
    dct:format    "JSON" ;
    dct:identifier "aa43ef8c-e9bf-4648-b586-a816ca5337c5" ;
    dct:issued    "2022-04-28T21:47:09Z"^^xsd:dateTime ;
    dct:language   "Italian" ;
    dct:license    "Creative Commons Attribuzione 4.0 Internazionale (CC BY 4.0) LINK" ;
    dct:title     "Pubblicazioni (2022)" ;
    dcat:accessURL "https://dati.trentino.it/dataset/pubblicazioni-universita-di-trento/resource/aa43ef8c-e9bf-4648-b586-a816ca5337c5" ;
    dcat:downloadURL "https://dati.unitn.it/du/MindProduct/2022"^^xsd:anyURI .

<https://www.epos-eu.org/epos-dcat-ap#Distribution/fb7437d7-8ffd-4b4a-85e5-c0bb6a165dfd>
    rdf:type      dcat:Distribution ;
    dct:description "Staff of the university of Trento" ;
    dct:format    "JSON" ;
    dct:identifier "57720d62-a7d1-47a4-b6d6-afb76fd27e88" ;
    dct:issued    "2020-02-20T21:57:20Z"^^xsd:dateTime ;
    dct:license    "Creative Commons Attribuzione 4.0 Internazionale (CC BY 4.0) LINK" ;
    dct:title     "Staff of the University of Trento" ;
    dcat:accessURL "https://dati.trentino.it/dataset/staff-of-the-university-of-trento/resource/57720d62-a7d1-47a4-b6d6-afb76fd27e88" ;
    dcat:downloadURL "https://dati.unitn.it/du/Person/en"^^xsd:anyURI .

```



4 Purpose Formalization

4.1 Phase Description

After having in mind the purpose and DoI, the next step of the utilized Knowledge graph engineering methodology is the formalization of the purpose of the project. It could be a crucial point of all the process since the final results would be affected by clarity and characterization of the project purpose. Purpose formulation involves definition of Scenarios, Personas, Competency questions (CQs) and Entities identified.

4.2 Scenarios, Personas, Competency Questions and Entities

- **Scenarios:**

In this section, the intention is to define the possible scenarios in which there would be personas who could interact with the outcome of this project. They also try to give a time and place for the usage cases.

I Spring, school year, familiar with Unitn, Italy. The time when high-school students are searching for some information about the courses of study, academic programs and some detail about the University.

II Spring, school year, already inside Unitn, Italy. At the beginning of the year, some researchers explore papers or articles on different fields to find out new goals of research. Moreover they ask some information to the professors of the departments for further requirements.

III Autumn, beginning of school year, already inside Unitn, Trento, Italy. At the beginning of academic year, professors, following the programs of our department, start to organize some didactic experiences for students who have done the final part of their degree and would like to begin a postgraduate experience.

IV Summer, before start of school year, not familiar with Unitn, outside Italy. During the summer period, foreign students decide to start a bachelor experiences abroad. Moreover, they would like to enrich their personal project and academic career studying in other parts of the world.

V Summer, before start of school year, familiar with Unitn, outside Italy. Editors and journalists of specialized magazines are in search for papers and their related authors to discuss to them about their results. Moreover the interviews gathered on these events will be published as a "summer edition" magazine.

VI Winter, school year, familiar with Unitn, Italy. At the end of the year Investors and communicators organize conferences and open events to talk about the innovative research promoted in this last year. They decide to invite the highly-recognized researchers to explain the progress in various scientific fields.

VII Autumn, beginning of school year, not familiar with Unitn, Italy. When the project's programs are formalized by universities, professors are interested to provide new collaborations with other institutes in order to expand the efforts in some scientific fields and obtain



new results and prediction of them. They would like to start these collaborations the next academic year on September.

VIII Spring, school year, familiar with Unitn, Italy. At the end of the year the companies explore publications and articles to find out new applications or strategies for their purposes. They look for new information, in order to enforce them at the beginning of the new year.

IX Summer, before start of the school year, not familiar with Unitn, Italy. Before the beginning of the academic year families are interested to see what could be the possible opportunities that their children might start after the master's degree. They start to gather some information about the recent programs of research and the departments of some institutes.

X Spring, school year, not familiar with Unitn, outside Italy (EU). Following the new funds arranged by European Union, ambassadors extend the project of universities abroad including some others European countries and Institutes. They take into account the PhD publications and experimental thesis of last years. The funding for these collaborations will take place at the beginning of the next year.

XI Autumn, beginning of school year, inside Unitn, Italy. The post-doc academics, during last part of their research-project are seeking an internship experience abroad connected with some thesis or scientific work. They are available to perform some post-analysis, starting to some publications of the last years. However they prefer to start this new experience immediately with the beginning of the new academic year.

- **Personas:**

Personas are the possible users of the Knowledge Graph project outcome: they are supposed to exploit the final result thanks to this work. This section in particular defines different types of person with their different needs. The UniTn website might be interpreted as the common center in which all kind of users have searched or are going to search even if following different paths the overall results are different for each ones.

Marco Marco Pierotti, 19 (a high-school student at Bassano del Grappa in his final year, indecisive of what discipline to study and where to go to University).

Elena Elena Ranzani, 25 (an assistant for already 3 years at Unitn who wants to do now a PhD and continue her research in biology).

Donatello Donatello Ferrari, 63 (a full-time professor at Unitn who is looking for students/alumni from the Psychology department to help him).

Ivan Ivan Krumov, 21 (a Bachelor student in Bulgaria, who wants to go to Italy being strongly interested in AI to begin a Master taught in English, and find help for a project he has started by himself).

Ginnie Ginnie Anderson, 32 (a British editor in a new innovative Maths journal, whose job is to exploit different European universities' math-related papers and potentially sending invitations to the authors to provide material to their journal in English).

Paolo Paolo Lanza, 44 (an investor from Parma highly passionate about scientific research who would also like to organize a physics conference).

Francesco Francesco Tiribocchi, 55 (full-time professor at UniBo, who is seeking a Machine Learning or Artificial Intelligence PhD/research in order to create a collaboration with the DISI department).

Guido Guido Domeneghetti, 34 (an external partner of UniTn who works in a company high-connected with university activities).

Giuseppe Giuseppe Ruggero, 49 (a parent of a bachelor student in Rome, who want to know about the fields of research at UniTn).

Jane Jane Muller, 31 (an European Ambassador who is seeking collaboration with UniTn).

Serena Serena Colla, 27 (a post-doc student of UniTn who wants to make a internship experience).

- **Competency Questions (CQs):**

I The father of Marco suggests him to look at what faculties are there at the University of Trento, if they are close to the city center or not, how popular and recognized in Italy they are and if he finds something that interests him, to contact and ask current students/alumni/professors.

- What are the courses of study offered by the University of Trento?
- How are courses of study grouped in faculties? Which are the most prominent ones?
- What kinds of department are there?
- Are the department close or distant to the center of the city?
- How recognized is the level of academic pathway of Trento with respect to the other universities?

II Elena would like to get in touch with several professors who work in the biological faculty to ask for possibilities and requirements, and to search for different papers in the sector in order to gain inspiration for her own research during the potential PhD, as well as to see the important journals who publish them and try send her already written works to them for publishing.

- Which department could offer an important opportunity to research in biological field?
- Are there some publication or article which could help on her PhD project?
- Are there groups of research or authors to work with?
- Are there experimental theses to expand the research with?

III Prof. Ferrari would like to offer internship opportunities and possibly offers for PhD for students who have already done their theses in psychology, and would like to contact them personally.

- Which are the theses published under the Psychology department?
- Are there some publications connected to these works?
- Who are the advisors and co-advisors involved in these activities?
- Where is it possible to find out the emails of the students?

-
- Are there any theses made in collaboration with other departments?

IV Ivan explores publications on his topic of interest and reads a very nice one written from a professor at Unitn. He is now eager to check other recent publications of the author or similar that are in English, find his contact and ask him if he could find help for his project, and do a master in AI in English at UniTn.

- Which are the publications about Artificial Intelligence published in the last year?
- Are there theses or other articles related to them?
- Who are the staff from UniTn involved on these project?
- How are the Master degrees organized at UniTn?
- Is there some information about the department responsible of this didactic offer?

V Miss Anderson is exploring different mathematical articles, conferences and seeks contributors.

- Which are the crucial publications of this year regarding Mathematics and its applications?
- Who are the authors and students involved in these types of project?
- Which are the departments' heads of these academic experiences?
- Are there other people indirectly involved in this organization?
- Are there experimental theses connected with these studies?

VI Mister Lanza would like to talk with organizational units staff, the rector and directors in order to propose to become an investor for the Physics department, as well as to invite highly-recognized professors to take part in the international conference he is organizing.

- Who are the staff persons involved in the Department of Physics?
- Which are the experimental theses or publications highly-recognized in the last year?
- Who are the professors and advisors involved in these project?
- Where it is possible to find out information about the directional staff of the Physics Department?

VII Prof. Tiribocchi would like to know about some PhD or project performing by DISI department of the University of Trento in order to organize and provide some collaboration to expand the research. In addition he is searching the e-mail of the researches and professors involved in that projects.

- Which are the last research projects organized by DISI department?
- Is there any information of this department? Where is geographically located?
- Who are the advisors involved in the last project?
- Are there any phone-numbers or e-mails to contact this Department?

VIII Guido explores publications about the last research in a certain domain in order to find out applications that could be implemented on his company; he wants to use new kinds of tested treatments to enhance the productivity of his employees.

- Are there some publications or articles useful for the company?
- What are the departments responsible of these project?
- Who are the directors? Who are the professors or student involved in them?
- Is there any way to contact them?
- Are there any publications available for not academical persons?

IX Giuseppe wants to know about PhD/internship offered by the departments of University of Trento, that's because his son has an intention to attend one of these programs; moreover he wants to take information about the professors involved in that project and some general details about the departments and their goals on the research.

- What are the Departments which organized these activities?
- Who are the professors responsible of the PhD projects?
- Where are the departments located?
- Which is the last research taken into account by Unitn?
- Are there courses of study related with more departments?

X Miss Muller, following an European Project, is seeking experimental theses and others project related to PhD publications in order to extend the research abroad including some others European countries and universities, starting to particular topic seen and studied by UniTn students; she is interested to understand who are the authors of these publications and which are the departments involved into it and their directors.

- Which are the most attractive projects and publications done by the university?
- Who are the heads of these projects?
- Are they associated with any department?
- Who are the directors of these departments? What are their contacts?
- Are there any theses which deal with the concepts of those projects?
- Are there the contact of those involved on these theses works?

XI Serena would like to make an internship experience after her post-doc on Biological Science, so she is seeking a good opportunity offered by departments of UniTn; she is asking for the contact of professors and researchers involved in PhD or thesis-projects or possibly she is looking for some publications on which to start some post analysis.

- Which are the departments involved in some Biological field?
- Are there any opportunities to continue the research there?
- What kind of research has been developed during the last year? Are there any publications about it?

-
- Who are the advisors, researchers or students involved in these programs? How is it possible to contact them?
 - **Entities identified:** the terms representing the entities to consider in the KGE project, classified using the popularity categories.

Common Entities

- I. Person (with data properties name, surname, some contact information; and with object property such as the occupation the person has)
- II. Functional Unit
- III. Creative work (including in itself publication, thesis, etc.) (with data properties title, time of publication, some kind of subject what is it about; and with object properties referring to author of the specific work, and if available a way to access it)

Core Entities

- I. Student (with data properties name and surname, possibly the course he/she is attending)
- II. Professor
- III. Faculty (Department) (with data properties name of department, address, website, some reference person; with object properties the degrees which are under this department)
- IV. File (an object which identifies a creative work, with data properties such as name, access link, license, type)
- V. Author (with data properties name, surname, possible occupation and contact)

Contextual Entities

- I. Publication (an entity representing specific scientific paper, with data properties such as the title, the year it was published, the language in which it is written; object property would be the specific file if present)
- II. Thesis (with data properties name, year, degree it refers to; object properties would be the student who has written it, as well as the supervisors who would also take part of the University personnel)
- III. Course of Study (Degree Program) (having as data properties name, type, e.g. Bachelor or Master, and object property the department it belongs to)
- IV. Position (role) (further specifying a specific position a person takes, having as data property the name of the specific role/function, and object property in what department it is situated)
- V. Supervisor/Co-supervisor (of a thesis) (an object regarding a Thesis, with again name, surname and possible contact information)

The complete file with a table consisting of the concepts above (Inception sheet) is available on the github repository of the project.

Those are the entities identified at the moment. In subsequent stages and phases of the project development and after further contemplation on the purpose itself, there could be more entities with respective properties which may prove to be useful for the users. However, at the

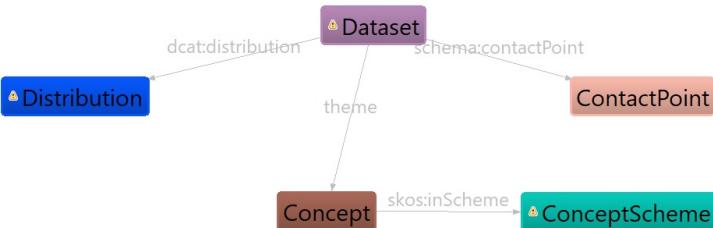


moment for the initial construction of the Knowledge graph, we believe these to be the basis to begin with.

Other entities that could be extracted subsequently would need further considerations about sources from where to collect information about them keeping in mind the specific purpose. Such potential entities and data sources would also need to be verified in terms of representativeness, trustfulness and completeness.

4.3 Metadata Purpose Formalization Sheet

The metadata about the Inception sheet can be accessed here.



```

1 @prefix : <https://www.epos-eu.org/epos-dcat-ap#> .
2
3 @base <https://www.epos-eu.org/epos-dcat-ap#ContactPoint/32b31b21-a58a-4dcf-b2a3-e5fc5f31dc90>
4
5 rdfs:type schema:ContactPoint ;
6 schema:availableLanguage "it" ;
7 schema:contactType "Student" ;
8 schema:email "silena.atanasova@studenti.unitn.it" .
9
10 <https://www.epos-eu.org/epos-dcat-ap#Dataset/4f62bf8f-6014-461e-a123-0b784d82314d>
11
12 rdfs:type dcat:Dataset ;
13 dct:description "Inception sheet for a Knowledge Graph Engineering project following the iTelos methodology" ;
14 dct:identifier "PFSheet-DUTPS_finalversion" ;
15 dct:issued "2023-12-03T14:08:08Z^^xsd:dateTime" ;
16 dct:language "en" ;
17 dct:title "Inception Sheet" ;
18 dct:contactPoint <https://www.epos-eu.org/epos-dcat-ap#ContactPoint/32b31b21-a58a-4dcf-b2a3-e5fc5f31dc90> ;
19 dct:distribution <https://www.epos-eu.org/epos-dcat-ap#Distribution/602d0081-0661-4e7b-b76f-626411d55c22> ;
20 dct:keyword "Inception", "sheet", "iTilos", "purpose formalization", "knowledge graph" ;
21 dct:theme <https://www.epos-eu.org/epos-dcat-ap#Concept/32ad0b01-8fde-430d-9666-ebe9a156f385> .
22
23 <https://www.epos-eu.org/epos-dcat-ap#Concept/3e9d9d81-8fde-430d-9666-ebe9a156f385>
24
25 rdfs:type skos:Concept ;
26 skos:definition "The purpose formalization is the phase when the initial user purpose is being researched to extract important entities for a knowledge graph" ;
27 skos:inScheme <https://www.epos-eu.org/epos-dcat-ap#ConceptScheme/8cceba54-0ccc-4804-b6e5-90c1f1181c91> ;
28 skos:prefLabel "Purpose formalization" .
29
30 <https://www.epos-eu.org/epos-dcat-ap#Distribution/602d0081-0661-4e7b-b76f-626411d55c22>
31
32 rdfs:type dcat:Distribution ;
33 dct:description "An inception sheet including scenarios, personas, competency questions and entities for a KGE project following iTelos. This particular concerns the project Digital University Thesis, Publications, and Staff" ;
34 dct:format "xlsx" ;
35 dct:identifier "PFSheet-DUTPS" ;
36 dct:language "en" ;
37 dct:accessURL "https://github.com/atanasova16/TrentinoDUnitPS/blob/main/Teleologies/Inception/PFSheet_DUTPS_finalversion.xlsx"^^xsd:anyURI .
38
39 <https://www.epos-eu.org/epos-dcat-ap#ConceptScheme/8cceba54-0ccc-4804-b6e5-90c1f1181c91>
40
41 rdfs:type skos:ConceptScheme ;
42 dct:description "The purpose formalization specific for Digital University concept" .
43
  
```

5 Inception

5.1 Phase Description

The step directly following Purpose Formalization is the inception one. It involves initial preparation of individual datasets, as well as ontologies related to those datasets.

- **Input:** a set of data sources identified previously, plus the initial user's Purpose
- **Output:** semi-formal resources

To achieve that different activities were performed. First was the preparation of the data sources: thanks to various actions of collecting, cleaning and formatting the datasets retrieved from Open data Trentino portal, the structure of each dataset was planned following the definition of CQs seen before. These results can be noticed below, under the sub-section “Data Management”.

Once organized in a proper way, the second step was the association of datasets themselves to precise schemas, in order to provide knowledge layer resources for the information shown by data sources. The tool used is Protégé, with which small schemas for each dataset were extracted. This is described below in “Creation of Ontologies Dataset by Dataset” sub-section.

The last part of Inception phase that was performed is about connection of datasets to those schemas. To do that, once each dataset was defined, the Karma data linker was used for each pair of data source associated with its own Ontology to merge them in a single object. The workflow could be exhibited in the section “Connecting data with ontology”.

The development of Inception phase is strongly connected to the Purpose formalization step performed before, since some decisions were made with respect to other ones for enhancing a precise way to proceed. On the other side, it provided a better understanding of the data collected and what schemas were more effective according the datasets considered.

5.2 Data Management

After first initial examination of the resources, it was decided that the time-period to consider would be years 2020-2022. It was thought that such a period would be representative of the recent as well as the current ongoing research that is happening at UniTn on some hot topics that themselves change rapidly (above all in terms of technologies, IT, AI that are constantly changing), while at the same time, have an overview of also works in some other more classical and traditional fields which do not exhibit such a big change from past years to the present. However, it is kept in mind that at later stages it might be useful to enlarge the data layer with sources of information for the years before as well to have a more general overview of what has been studied at the university.

The concrete distributions coming from Open Data Trentino portal were collected in the format they come in - JSON, by downloading the corresponding files. Then, since they all exhibited a nested structure, some modifications needed to be made in order to have a cleaner version of these datasets and then deal with them more efficiently, which would eventually help the reusability.

- For the Staff dataset, it was considered as containing common and core concepts. Each staff person (that could be seen as a Staff Person entity) has its properties and it is in a sense a subclass of Person. Then, from the staff distribution, a new dataset positions containing all the different roles (together with the specific department in which they are situated) was extracted. New identifiers were made for each position and put in both datasets. In fact, at the Data Integration phase, this was modified further.
- About Theses, it is considered Thesis to be a contextual entity for our purpose and so it is quite important in this Knowledge Graph. For each thesis entity there are several fields,

among which such about students, supervisors, co-supervisors and examiners. It was noticed the field about students has as a value list of dictionaries with information about students, but after checking, it was seen that the length of those lists was always one, so it was converted to just a single value. They all have the same structure (containing name, surname, and ID). Supervisors, co-supervisors and examiners further would be also instances of Staff persons as well as they have id-s which for those present in the two datasets, are the same. For the purpose of this phase of the methodology, a sub-dataset was extracted for each of those fields, with adding id-s referring to the persons at the original Theses dataset. In this way, it would be easier to identify the properties to each of these entity types (which were identified as: student - core, while the remaining ones as contextual).

- In the case of Publications, also considering it a contextual entity, the datasets obtained (for the three years) are richer in information. At first they were handled separately at this initial stage, by keeping track of the year the datasets refer to. So it was separately dealt with every dataset, also for dividing the nested structures into smaller components. But later it was decided to put the datasets together in one collection, since there is the field which refers to the year for each publication. Then again, a similar procedure as the above-mentioned was committed for the authors of the publishings (authors of the publishings identified as core entities in this particular case, belonging to the common class Person). Moreover, a process was done for this also for the field "file", which contains information about the specific file of the publishing. All of the existing ones were transferred to another dataset of files, and using the link field values to the specific files as identifiers since they are unique. At a later stage of development, as explained, this was actually changed. Another small modification was done in order to transform the values of the field "openAccess" from a string to a boolean type. Quite a lot of publications in fact were missing values for the field "file" and the information regarding it, so in that case, as a value was put "NOT_PRESENT". Doing the described procedure provides a way in which the entity type File can be represented by itself and reused in some further analysis or projects if needed with the ontology created or by enriching it, thus integrating it. Thus it would then not be a contextual for this purpose, but as a core one, considering the purpose of interest. Another consideration that could be made is to divide the different publications according to their type which would then be more differentiated subclasses, but this was not done because of the consideration that in such a way things would get too specific and reusability may be reduced.
- Then, when it comes to the courses distribution, it was considered in order to be able to link each specific course of study (degree program, a contextual entity) to a department at hand. This dataset contains all the courses with the professors who teach them and the degree program to which they belong. For the purposes of this project, it was thought useful to get the degree courses in order to then link them with the theses, and to departments. So only those two fields (degrees and departments) were extracted, the duplicates removed, and systemized, also adding an ID for each degree. Finally, from the name of the degree, the type was extracted (LT - Bachelor, LM - master, and some other possibilities).

Another consideration that was taken into account is the possibility to extract the common entity Creative work (which is present in the schema.org resource), to which both publication

and thesis would belong to. They would also share some of the properties, like title and year.

Instead, the data about the different departments (a department/or a faculty/ in this case being a core entity) was scraped by the web portal of the digital university using the Python programming language and the library *beautifulsoup4*, gathering the information that is present in each faculty's vignette (director, address, and website). The data was put in a .csv file. One of the departments (Centre for Medical Sciences - CISMED) was missing a link to its corresponding website, but it was added manually. Since, however, this dataset did not contain identifiers for each department, those ones were extracted from the dataset about degrees and departments (the last above), and each department present in this dataset was found and linked to the department ID that could be found in the other one. This degrees dataset was further used to identify if there is a missing department, and the one found was the language center of UniTn, which was then inserted.

The approach of dividing the datasets due to the exhibited nested structure should prove practically a good practice for the above-mentioned motive for reusability, as well as for a more systemized organization of data and the corresponding properties. In this way, the connection between the different Entity Types would be done in the later stage, as currently, the ontologies are built on a dataset level.

Here is a specific example of the work conducted with the Staff dataset.

```
[{"id": "guest00008627_1",
  "name": "Maurizio",
  "surname": "Ferrari",
  "phone": [],
  "position": [{"role": "Guest",
    "unitName": "Department of Physics",
    "unitid": "ST00008627"}],
  {"id": "1547a72aafb2b91e063ef3d978537a7a",
  "name": "Anna",
  "surname": "Scalfi",
  "phone": [],
  "position": [{"role": "Guest",
    "unitName": "Department of Sociology and social research",
    "unitid": "ST00008630"}]},
```

Figure 1: Extract of the original json dataset about staff of Unitn

```
[{"id": "guest00008627_1",
  "role": "Guest",
  "situated_at": "Department of Physics",
  "unit_id": "ST00008627"},
 {"id": "guest00008630_2",
  "role": "Guest",
  "situated_at": "Department of Sociology and social research",
  "unit_id": "ST00008630"},

 {"id": "teasto00008632_3",
  "role": "Teaching fellow",
  "situated_at": "Department of Humanities",
  "unit_id": "ST00008632"},
 {"id": "teasto00008621_4",
  "role": "Teaching assistant",
  "situated_at": "Department of Economics and Management",
  "unit_id": "ST00008621"},
```

Figure 2: Extract of how is defined the 'positions' dataset

```
{"id": "9c0c137493b2ec94df0ad18f8c79fd40",
  "name": "Bruno",
  "surname": "Callegher",
  "phone": [],
  "position": ["teasto00008632_3"]},
 {"id": "3f561286a93650282713dc900e008f98",
  "name": "Moreno",
  "surname": "Ferrarese",
  "phone": [],
  "position": ["teasto00008621_4", "teasto00008630_5"]},
```

Figure 3: Extract of how the modified staff dataset looks like

5.3 Some exploration of the datasets defined

It may be useful to have a look at how the datasets defined look like and in particular explore some distributions of the variables. In the case for this project, almost all of the variables are of character type (strings), therefore some bar plots are exhibited just to have a brief idea of how the data looks like. It is done for the attributes which are seen to be descriptive and to which makes sense to have such an overview.

Theses dataset contains a lot of fields referring to persons. Below, instead, is a bar plot of what the numbers of theses are there in the three years considered. 2022 has a bit less than the others, as this year has not yet finished.



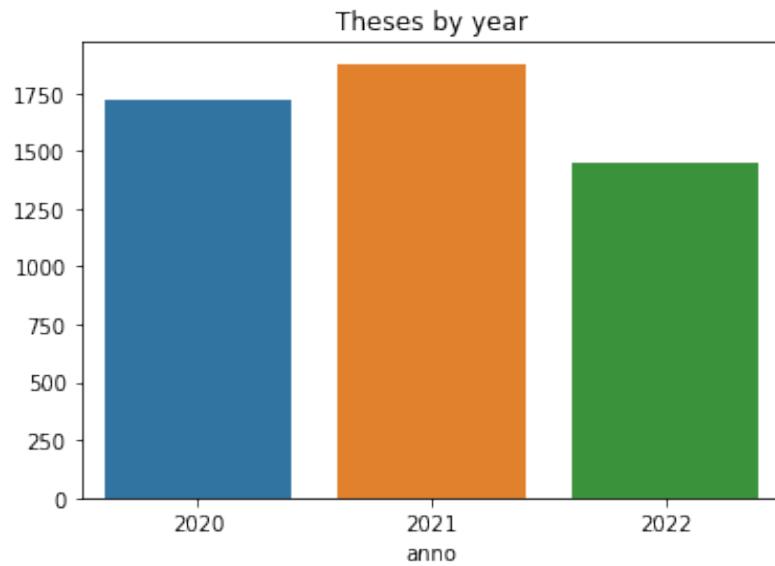


Figure 4: Theses by the years

Publications can also be divided in the groups from which they come from, as well as to their type. It can be noticed that publications from 2022 are lower in size, due to the fact that the year 2022 has not yet finished. In terms of type, there is a significant number of journal articles, followed by chapters and conference articles.

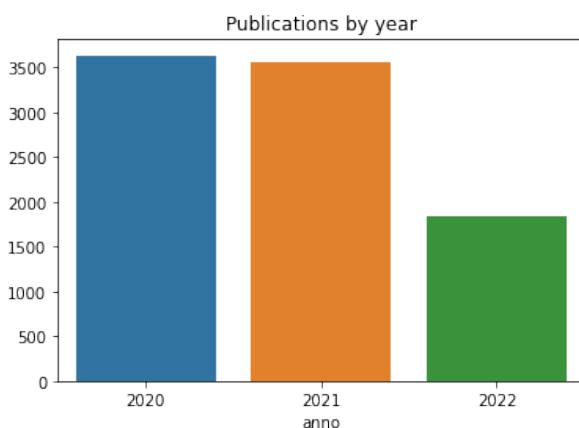


Figure 5: Publications by year when published

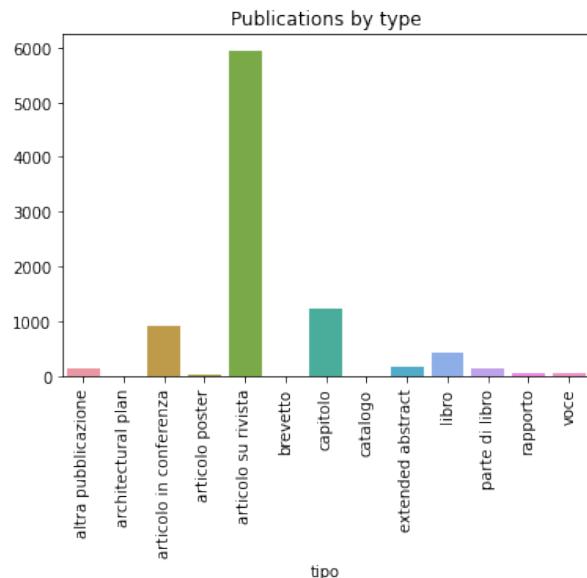


Figure 6: Publications according to their type

Positions exhibit quite a high variability in the values for most of the fields.



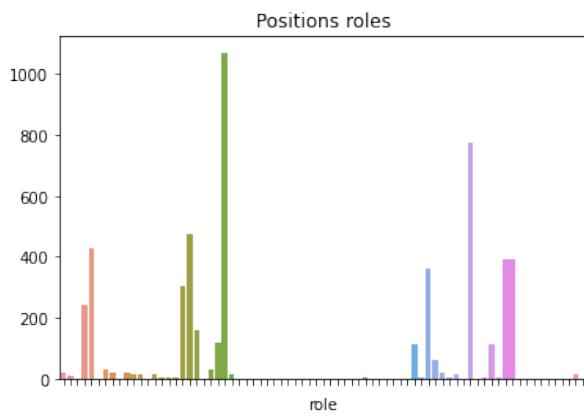


Figure 7: Diversity in roles

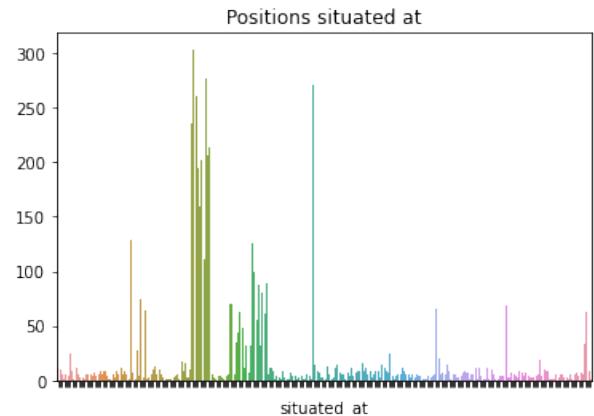


Figure 8: Diversity in place where working

Degrees dataset was examined in terms of distribution for the type of a course study program (Master (LM) and Bachelor (LT) programs significantly highly represented), as well as according to the department from which they are offered.

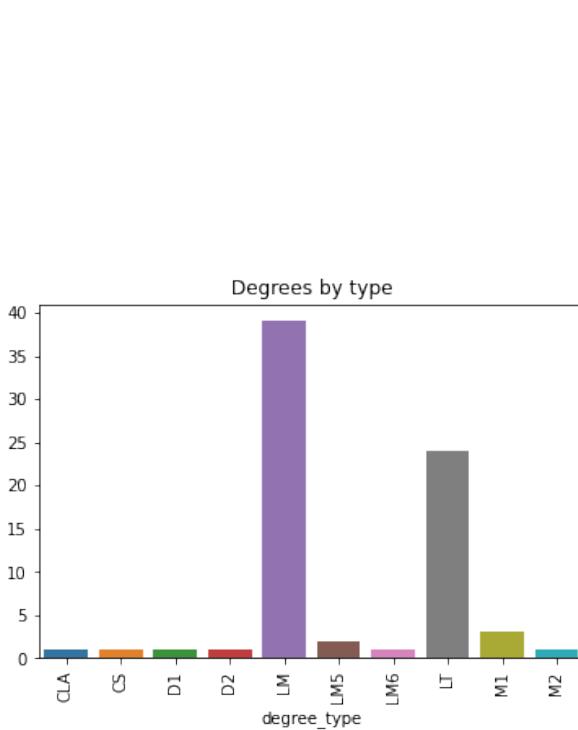


Figure 9: Degrees according to type

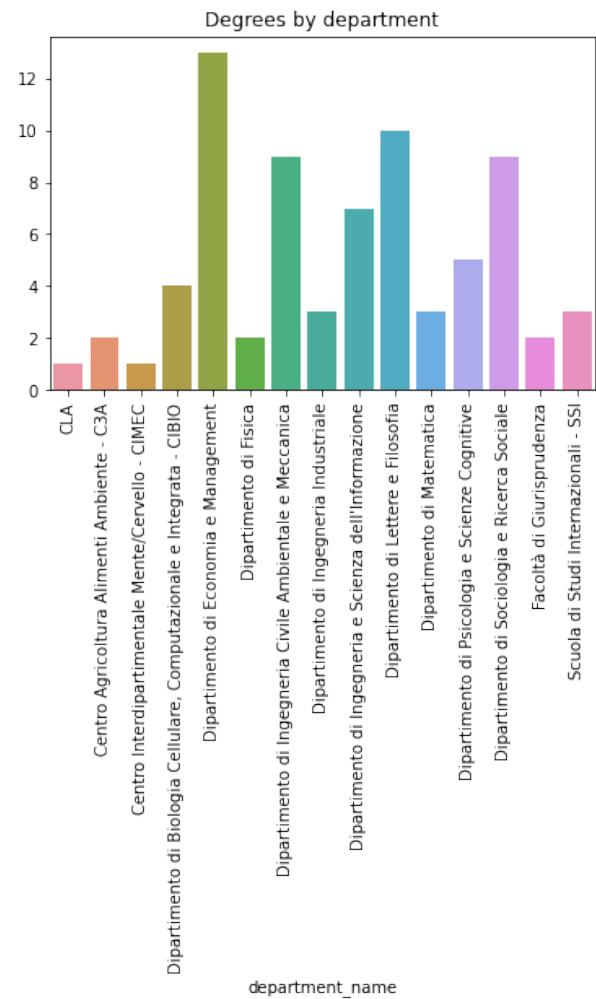
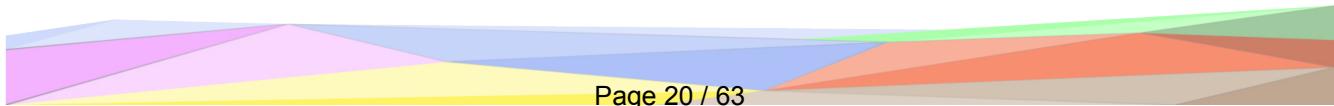


Figure 10: Degrees according to department

Finally, **Departments** dataset contains as a string the addresses of the faculties. Using geocoding, these are shown below on the map in order to have at least a visual perspective of the spatial aspect of the project.



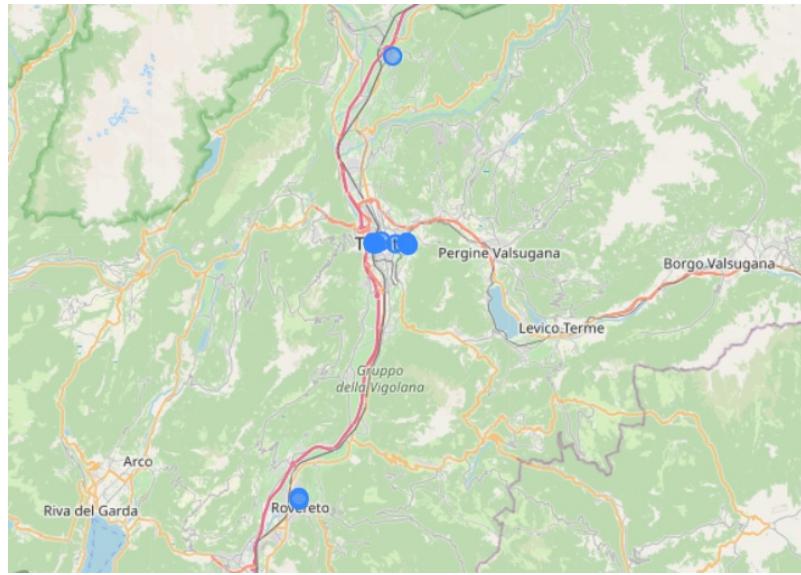


Figure 11: Mapping of the departments dataset

5.4 Creation of Ontologies Dataset by Dataset

After having identified, modified, and systemized these data resources, it was proceeded with the definition of the ontologies for each one dataset extracted. To identify the specific entities at hand and their properties, we relied on what are the fields in the JSON files (or columns in the .CSV files) finalized after initial data management. The software Protege was used in order to complete this stage.

Following are the identified objects (the work has been done dataset by dataset) with the corresponding properties extracted. They are coded as data properties in the ontologies defined with Protege. In this way, there is just one Entity Type per dataset and no object properties. Instead though the id to which a particular value of a field is pointing to, would actually be another etype, so it would be later defined as object property and defining nested ontology in the dataset with a data property with ID as a value, and the same exact ID being a data property in a new extracted dataset.

Degree	Department	Staff Person	Position
has_degree_id	has_department_id	has_name	has_position_id
has_degree_name	has_department_name	has_surname	has_position_role
has_degree_type	has_department_director	has_person_id	is_situated_at_department
	has_department_address	has_phone	is_situated_at_department_id
	has_department_website		

Publication	File	Thesis	Author
has_pub_id	has_file_link	has_title	has_name
has_publication_title	has_file_name	is_from_year	has_surname
has_publication_type	has_format	has_degree_name	has_id
is_from_year	has_license		
has_language	has_version		
has_file_link	has_open_access		
has_citation			

Student	Supervisor	Cosupervisor	Examiner
has_name	has_name	has_name	has_name
has_surname	has_surname	has_surname	has_surname
has_id	has_id	has_id	has_id

5.5 Connecting data with ontology

The final step for the Inception is to bring everything together using the Karma data mapping, which maps datasets with corresponding schemas (from Protege). The files in hand used as input for the program are in .json and .csv format (data resources), and .owl format (ontologies defined with Protege).

What was done is mapping each field identifying it as a property of the corresponding class. A nice trait of Karma is the possibility to handle .json formats with lists as values.

On the figures below there are examples of how the mapping procedure were done for the Department and Degree entities and the corresponding data properties. For degree, department_id was not mapped, as it will afterwards be used for the object property of this Etype.

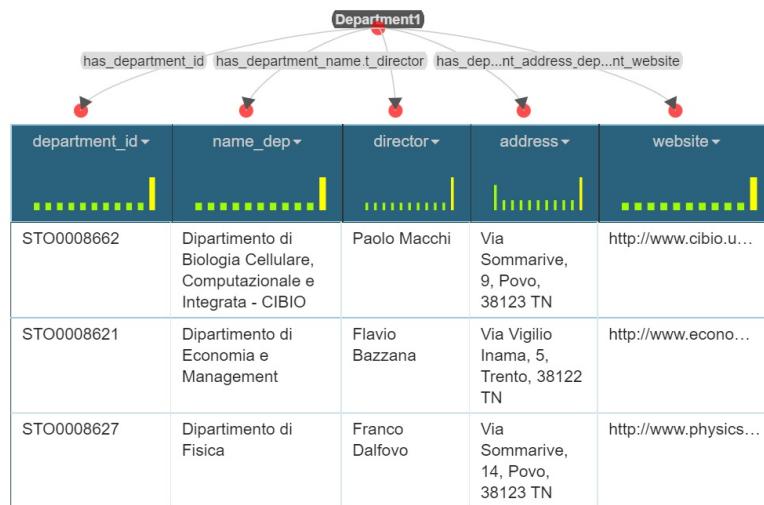


Figure 12: Mapping of the departments dataset



department_name ▾	department_id ▾	degree_name ▾	degree_id ▾	degree_type ▾
Dipartimento di Ingegneria Civile Ambientale e Meccanica	STO0008626	Ingegneria civile (LT)	ingegneriacivile(lt)_1	LT
Dipartimento di Economia e Management	STO0008621	Gestione aziendale (LT)	gestioneaziendale...	LT
Dipartimento di Lettere e Filosofia	STO0008632	Beni culturali (LT)	beniculturali(lt)_3	LT
Dipartimento di Lettere e Filosofia	STO0008632	Filosofia (LT)	filosofia(lt)_4	LT
Dipartimento di Lettere e Filosofia	STO0008632	Studi storici e filologico-letterari	studistoriciefilologi... letterari(lt)_5	LT

Figure 13: Mapping of the degrees dataset

Next figure shows the procedure for the Staff person entity which exhibits a nested structure, which was handled as discussed earlier by creating another dataset of the positions.

identifier ▾	name ▾	surname ▾	phone ▾ values ▾	position ▾ values ▾
3d22af80edc6910...	Maurizio	Ferrari		guessto0008627_1
1547a72aafb2b91...	Anna	Scalfi		guessto0008630_2
9c0c137493b2ec9...	Bruno	Callegher		teasto0008632_3
3f561286a936502...	Moreno	Ferrarese		teasto0008621_4 teasto0008630_5
b69872e46afdf8f11...	Paolo	Barbieri	0461 281326 0461 281437	fulsto0000371_6 fulsto0008630_7
42ab203bc6bbc9...	Serena	Manara		ressto0008662_8
d37dbdbf4b89959...	Irene	Villa		teasto0008628_9 ressto0008628_10

Figure 14: Mapping of the staff dataset

id ▾	role ▾	situated_at ▾	unit_id ▾
guessto0008627_1	Guest	Department of Physics	STO0008627
guessto0008630_2	Guest	Department of Sociology and social research	STO0008630
teasto0008632_3	Teaching fellow	Department of Humanities	STO0008632
teasto0008621_4	Teaching assistant	Department of Economics and Management	STO0008621
teasto0008630_5	Teaching fellow	Department of Sociology and social research	STO0008630
fulsto0000371_6	Full professor	Doctoral School of Social Sciences	STO0000371
fulsto0008630_7	Full professor	Department of Sociology and social research	STO0008630

Figure 15: Mapping of the positions dataset

In the case of the Etypes Student, Author, Supervisor, Co-supervisor, Examiner, for all of them was used the same kind of schema with only the three properties specified earlier. At the current stage, there is no other information available for each one of them (specifically for students, as for the others they may be extracted if those people are part of the university staff). However, this may represent a particular weak point because of the lack of more information. This remains an open issue. Below a mapping of an example for examiners.

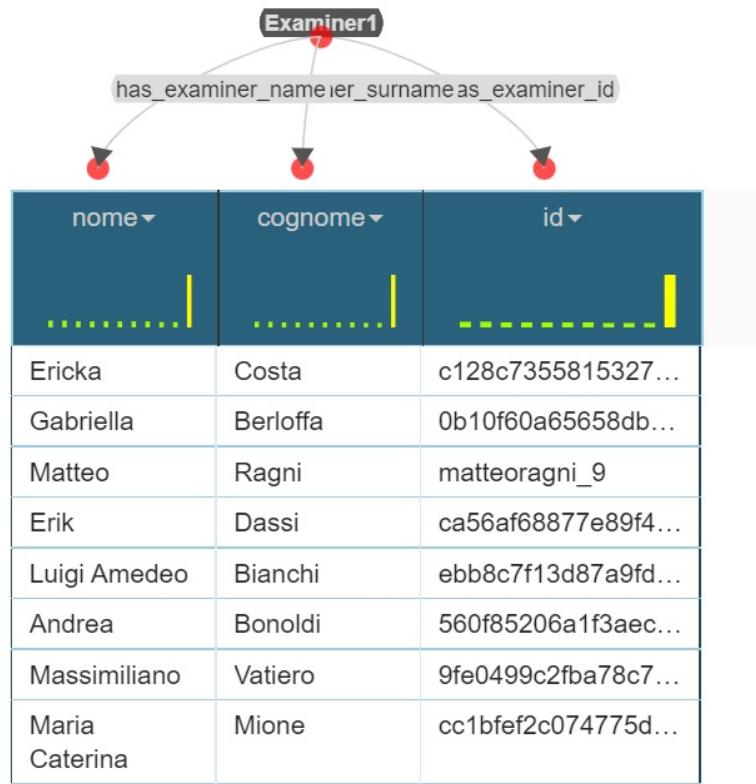


Figure 16: Mapping of the examiners dataset

The program outputs a .ttl file that is the RDF. In each case, the specific fields were mapped to the appropriate properties identified earlier when constructing the schema.

As output there are also reports shown below.

```

## Semantic Types
| Column | Property | Class |
| ----- | ----- | ----- |
| _anno_ | `http://knowdive.disi.unitn.it/etype#is_from_year` | `Thesis1` |
| _corsoStudi_ | `http://knowdive.disi.unitn.it/etype#has_degree_name` | `Thesis1` |
| _thesis_id_ | `http://knowdive.disi.unitn.it/etype#has_thesis_id` | `Thesis1` |
| _titolo_ | `http://knowdive.disi.unitn.it/etype#has_title` | `Thesis1` |

## Semantic Types
| Column | Property | Class |
| ----- | ----- | ----- |
| _cognome_ | `http://knowdive.disi.unitn.it/etype#has_student_surname` | `Student1` |
| _id_ | `http://knowdive.disi.unitn.it/etype#has_student_id` | `Student1` |
| _nome_ | `http://knowdive.disi.unitn.it/etype#has_student_name` | `Student1` |

## Semantic Types
| Column | Property | Class |
| ----- | ----- | ----- |
| _cognome_ | `http://knowdive.disi.unitn.it/etype#has_supervisor_surname` | `Supervisor1` |
| _id_ | `http://knowdive.disi.unitn.it/etype#has_supervisor_id` | `Supervisor1` |
| _nome_ | `http://knowdive.disi.unitn.it/etype#has_supervisor_name` | `Supervisor1` |

## Semantic Types
| Column | Property | Class |
| ----- | ----- | ----- |
| _cognome_ | `http://knowdive.disi.unitn.it/etype#has_cosupervisor_surname` | `Cosupervisor1` |
| _id_ | `http://knowdive.disi.unitn.it/etype#has_cosupervisor_id` | `Cosupervisor1` |
| _nome_ | `http://knowdive.disi.unitn.it/etype#has_cosupervisor_name` | `Cosupervisor1` |

## Semantic Types
| Column | Property | Class |
| ----- | ----- | ----- |
| _cognome_ | `http://knowdive.disi.unitn.it/etype#has_examiner_surname` | `Examiner1` |
| _id_ | `http://knowdive.disi.unitn.it/etype#has_examiner_id` | `Examiner1` |
| _nome_ | `http://knowdive.disi.unitn.it/etype#has_examiner_name` | `Examiner1` |

## Semantic Types
| Column | Property | Class |
| ----- | ----- | ----- |
| _anno_ | `http://knowdive.disi.unitn.it/etype#is_from_year` | `Publication1` |
| _citazioni_ | `http://knowdive.disi.unitn.it/etype#has_citation` | `Publication1` |
| _pub_id_ | `http://knowdive.disi.unitn.it/etype#has_pub_id` | `Publication1` |
| _tipo_ | `http://knowdive.disi.unitn.it/etype#has_publication_type` | `Publication1` |
| _titolo_ | `http://knowdive.disi.unitn.it/etype#has_publication_title` | `Publication1` |
| _values_ | `http://knowdive.disi.unitn.it/etype#has_language` | `Publication1` |

## Semantic Types
| Column | Property | Class |
| ----- | ----- | ----- |
| _formato_ | `http://knowdive.disi.unitn.it/etype#has_format` | `File1` |
| _licenza_ | `http://knowdive.disi.unitn.it/etype#has_license` | `File1` |
| _link_ | `http://knowdive.disi.unitn.it/etype#has_file_link` | `File1` |
| _nome_ | `http://knowdive.disi.unitn.it/etype#has_file_name` | `File1` |
| _openAccess_ | `http://knowdive.disi.unitn.it/etype#is_open_access` | `File1` |
| _versione_ | `http://knowdive.disi.unitn.it/etype#has_version` | `File1` |

## Semantic Types
| Column | Property | Class |
| ----- | ----- | ----- |
| _cognome_ | `http://knowdive.disi.unitn.it/etype#has_author_surname` | `Author1` |
| _id_ | `http://knowdive.disi.unitn.it/etype#has_author_id` | `Author1` |
| _nome_ | `http://knowdive.disi.unitn.it/etype#has_author_name` | `Author1` |

```

```

## Semantic Types
| Column | Property | Class |
| ----- | ----- | ----- |
| _identifier_ | `http://knowdive.disi.unitn.it/etype#has_person_id` | `Staff_person1` |
| _name_ | `http://knowdive.disi.unitn.it/etype#has_name` | `Staff_person1` |
| _surname_ | `http://knowdive.disi.unitn.it/etype#has_surname` | `Staff_person1` |
| _values_ | `http://knowdive.disi.unitn.it/etype#has_phone` | `Staff_person1` |

## Semantic Types
| Column | Property | Class |
| ----- | ----- | ----- |
| _id_ | `http://knowdive.disi.unitn.it/etype#has_position_id` | `Position1` |
| _role_ | `http://knowdive.disi.unitn.it/etype#has_position_role` | `Position1` |
| _situated_at_ | `http://knowdive.disi.unitn.it/etype#is_situated_at_department` | `Position1` |

## Semantic Types
| Column | Property | Class |
| ----- | ----- | ----- |
| _degree_id_ | `http://knowdive.disi.unitn.it/etype#has_degree_id` | `Degree1` |
| _degree_name_ | `http://knowdive.disi.unitn.it/etype#has_degree_name` | `Degree1` |
| _degree_type_ | `http://knowdive.disi.unitn.it/etype#has_degree_type` | `Degree1` |

## Semantic Types
| Column | Property | Class |
| ----- | ----- | ----- |
| _address_ | `http://knowdive.disi.unitn.it/etype#has_department_address` | `Department1` |
| _department_id_ | `http://knowdive.disi.unitn.it/etype#has_department_id` | `Department1` |
| _director_ | `http://knowdive.disi.unitn.it/etype#has_department_director` | `Department1` |
| _name_dep_ | `http://knowdive.disi.unitn.it/etype#has_department_name` | `Department1` |
| _website_ | `http://knowdive.disi.unitn.it/etype#has_department_website` | `Department1` |

```

6 Informal Modeling

6.1 Phase Description

The following section of the report describes all the steps done in the Informal modeling phase of the iTelos Methodology. During it, a teleology is being built by combining objects, actions, and functions in a meaningful and effective way.

- **Input:** E-types obtained by Purpose formalization
- **Output:** ER model formalization

6.2 Initial considerations

From the CQs formalization step the following main E-types regarding this project: Staff person, Degree, Publication, Thesis and Department were extracted. Various connections as well as hierarchies among some of these entities were discovered.

The purpose of this stage now is to identify exactly the relations, while the hierarchies will be dealt with at the next stage. To provide an illustration of such relations in our project, for instance, for hierarchy: different types of persons present in our specifications are subcategories of Person. On the other hand, for functional and actionable relations: types of persons (advisor,



co-advisor, examiner and student) are entities associated to E-type Thesis by their role related to it. In a similar way, Thesis is connected to the E-type Degree such as a completion of the academic career and the courses of study are organized in Department. Instead, a specific Position (of a Staff person) is in some department. In this case, there are indeed some open issues, as the department in which a Staff person works is not always one of those that are in the data sources about departments, as it is not of the main ones (the faculties). Therefore, they would not provide the specific information more.

All The E-types above mentioned, in particular Staff person and Department, are directly or indirectly related with E-type University, the initial object in which the Purpose of our project is formalized.

6.3 Schema Level E-type

All the entities identified during the previous phase from the Competency questions and modeling sheet are divided into three categories.

The Common category describes all the E-types that are more broad elements and with a high level of usage, also in different fields. E-types that are crucial for answering the competency questions are organized as core. Finally, some of the E-types that are important to bring more completeness to the knowledge graph are identified as contextual E-types. These types of entities from each category are listed below.

Common E-types

- **Staff Person:** a common type entity describing any human who is working in a particular place.

Data Properties:

- name: name of the person.
- surname: surname of the person.
- person ID: identification code of the person.
- phone: phone number of the person.

- **University:** a common type entity describing the spatial location representation of a university in the world.

Data Properties:

- region latitude: the latitudinal coordinates of the building.
- region longitude: the longitudinal coordinates of the building.
- region altitude: the altitude level in which is located the building.
- start time: when the university starts to exist.
- end time: the current year in which the university still exists.

Core E-types



-
- **Student:** a core type entity describing any university student.

Data properties:

- name: the name of the student
- surname: the surname of the student
- student ID: identification code of the student

- **Department:** a core type entity describing a university faculty.

Data Properties:

- department ID: identification code of department
- name: name of department
- director: the name of the director of department
- address: address of department
- website: website of department

- **Author:** a core type entity describing the general information about this type of staff person.

Data Properties:

- name: the name of the Author
- surname: Surname of the Author
- author ID: identification code of the Author

- **File:** a core type entity describing the general information about a representation of a document.

Data Properties:

- link: the link to access on file
- name: the name of the file
- format: the representation of the file
- license: the authorization to access on file
- version: the version of the file
- open access: whether this file has open access or not

Contextual E-type

- **Thesis:** a contextual type entity describing the thesis document.

Data properties:

- title: the title of the thesis
- year: the year in which the thesis has been written.
- degree name: the name of the course of study related to the thesis

-
- **Degree:** a contextual type entity describing any general study courses offered by university.

Data Properties:

- ID: identification code of degree
- name: the name of the course of study
- type: the type of degree
- department ID: the id of the department offering the degree
- department name: the name of the department by which the degree is offered

- **Supervisor:** a contextual type entity to describe a professor who helps the student to write a thesis.

Data Properties:

- name: the name of supervisor
- surname: the surname of supervisor
- ID: the identification code of supervisor

- **Co-supervisor:** a contextual type entity to describe a person who helps student to write a thesis, but is not the main one doing it.

Data Properties:

- name: the name of co-supervisor
- surname: the surname of co-supervisor
- ID: the identification code of co-supervisor

- **Examiner:** a contextual type entity to describe the person who valuate the student career and his final work.

Data properties:

- name: the name of examiner
- surname: the surname of examiner
- ID: the identification code of examiner

- **Publication:** a contextual type entity to describe the work of publishing followed by research and professors.

Data Properties:

- ID: identification code of the publication
- title: the title of the publication
- type: the type of the publication
- year: the year of the publication
- language: the language in which the publication has been written
- citation: the citation on the publication of other works

- **Position:** a contextual type entity to describe the specific role a staff person takes.

Data Properties:

- ID: identification code of the position
- role: the type of the job
- department: to which department this job is affiliated to.

6.4 Description of ER Model

The entities defined before were modeled using appropriate relations between them. Each entity has the data properties as identified previously, but also certain object properties that explain the role of it, as well as the specific functions and actions related to it.

As suggested by the project tutor, two entity relationship diagrams were developed, one is a more general version including more objects and properties, as driven by the domain and by the competency questions. The other is a more specific for this project, keeping in mind the particular purpose, as well as the data sources used for the project, that are to be aligned accordingly. It has a smaller number of objects and properties. This also would concentrate more on the integration of data and knowledge for the usage of the KG.

The more general ER diagram is shown first.

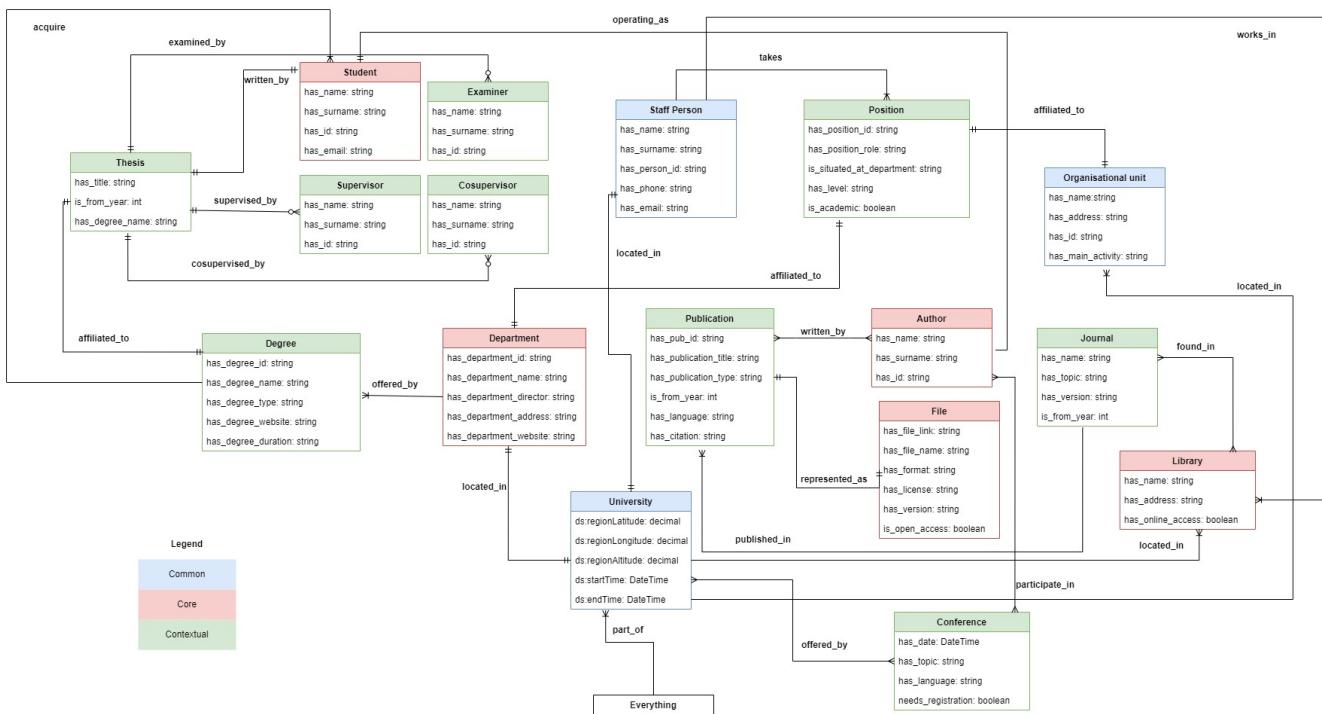


Figure 17: Entity Relationship Diagram - a more general version (made with draw.io)

This model is broader and a generalized one. As it can be seen from the graph, it includes certain objects not previously listed (Organizational unit, Conference, Journal, Library). Further,

there are several other data and object properties which enrich them. Afterwards those were not considered, because of the need to concentrate on the specific purpose, still keeping in mind the reusability issue, but also regarding the data as well. It is actually an open issue of consideration whether to include any of those (like Library). From the one hand, they could give more information to satisfy specific queries of producers and consumers. From the other instead, as mentioned, it may be they go a bit too far from the main reasoning behind this specific Knowledge graph. There would also be present an issue in the data layer for example for the Journal entity, as usually journals are with restricted access and no meaningful information could be found, neither any identification.

With this considerations in mind, the concrete and more specific ER model is presented below.

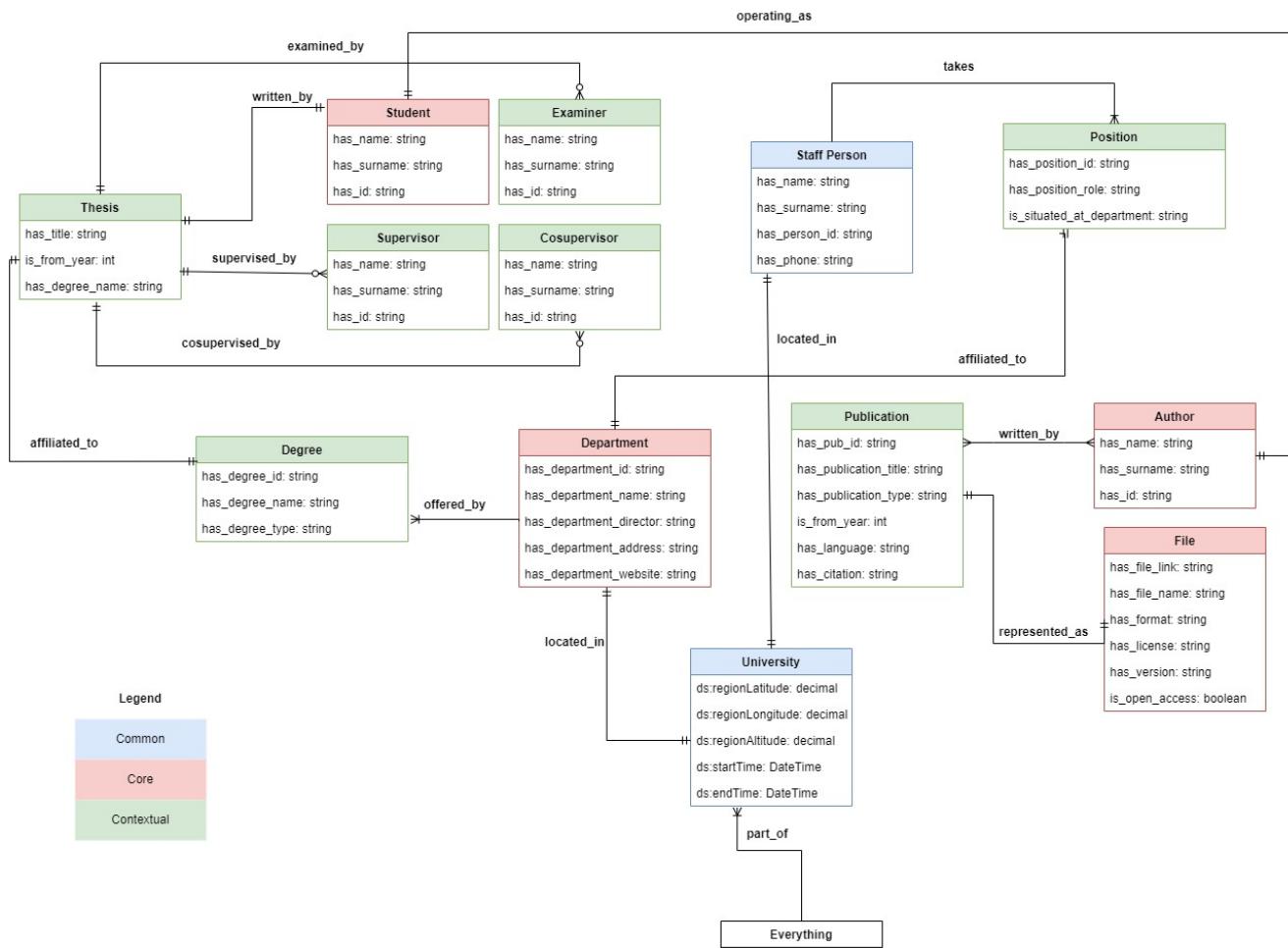


Figure 18: Entity Relationship Diagram - a more specific version (made with draw.io)

The ER model represents a good visual of what is actually being built as a Knowledge layer of the graph. Here, only the identified entities and properties are being shown in this model, still with respect to the specific user purpose. It is also the one which is used to define the teleology later in Protege by importing the ontologies defined earlier and by creating object properties defining from which to which e-types they correspond to.



As can be seen from the diagram, the entities are divided in common, core and contextual, and all of them are somehow connected with links showing some object properties (defining spatio-temporal relations, actions, or functions).

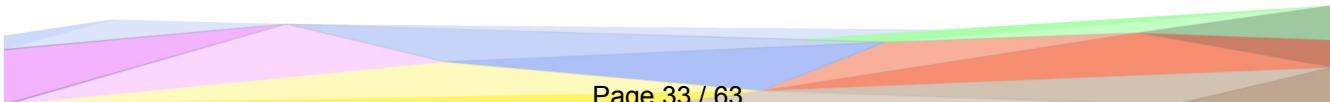
The object properties are the following:

- `written_by` (Thesis written by Student; Publication written by Author)
- `supervised_by` (Thesis being supervised by a particular person)
- `cosupervised_by` (The Co-supervisor for a Thesis of a student)
- `examined_by` (who the Examiner for this Thesis is)
- `affiliated_to` (two respective domains and ranges: Thesis to a Degree, meaning what study program was the thesis allowing the student to finish; Position to Department as the job place)
- `offered_by` (Degree program is offered by a specific Department of the university)
- `takes` (a Staff Person having a particular position)
- `operating_as` (Student who is also an Author of a publication)
- `represented_as` (How a Publication is digitally stored as a File)
- `located_in` (two domains and ranges: Department in a University; Staff Person in a University)
- `part_of` (wrapping University to Everything)

It is important to make explicit the spatio-temporal context to expose the underlying design assumptions. That is why the entity University was added with data properties which frame the knowledge graph in time and space. Usually, a University could be seen as either a more non-tangible object, providing the services of teaching, and where students consume this kind of services, and University itself collecting courses, research activities, etc. At the same time the tangible buildings, offices, departments. At data level, for what concerns our project, the instance of this E-type would be University of Trento, with its specific coordinates. Still, for the concrete project at hand, it is about *Digital University*, so in that case time space is needed as for example to identify personnel and faculties, however it is not of extremely high importance. That is further due to the fact that the Domain of Interest concerns *Research activities*.

Some entities are described with Spatial properties which provide to ER modeling a more precise position in space. For example the core E-type department is associated in terms of spatial feature with the common E-type University, which as "located in". This is due to the fact that each faculty has its physical address and its location can be identified. Same for a staff person as they would be working at this specific university.

There are different types of relationship: one to one, one to many, many to many, either mandatory and optional (optional meaning that for example it is not necessary to have such a relation, as is the case with Thesis and Cosupervisor). The relationship between E-type Author and Publication is described as a many-to-many relationship, due to the fact that many publications can be written by many authors and vice versa. While others connections represent



one-to-one relation, as student and thesis; or one-to-many relation, as Staff person and Position. Those kinds of relationships are presented using the different kinds of *arrows (connections)* visually in the graph.

Many decisions had to be taken here in terms of how to construct this model, and they were discussed with the tutor. For example, Student is connected to Author, as a master student could be the author of a publication. Then, even though not all publications actually have values for their representation as files, it is still useful for the purpose to include them, since File contains more information that could be of hand for users, e.g. a link to access the publication. Then, about position, it is connected with department as some of the positions which are academic positions, are affiliated to the specific scientific departments (e.g. Physics, Sociology, etc.), even though other roles are not academic therefore they will not be associated to neither of those departments. Still, as considered above, *research* being the main issue here for this project, the academic staff would play a role which is more important for users (like professors). Thesis, on the other hand, is connected with all the types of persons who are involved with it, as well as with the degree program for which it has been written. In such a way, a Student could be connected through Thesis to Degree and subsequently to Department. And this actually is one of the main powers of a knowledge graph connecting nodes with various edges.

In terms of level of abstraction, there exist three levels: object, action, and function. Object is the set of representation of how we perceive a certain thing. In this case, objects are modeled entity types identified, but objects would also be the instantiations of those. Functions are behaviours the objects are expected to have (usually it is a purpose or a role) and they form causalities, having usually roles of producers and consumers. In this case, Supervisor and Author are functions of a Staff person, while Staff Person has a producer function w.r.t. University. Student, instead has a consumer function w.r.t. University. In the more general case shown above, for example Thesis and Publication are functions of Creative work. Then, the third level of abstraction, actions, represents how objects change in time. Example of actions include a Supervisor supervising a thesis of a student, an Author writing a publication, or a Department offering a Degree program.

The draw.io application also allows to export an XML file with the metadata corresponding to the diagram created. Those files are stored. Below an instance of the file is shown. The full file can be accessed [here](#).

```

<mxfile host="app.diagrams.net" modified="2022-11-22T13:50:19.840Z" agent="5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/107.0.0.0 Safari/537.36" etag="sP9yNp0lyDXJUBug8hJu" version="20.5.3" type="google">
  <diagram id="R21EEUdfHJ1lhrx00" name="specific">
    <model id="2272" dx="675" grid="1" gridSize="10" guides="1" tooltips="1" connect="1" arrows="1" fold="1" page="1" pageScale="1" pageWidth="850" pageHeight="1100" math="0" shadow="0" extFonts="PermanentMarker">
      <root>
        <mxCell id="0"/>
        <mxCell id="1">
          <mxCell id="HrBg3700080e0ev1a9tPf-19" value="file" style="sublantern;fontStyle=1;childLayout=stackLayout;horizontal1=1;startSize=26;horizontalStack=0;resizeParent=1;resizeParentMax=0;resizeLast=0;collapsible=1;marginBottom=0;fillColor=#f8cecc;strokeColor=#b85450;" parent="1" vertex="1">
            <mgGeometry x="570" y="610" width="160" height="182" as="geometry"/>
          </mxCell>
        <mxCell id="HrBg3700080e0ev1a9tPf-20" value="has_file_name" style="text;strokeColor=None;fillColor=None;align=left;verticalAlign=top;spacingLeft=4;spacingRight=4;overflow=hidden;rotatable=0;points=[[0,0.5],[1,0.5]];portConstraint=eastwest;" parent="HrBg3700080e0ev1a9tPf-19" vertex="1">
          <mgGeometry y="26" width="160" height="26" as="geometry"/>
        </mxCell>
        <mxCell id="HrBg3700080e0ev1a9tPf-21" value="has_file_name" style="text;strokeColor=None;fillColor=None;align=left;verticalAlign=top;spacingLeft=4;spacingRight=4;overflow=hidden;rotatable=0;points=[[0,0.5],[1,0.5]];portConstraint=eastwest;" parent="HrBg3700080e0ev1a9tPf-19" vertex="1">
          <mgGeometry y="50" width="160" height="26" as="geometry"/>
        </mxCell>
        <mxCell id="HrBg3700080e0ev1a9tPf-22" value="has_format" style="text;strokeColor=None;fillColor=None;align=left;verticalAlign=top;spacingLeft=4;spacingRight=4;overflow=hidden;rotatable=0;points=[[0,0.5],[1,0.5]];portConstraint=eastwest;" parent="HrBg3700080e0ev1a9tPf-19" vertex="1">
          <mgGeometry y="74" width="160" height="26" as="geometry"/>
        </mxCell>
        <mxCell id="HrBg3700080e0ev1a9tPf-23" value="has_license" style="text;strokeColor=None;fillColor=None;align=left;verticalAlign=top;spacingLeft=4;spacingRight=4;overflow=hidden;rotatable=0;points=[[0,0.5],[1,0.5]];portConstraint=eastwest;" parent="HrBg3700080e0ev1a9tPf-19" vertex="1">
          <mgGeometry y="104" width="160" height="26" as="geometry"/>
        </mxCell>
        <mxCell id="HrBg3700080e0ev1a9tPf-24" value="has_version" style="text;strokeColor=None;fillColor=None;align=left;verticalAlign=top;spacingLeft=4;spacingRight=4;overflow=hidden;rotatable=0;points=[[0,0.5],[1,0.5]];portConstraint=eastwest;" parent="HrBg3700080e0ev1a9tPf-19" vertex="1">
          <mgGeometry y="130" width="160" height="26" as="geometry"/>
        </mxCell>
        <mxCell id="HrBg3700080e0ev1a9tPf-25" value="is_open_access" style="text;strokeColor=None;fillColor=None;align=left;verticalAlign=top;spacingLeft=4;spacingRight=4;overflow=hidden;rotatable=0;points=[[0,0.5],[1,0.5]];portConstraint=eastwest;" parent="HrBg3700080e0ev1a9tPf-19" vertex="1">
          <mgGeometry y="156" width="160" height="26" as="geometry"/>
        </mxCell>
        <mxCell id="HrBg3700080e0ev1a9tPf-26" value="Publication" style="sublantern;fontStyle=1;childLayout=stackLayout;horizontal1=1;startSize=26;horizontalStack=0;resizeParent=1;resizeParentMax=0;resizeLast=0;collapsible=1;marginBottom=0;fillColor=#d5e8d4;strokeColor=#82b366;" parent="1" vertex="1">
          <mgGeometry x="260" y="182" width="160" height="182" as="geometry"/>
        </mxCell>
        <mxCell id="HrBg3700080e0ev1a9tPf-27" value="has_pub_id" string="string" style="text;strokeColor=None;fillColor=None;align=left;verticalAlign=top;spacingLeft=4;spacingRight=4;overflow=hidden;rotatable=0;points=[[0,0.5],[1,0.5]];portConstraint=eastwest;" parent="HrBg3700080e0ev1a9tPf-32" vertex="1">
          <mgGeometry y="26" width="170" height="26" as="geometry"/>
        </mxCell>
        <mxCell id="HrBg3700080e0ev1a9tPf-28" value="has_publication_title" string="string" style="text;strokeColor=None;fillColor=None;align=left;verticalAlign=top;spacingLeft=4;spacingRight=4;overflow=hidden;rotatable=0;points=[[0,0.5],[1,0.5]];portConstraint=eastwest;" parent="HrBg3700080e0ev1a9tPf-32" vertex="1">
          <mgGeometry y="52" width="170" height="26" as="geometry"/>
        </mxCell>
        <mxCell id="HrBg3700080e0ev1a9tPf-29" value="has_publication_type" string="string" style="text;strokeColor=None;fillColor=None;align=left;verticalAlign=top;spacingLeft=4;spacingRight=4;overflow=hidden;rotatable=0;points=[[0,0.5],[1,0.5]];portConstraint=eastwest;" parent="HrBg3700080e0ev1a9tPf-32" vertex="1">
          <mgGeometry y="78" width="170" height="26" as="geometry"/>
        </mxCell>
        <mxCell id="HrBg3700080e0ev1a9tPf-30" value="is_from_year" int="int" style="text;strokeColor=None;fillColor=None;align=left;verticalAlign=top;spacingLeft=4;spacingRight=4;overflow=hidden;rotatable=0;points=[[0,0.5],[1,0.5]];portConstraint=eastwest;" parent="HrBg3700080e0ev1a9tPf-32" vertex="1">
          <mgGeometry y="104" width="170" height="26" as="geometry"/>
        </mxCell>
        <mxCell id="HrBg3700080e0ev1a9tPf-31" value="is_supervisor" string="string" style="text;strokeColor=None;fillColor=None;align=left;verticalAlign=top;spacingLeft=4;spacingRight=4;overflow=hidden;rotatable=0;points=[[0,0.5],[1,0.5]];portConstraint=eastwest;" parent="HrBg3700080e0ev1a9tPf-32" vertex="1">
          <mgGeometry y="130" width="170" height="26" as="geometry"/>
        </mxCell>
        <mxCell id="HrBg3700080e0ev1a9tPf-32" value="Supervisor" string="string" style="text;strokeColor=None;fillColor=None;align=left;verticalAlign=top;spacingLeft=4;spacingRight=4;overflow=hidden;rotatable=0;points=[[0,0.5],[1,0.5]];portConstraint=eastwest;" parent="HrBg3700080e0ev1a9tPf-32" vertex="1">
          <mgGeometry y="156" width="170" height="26" as="geometry"/>
        </mxCell>
        <mxCell id="HrBg3700080e0ev1a9tPf-33" value="Student" string="string" style="text;strokeColor=None;fillColor=None;align=left;verticalAlign=top;spacingLeft=4;spacingRight=4;overflow=hidden;rotatable=0;points=[[0,0.5],[1,0.5]];portConstraint=eastwest;" parent="HrBg3700080e0ev1a9tPf-32" vertex="1">
          <mgGeometry y="182" width="170" height="26" as="geometry"/>
        </mxCell>
        <mxCell id="HrBg3700080e0ev1a9tPf-34" value="Degree" string="string" style="text;strokeColor=None;fillColor=None;align=left;verticalAlign=top;spacingLeft=4;spacingRight=4;overflow=hidden;rotatable=0;points=[[0,0.5],[1,0.5]];portConstraint=eastwest;" parent="HrBg3700080e0ev1a9tPf-32" vertex="1">
          <mgGeometry y="208" width="170" height="26" as="geometry"/>
        </mxCell>
        <mxCell id="HrBg3700080e0ev1a9tPf-35" value="Positon" string="string" style="text;strokeColor=None;fillColor=None;align=left;verticalAlign=top;spacingLeft=4;spacingRight=4;overflow=hidden;rotatable=0;points=[[0,0.5],[1,0.5]];portConstraint=eastwest;" parent="HrBg3700080e0ev1a9tPf-32" vertex="1">
          <mgGeometry y="234" width="170" height="26" as="geometry"/>
        </mxCell>
        <mxCell id="HrBg3700080e0ev1a9tPf-36" value="Publication" string="string" style="text;strokeColor=None;fillColor=None;align=left;verticalAlign=top;spacingLeft=4;spacingRight=4;overflow=hidden;rotatable=0;points=[[0,0.5],[1,0.5]];portConstraint=eastwest;" parent="HrBg3700080e0ev1a9tPf-32" vertex="1">
          <mgGeometry y="260" width="170" height="26" as="geometry"/>
        </mxCell>
        <mxCell id="HrBg3700080e0ev1a9tPf-37" value="File" string="string" style="text;strokeColor=None;fillColor=None;align=left;verticalAlign=top;spacingLeft=4;spacingRight=4;overflow=hidden;rotatable=0;points=[[0,0.5],[1,0.5]];portConstraint=eastwest;" parent="HrBg3700080e0ev1a9tPf-32" vertex="1">
          <mgGeometry y="286" width="170" height="26" as="geometry"/>
        </mxCell>
        <mxCell id="HrBg3700080e0ev1a9tPf-38" value="Staff_person" string="string" style="text;strokeColor=None;fillColor=None;align=left;verticalAlign=top;spacingLeft=4;spacingRight=4;overflow=hidden;rotatable=0;points=[[0,0.5],[1,0.5]];portConstraint=eastwest;" parent="HrBg3700080e0ev1a9tPf-32" vertex="1">
          <mgGeometry y="312" width="170" height="26" as="geometry"/>
        </mxCell>
        <mxCell id="HrBg3700080e0ev1a9tPf-39" value="Examiner" string="string" style="text;strokeColor=None;fillColor=None;align=left;verticalAlign=top;spacingLeft=4;spacingRight=4;overflow=hidden;rotatable=0;points=[[0,0.5],[1,0.5]];portConstraint=eastwest;" parent="HrBg3700080e0ev1a9tPf-32" vertex="1">
          <mgGeometry y="338" width="170" height="26" as="geometry"/>
        </mxCell>
        <mxCell id="HrBg3700080e0ev1a9tPf-40" value="Department" string="string" style="text;strokeColor=None;fillColor=None;align=left;verticalAlign=top;spacingLeft=4;spacingRight=4;overflow=hidden;rotatable=0;points=[[0,0.5],[1,0.5]];portConstraint=eastwest;" parent="HrBg3700080e0ev1a9tPf-32" vertex="1">
          <mgGeometry y="364" width="170" height="26" as="geometry"/>
        </mxCell>
        <mxCell id="HrBg3700080e0ev1a9tPf-41" value="Cosupervisor" string="string" style="text;strokeColor=None;fillColor=None;align=left;verticalAlign=top;spacingLeft=4;spacingRight=4;overflow=hidden;rotatable=0;points=[[0,0.5],[1,0.5]];portConstraint=eastwest;" parent="HrBg3700080e0ev1a9tPf-32" vertex="1">
          <mgGeometry y="390" width="170" height="26" as="geometry"/>
        </mxCell>
        <mxCell id="HrBg3700080e0ev1a9tPf-42" value="ds_University" string="string" style="text;strokeColor=None;fillColor=None;align=left;verticalAlign=top;spacingLeft=4;spacingRight=4;overflow=hidden;rotatable=0;points=[[0,0.5],[1,0.5]];portConstraint=eastwest;" parent="HrBg3700080e0ev1a9tPf-32" vertex="1">
          <mgGeometry y="416" width="170" height="26" as="geometry"/>
        </mxCell>
        <mxCell id="HrBg3700080e0ev1a9tPf-43" value="Author" string="string" style="text;strokeColor=None;fillColor=None;align=left;verticalAlign=top;spacingLeft=4;spacingRight=4;overflow=hidden;rotatable=0;points=[[0,0.5],[1,0.5]];portConstraint=eastwest;" parent="HrBg3700080e0ev1a9tPf-32" vertex="1">
          <mgGeometry y="442" width="170" height="26" as="geometry"/>
        </mxCell>
        <mxCell id="HrBg3700080e0ev1a9tPf-44" value="Thesis" string="string" style="text;strokeColor=None;fillColor=None;align=left;verticalAlign=top;spacingLeft=4;spacingRight=4;overflow=hidden;rotatable=0;points=[[0,0.5],[1,0.5]];portConstraint=eastwest;" parent="HrBg3700080e0ev1a9tPf-32" vertex="1">
          <mgGeometry y="468" width="170" height="26" as="geometry"/>
        </mxCell>
      </root>
    </model>
  </diagram>
</mxfile>

```

Figure 19: Instance of XML file with data for the ER diagram

6.5 Implementation of Informal Modeling in Protege

The informal modeling of the Teleology is done with the Protege software as mentioned above. Now there are identified all the E-types (Objects), with the respective data properties, and object properties (which are the links in the diagram).

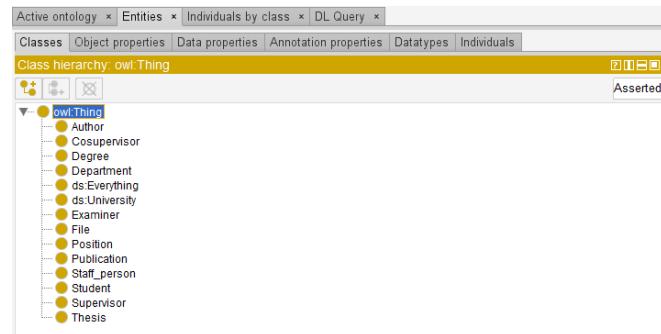


Figure 20: Teleology entities



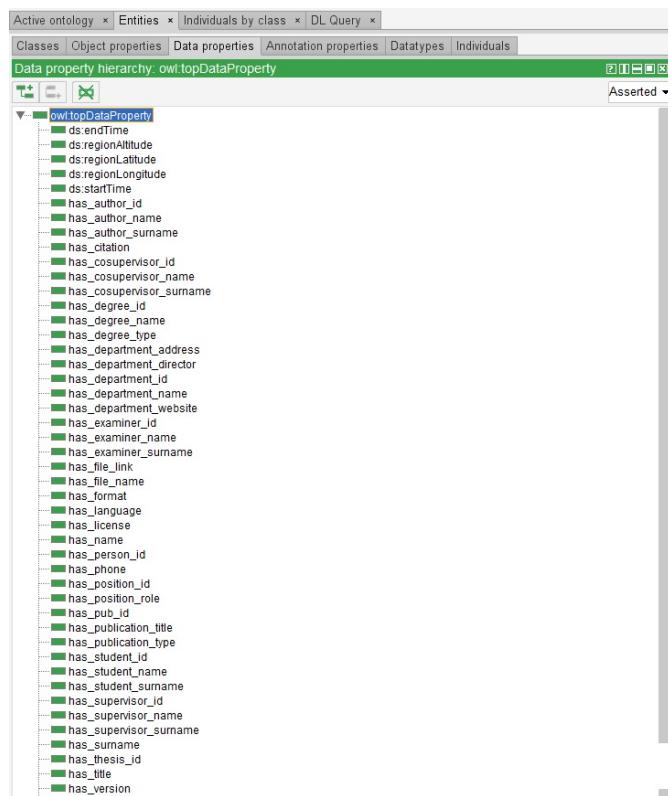


Figure 21: Teleology data properties



Figure 22: Teleology object properties

The domains and ranges are identified for each once-present object property, and for those present multiple times, domain and ranges are identified as object restrictions. An example of how it has been done follows below.

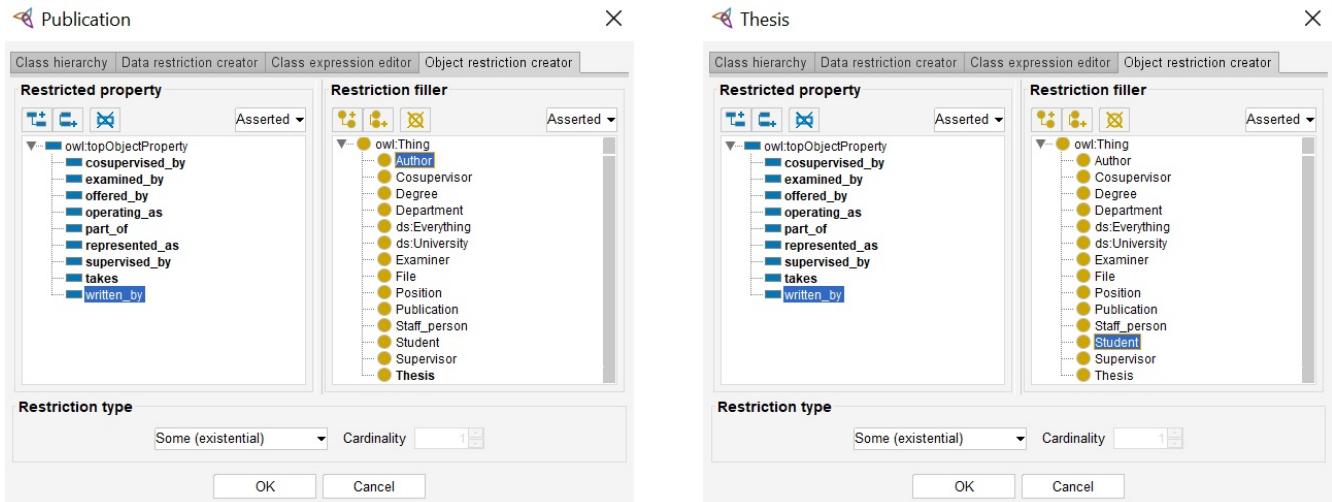
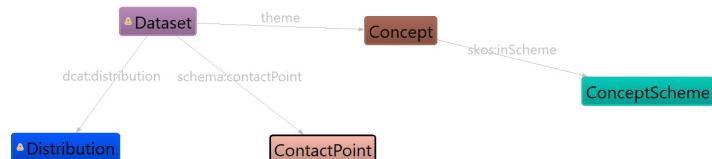


Figure 23: Object restrictions

6.6 Metadata

As usually, metadata have been generated for the teleology, which can be seen below and accessed here.



```

i 1@#prefix : <https://www.epos-eu.org/epos-dcat-ap#> .□
27
28@ <https://www.epos-eu.org/epos-dcat-ap#ContactPoint/9d7e7259-c526-44cd-bcc5-32acb63af9bd>
29    rdf:type          schema:ContactPoint ;
30    schema:availableLanguage "en" ;
31    schema:contactType "Student" ;
32    schema:email      "milena.atacasova@studenti.unitn.it" .
33
34@ <https://www.epos-eu.org/epos-dcat-ap#Concept/ad3e7bcc-453e-41f7-ba3d-716e7a22e656>
35    rdf:type          skos:Concept ;
36    skos:definition  "Teleology concerns the informal modeling phase of building a knowledge graph" ;
37    skos:inScheme     <https://www.epos-eu.org/epos-dcat-ap#ConceptScheme/44404d43-2a65-4bab-aa69-650874c665d8> ;
38    skos:prefLabel   "Teleology" .
39
40@ <https://www.epos-eu.org/epos-dcat-ap#Distribution/95552f13-a1cb-44d3-9c89-66877fe937d1>
41    rdf:type          dcat:Distribution ;
42    dct:description  "A teleology model built for the project Digital University Theses, Publications and Staff" ;
43    dct:format       "owl" ;
44    dct:identifier   "http://www.semanticweb.org/user/ontologies/2022/10/teleology/1.0.0"^^xsd:anyURI ;
45    dct:issued       "2022-11-23T22:25:21Z"^^xsd:dateTime ;
46    dct:language     "en" ;
47    dcat:accessURL   "https://github.com/atanasova16/TrentinoDUnitTPS/tree/main/Teleologies/Informal%20Modeling"^^xsd:anyURI .
48
49@ <https://www.epos-eu.org/epos-dcat-ap#ConceptScheme/44404d43-2a65-4bab-aa69-650874c665d8>
50    rdf:type          skos:ConceptScheme ;
51    dct:description  "Entities, data and object properties for the teleology of the project Digital University" ;
52    dct:title        "Teleology Digital University Theses, Publications, and Staff" .
53
54@ <https://www.epos-eu.org/epos-dcat-ap#Dataset/a5ecca98-9095-455c-9418-334536841165>
55    rdf:type          dcat:Dataset ;
56    dct:description  "Teleology model for the project Digital University Theses, Publications and Stuff." ;
57    dct:identifier   "https://knowdive.disi.unitn.it/etype#" ;
58    dct:language     "en" ;
59    dct:title        "Teleology" ;
60    dcat:contactPoint <https://www.epos-eu.org/epos-dcat-ap#ContactPoint/9d7e7259-c526-44cd-bcc5-32acb63af9bd> ;
61    dcat:distribution <https://www.epos-eu.org/epos-dcat-ap#Distribution/95552f13-a1cb-44d3-9c89-66877fe937d1> ;
62    dcat:keyword     "Protege" , "entities" , "teleology" , "knowledge graph" , "properties" , "itilos" ;
63    dcat:theme       <https://www.epos-eu.org/epos-dcat-ap#Concept/ad3e7bcc-453e-41f7-ba3d-716e7a22e656> .
64

```

Now that the ER model is ready, in the next phase it would be continued with the formal modeling where an ETG is generated, including also a common ontology, in order to increase and benefit from reusability.

7 Formal Modeling

7.1 Phase Description

Once the ER model is defined, the following step is the creation of the EER model, also named teleontology.

Teleontologies are fundamental structures in the KG process since they are necessary in order to provide in the model model the reusability property; indeed they can be seen as extended versions of Entities – Relationships model and define the final KG system.

In the previous ER model, E-types shown don't represent the general concepts and ideas of the world and this ends up to leave a gap on the process that has to be filled by the selection of an ontology.

Ontology provides an explicit specification of a formal structure of reality or a piece of that, without any link with the actual situation of a specific field and the vocabulary used.

The information coming from the chosen ontology will be joined with the teleology in order to generate the Teleontology structure, known as Extended Entity-Relationship model (EER model).

- **Input:** ER model produced during the last phase, plus the selected Ontology
- **Output:** Teleontology (EER model) thanks to the implementation of Ontology to Teleology.

7.2 Ontology

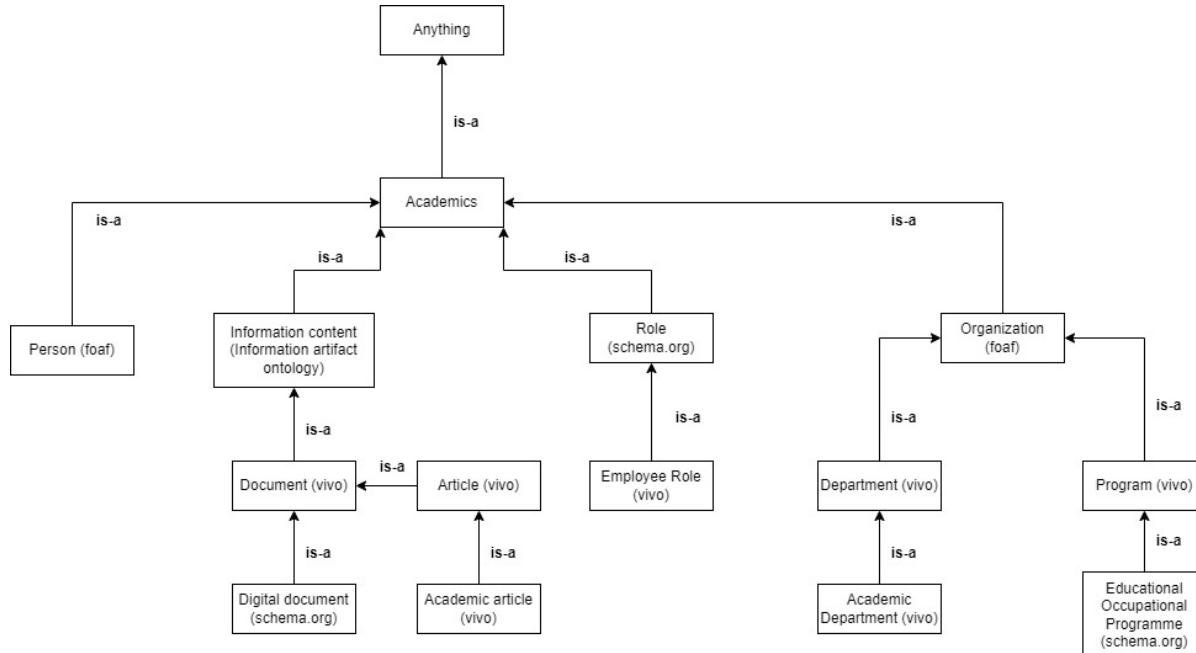


Figure 24: Ontology

The ontology shown is the result of our research for a schema that might be feasible having in mind the purpose. During the formulation it was considered that all the entities in the ontology were designed with the idea to collect and align them in a proper way with ER model.

One particular schema pre-defined was not enough to consider and relate all of the entity-types and concepts intended for the study of this project, therefore several knowledge sources were explored and different schema were combined in a way to identify one for Digital University Theses, Publications, and Staff, still not losing reusability. The schemas considered are:

- FOAF
- VIVO
- IAO
- Schema.org

A graph was obtained showing an exhaustive organization of the specific domain of knowledge where all the entities indicated are connected in a hierarchical structure with the top - level entity “Academics” which in turn is associated to everything.

As it is shown, the ontology contains the more general kinds of entities of the research activities domain and their interactions. Indeed all the entity types show a hierarchy among them and

the direction of relations starts from the leaves and then goes up until the “everything” describing a taxonomy distribution.

The Academics node is in a general way the representation in which the concepts of the domain of interest (and consequently purpose and the competency question) are involved. Academics is a non-tangible concept in which the users can ask and reach some questions about academic activities, from some other aspects it covers a fundamental role since it represents the main structure in which the employees and the academic locations are defined.

The ontology has a main feature defined as “conceptualization of a domain”: it allows the formal representation of reality as it is perceived. There are different levels of conceptualization in which the ontology proposed is organized.

The building process was done following a bottom-up approach, from the more specific concepts up to the more universal one wrapping each concept in the hierarchical structure. The approach followed allowed to look at teleology and see what kinds of classes are needed to properly represent the intended meaning and need. Moreover, for defining subclasses and therefore hierarchy in the model, the approach followed was to look at the types of hierarchy already present in the existing ontologies, if it is reasonable for the definitions defined so far.

As such, what was identified initially were the aspects of:

- **Academic Department:** A class from VIVO described as a "distinct, usually specialized educational unit within an educational organization". This class matches the meaning for the concept and it is of use for the purpose and that's why it was included.
- **Department:** then following the procedure mentioned above for grouping entities, Academic Department was mapped to Department from the same VIVO ontology ("A unit within a larger organization that addresses a specific subject or area of activity")
- **Educational Occupational Program:** For the purpose at hand, an entity describing the type of study is needed and after some research the concept from schema.org was found to be the most proper one: "A program offered by an institution which determines the learning progress to achieve an outcome, usually a credential like a degree or certificate"
- **Program:** in itself, the Educational Occupational Program is mapped to Program from VIVO ("An ongoing academic initiative not formalized with department or division status"). Then in themselves, **Department** and **Program** are both exhibiting an *is-a* relationship with the Organization type (description follows later).
- **Employee Role:** as Staff is one of the main concepts for the project, some research was done as to how exactly to represent the kind of job/career a person possess, and the solution chosen was to use the type Employee Role from VIVO ("A role inhering in a person that is realized when the bearer participates in an occupation by which a person earns a living or spends their time"). Its definition defines quite well what was intended for the purpose as it explains the kind of employment a person takes. It follows that it is mapped to the more general class **Role** with an *is-a* relation.
- **Digital document:** This type was considered to explain the representation of publications as files. Indeed, in schema.org it is "an electronic file or document", which is the concept searched for.

- **Academic article:** a very important class for this Knowledge graph, since it defines the concept of "*publication*" (described as "a scholarly academic article, typically published in a journal"). It is indeed what was searched for.
- **Article:** once more following hierarchy, Academic article shows an *is-a* to Article ("A written composition in prose, usually nonfiction, on a specific topic, forming an independent part of a periodical or book A written composition in prose, usually nonfiction, on a specific topic, forming an independent part of a book or other publication, as a newspaper or magazine"). Both Article and Digital document are connected with Document.
- **Document:** it is a broader concept from VIVO: "A bounded physical representation of a body of information designed with the capacity (and usually intent) to communicate". It captures the meaning and concepts of the types listed above in a higher level of generalization, useful to broaden the level of abstraction.

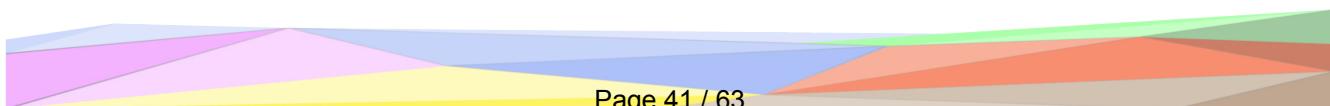
After having wrapped the classes above, and following the teleology created before the terms as first level of descriptive concepts which in themselves are connected to Academics are:

- **Person:** in this case standing all the people involved in academic affairs, for instance students, professors or researchers. The class Person was considered from the main general resource which models people *foaf*, which can provide additional properties. It is a very general concept, and broadly described as "representing people".
- **Information content:** for the purpose it represents all the possible way in which the information are carried through the academic channel. This class is coming from the Information Artifact Ontology (IAO). It was chosen to wrap Document (and therefore the lower-level types) in it as it provides a higher level of generality for this concept ².
- **Role:** it interprets the professional position which can be covered on this domain. This type was chosen from schema.org as a general class having the possibility to define a broad range of children classes to which Employee Role belongs to.
- **Organization:** it is a more abstract notion corresponding to some kinds of social institutions, expressing on the one hand the departments and on the other the programs.

As stated, the layers define the hierarchical structure in which Ontology is based on. Gradually the EER model organizes in a more general way the entities described before, starting from the most close concepts which will be connected to ER model, until the more general one which defines the domain of knowledge where the Purpose comes from. As a result a graph was obtained.

In order to satisfy the purpose and then select this Ontology, the Knowledge sources specified before were utilized. According to the domain, VIVO ontology is the most used to define the EER model since it contains the class of Document with the sub-classes of article and academic article, the entity Academic department, subclass of more general concept Department,

²In the IAO this class is defined as: "A generically dependent continuant that is about some thing". Examples of information content entities include journal articles, documents, graphs and data



the formal idea of Program and the sub-class Employee Role. Instead, *Schema.org* is an ontology defined to model some other concepts; the classes taken from there are Role, Digital document and Educational Occupational Program. On the other hand, *FOAF* ontology contains the attributes to model person and their relations: the main classes selected are Person and Organization.

All the classes of the ontologies proposed are connected between each other in order to better represent the formal organization of the academics at the University.

In terms of Organization's sub-classes, regarding modeling academics and research, to characterize the academic activities involved by university institute, we the *VIVO* type Program as a sub-class of Organization was considered and then, as a sub-class of it - Educational occupational program from *schema.org*. In a similar way, Role object identified from *schema.org* is explained as a parent class of Employee role to characterize the staff position inside the University structure. Instead, when it comes to the class named Information content, it wraps a broad range of concepts, between them the concept of document, which in itself includes our intended specifications for Article and Digital Document.

Below again is an extract from the *xml* file from the *draw.io* software and is available [here](#).

```
<mxfile host="app.diagrams.net" modified="2022-11-30T19:35:19.216Z" agent="5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/107.0.0.0 Safari/537.36"
  etag="j3Hyp53xKanck4n4e1FL" version="20.6.0" type="google">
<diagram id="wUgesENNNEPwLAgnNUK" name="ontology_final">
  <mxGraphModel dx="1967" dy="1630" grid="1" gridSize="10" guides="1" tooltips="1" connect="1" arrows="1" fold="1" page="1" pageScale="1" pageWidth="850" pageHeight="1100" math="0"
    shadow="0">
    <root>
      <mxCell id="0"/>
      <mxCell id="1" parent="0"/>
      <mxCell id="3cPxqGZLr0-C200Z15Re-1"
        style="edgeStyle=orthogonalEdgeStyle;rounded=0;orthogonalLoop=1;jettySize=auto;html=1;exitX=0.5;exitY=0;exitDx=0;exitDy=0;entryX=0.5;entryDx=0;entryDy=0;" edge="1"
        parent="1" source="3cPxqGZLr0-C200Z15Re-2" target="3cPxqGZLr0-C200Z15Re-27">
        <mxGeometry relative="1" as="geometry"/>
      </mxCell>
      <mxCell id="3cPxqGZLr0-C200Z15Re-2" value="Employee Role (vivo)" style="whiteSpace=wrap;html=1;align=center;" vertex="1" parent="1">
        <mxGeometry x="90" y="-270" width="100" height="40" as="geometry"/>
      </mxCell>
      <mxCell id="3cPxqGZLr0-C200Z15Re-3"
        style="edgeStyle=orthogonalEdgeStyle;rounded=0;orthogonalLoop=1;jettySize=auto;html=1;exitX=0.5;exitY=0;exitDx=0;exitDy=0;entryX=0.5;entryDx=0;entryDy=0;" edge="1"
        parent="1" source="3cPxqGZLr0-C200Z15Re-4" target="3cPxqGZLr0-C200Z15Re-28">
        <mxGeometry relative="1" as="geometry"/>
      </mxCell>
      <mxCell id="3cPxqGZLr0-C200Z15Re-4" value="Digital document (schema.org)" style="whiteSpace=wrap;html=1;align=center;" vertex="1" parent="1">
        <mxGeometry x="200" y="-170" width="100" height="40" as="geometry"/>
      </mxCell>
      <mxCell id="3cPxqGZLr0-C200Z15Re-5"
        style="edgeStyle=orthogonalEdgeStyle;rounded=0;orthogonalLoop=1;jettySize=auto;html=1;exitX=0.5;exitY=0;exitDx=0;exitDy=0;entryX=0.5;entryDx=0;entryDy=0;" edge="1"
        parent="1" source="3cPxqGZLr0-C200Z15Re-7" target="3cPxqGZLr0-C200Z15Re-29">
        <mxGeometry relative="1" as="geometry"/>
      </mxCell>
      <mxCell id="3cPxqGZLr0-C200Z15Re-6" style="edgeStyle=orthogonalEdgeStyle;rounded=0;orthogonalLoop=1;jettySize=auto;html=1;entryX=1;entryY=0.5;entryDx=0;entryDy=0;" edge="1"
        parent="1" source="3cPxqGZLr0-C200Z15Re-7" target="3cPxqGZLr0-C200Z15Re-28">
        <mxGeometry relative="1" as="geometry"/>
      </mxCell>
      <mxCell id="3cPxqGZLr0-C200Z15Re-7" value="Article (vivo)" style="whiteSpace=wrap;html=1;align=center;" vertex="1" parent="1">
        <mxGeometry x="-50" y="-270" width="100" height="40" as="geometry"/>
      </mxCell>
      <mxCell id="3cPxqGZLr0-C200Z15Re-8"
        style="edgeStyle=orthogonalEdgeStyle;rounded=0;orthogonalLoop=1;jettySize=auto;html=1;exitX=0.5;exitY=0;exitDx=0;exitDy=0;entryX=0.5;entryDx=0;entryDy=0;" edge="1"
        parent="1" source="3cPxqGZLr0-C200Z15Re-8" target="3cPxqGZLr0-C200Z15Re-22">
        <mxGeometry relative="1" as="geometry"/>
      </mxCell>
    </root>
  </mxGraphModel>
</diagram>
</mxfile>
```

Figure 25: Extract of ontology graph *xml*

7.3 Ontology Implementation in Protege

The diagram defined above was then implemented in the program Protege, defining all the entities with reference to the respective schemas.

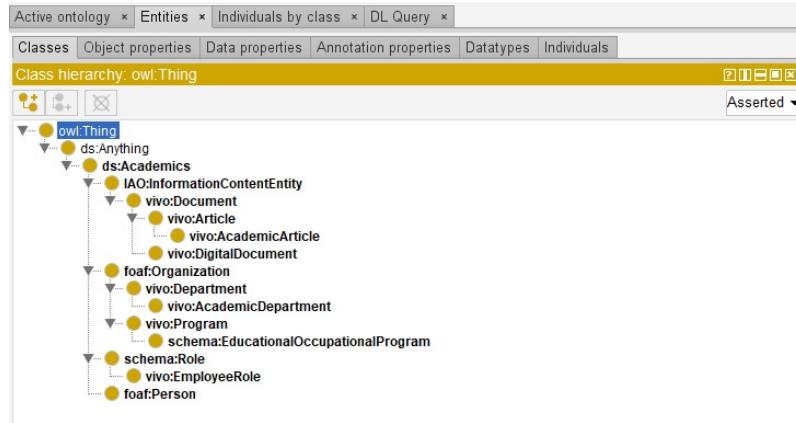


Figure 26: Entities in Ontology in Protege

And, as usual, metadata file about the creation of the ontology is shown below and with an access point here.

```

1@prefix : <https://www.epos-eu.org/epos-dcat-ap#> .
2
3<https://www.epos-eu.org/epos-dcat-ap#Concept/923e10c4-f659-49c4-9f44-3cf2f3f195ab>
4  rdf:type skos:Concept ;
5  skos:definition "Ontology for a Knowledge graph" ;
6  skos:inScheme <https://www.epos-eu.org/epos-dcat-ap#ConceptScheme/cadbf35a-1893-4c62-b1eb-5164f2848b1c> ;
7  skos:prefLabel "Ontology" .
8
9<https://www.epos-eu.org/epos-dcat-ap#ContactPoint/a2819333-1fd2-4796-b8f5-201090009a9d>
10  rdf:type schema:ContactPoint .
11
12<https://www.epos-eu.org/epos-dcat-ap#Distribution/67191ce0-11f1-4d6a-be0f-389ac825fbea>
13  rdf:type dcat:Distribution ;
14  dct:description "Ontology built by combining knowledge sources for the project Digital University Theses, Publications and Staff" ;
15  dct:format "owl" ;
16  dct:identifier "https://knowdive.disi.unitn.it/etype#" ;
17  dct:issued "2022-12-03T11:58:42Z^^xsd:dateTime" ;
18  dct:language "en" ;
19  dcat:accessURL "https://github.com/atanasova16/TrentinoDUnitPS/blob/main/Teleologies/Informal%20Modeling/ontology2.owl"^^xsd:anyURI .
20
21<https://www.epos-eu.org/epos-dcat-ap#ConceptScheme/cadbf35a-1893-4c62-b1eb-5164f2848b1c>
22  rdf:type skos:ConceptScheme ;
23  dct:description "Ontology created with Protege about entities from schemas for the Digital University project" ;
24  dct:title "Ontology Digital University" .
25
26<https://www.epos-eu.org/epos-dcat-ap#Dataset/c3cb5184-42a3-4349-9d03-02c22501e6ca>
27  rdf:type dcat:Dataset ;
28  dct:created "2022-12-02T18:52:29Z^^xsd:dateTime" ;
29  dct:description "The ontology constructed for the project about Digital University built from various ontologies" ;
30  dct:identifier "https://knowdive.disi.unitn.it/etype#" ;
31  dct:issued "2022-12-02T19:52:29Z^^xsd:dateTime" ;
32  dct:language "en" ;
33  dct:title "Ontology Digital University Theses, Publications and Staff" ;
34  dcat:contactPoint <https://www.epos-eu.org/epos-dcat-ap#ContactPoint/4ee429f9-0a4f-474f-9123-a030bd922fd6> ;
35  dcat:distribution <https://www.epos-eu.org/epos-dcat-ap#Distribution/67191ce0-11f1-4d6a-be0f-389ac825fbea> ;
36  dcat:keyword "protege", "ontology", "schemas", "data properties", "object properties", "entities", "knowledge graph" ;
37  dcat:theme <https://www.epos-eu.org/epos-dcat-ap#Concept/923e10c4-f659-49c4-9f44-3cf2f3f195ab> .

```

Figure 27: Metadata for ontology

7.4 Teleontology

Combining built teleology with identified ontology, the teleontology appears. The diagram follows.

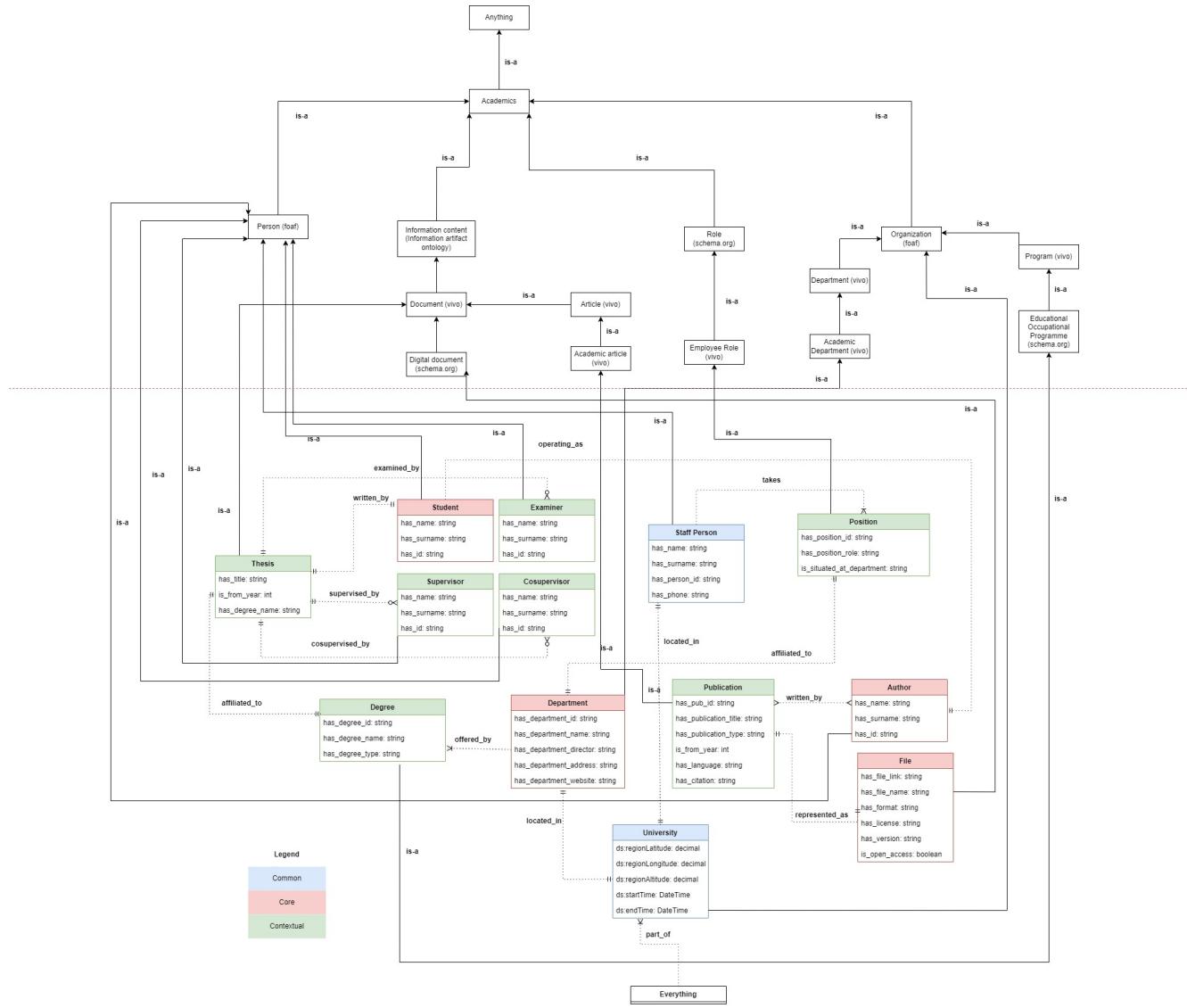


Figure 28: Teleontology

On the graph, it can be seen in detail the whole set of interactions in the form of object properties (for reference, look at the Informal modeling phase), as well as the hierarchies defined (in the sub-section about Ontology), all combined in a single diagram. Above the horizontal line is the ontology, and below it - the teleology.

A particular mapping of entities from teleology was done to the respective matches in the ontology. In such a way, entities could if needed further inherit properties from the parent classes should data resources were enriched.

The entity Person from the FOAF vocabulary is a common concept to which six of the definitions were mapped: Student, Supervisor, Cosupervisor, Examiner, Author, and Staff person. This is done following the thought they are people and have the characteristics, therefore they are some kind of subclasses. Moreover, apart from Staff person, all others do not have many data properties, so a more general concept captures them.

The E-type Position from the ER model is mapped to Employee Role because it identifies the position a person takes in a job. The description of Employee Role matches what was identified as a position, indeed a role for occupational activity.

File, on the other hand, for what was defined and how was it extracted, is actually the electronic representation of a Publication, and therefore has an *is-a* link to Digital Document.

Publication in itself is an Academic article because from the data sources, the publications are of those kinds having to do with *research*, already identified as a main purpose of the KG.

Thesis E-type is linked instead directly to Document - a connection which is defined due to the quite different nature of the concept of a thesis, but still it is a document, and is conforming to the definition.

The Degree E-type from the ER model is in fact an Educational Occupational Programme as it is the particular followed course of study combining in itself courses and activities from the university.

When it comes to Department, the linkage was quite straightforward, since a university department (faculty) is an Academic department as its nature.

Finally, university is an organization and that is why it has a link to this part of the Ontology. A university collects in itself many issues. For this case it is also used to wrap all the concepts in the spatial and time frame.

Overall, these linkages were considered in order to give the possibility both for constructing this particular Knowledge Graph, as well as to provide the possibility to afterwards enrich with more object and data properties, when more diverse data could potentially be collected. At the current stage this is thought as a solution fitting these two complementary goals.

Still, it needs to be considered that there could be some open issues also for this teleology building phase, as to whether there may be a better way to explain the relationships between the classes, and if there are some interactions not considered. Possibly further research could be conducted in this phase.

Below, the *xml* file for the diagram created with *draw.io*. Here is a link to it.

```

<?xml version="1.0" encoding="UTF-8"?>
<mxfile host="app.diagrams.net" modified="2022-11-30T19:52:01.296Z" agent="5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/107.0.0.0 Safari/537.36" etag="z14USInaXvKvcU1b4B" version="20.6.0" type="google">
  <diagram name="Teleontology_final" id="yKv0HmUBeavwIeB2Hr">
    <mxGraphModel dx="384" dy="2160" grid="1" guides="1" tooltips="1" connect="1" arrows="1" fold="1" page="1" pageScale="1" pageWidth="850" pageHeight="1100" math="0" shadow="0" extFonts="PermanentMarker">
      <!-- https://fonts.googleapis.com/css?family=Permanent+Marker -->
      <root>
        <mxCell id="D_w_EhE-Y8vd0Tdokrf-0"/>
        <mxCell id="D_w_EhE-Y8vd0Tdokrf-1" parent="D_w_EhE-Y8vd0Tdokrf-0"/>
        <mxCell id="D_w_EhE-Y8vd0Tdokrf-2" parent="D_w_EhE-Y8vd0Tdokrf-1" style="outline:none;fontStyle:italic;childLayoutStackLayout;horizontalAlign1;startSizE=26;horizontalStack=0;resizeParentMax=0;resizeLast=0;collapsible=1;marginBottom=0;fillColor=#f8ecc;strokeColor=#B85450;" vertex="1">
          <mxGeometry x="570" y="610" width="160" height="182" as="geometry"/>
        </mxCell>
        <mxCell id="D_w_EhE-Y8vd0Tdokrf-3" parent="D_w_EhE-Y8vd0Tdokrf-2" style="outline:none;fontStyle:italic;childLayoutStackLayout;horizontalAlign1;startSizE=26;horizontalStack=0;resizeParentMax=0;resizeLast=0;collapsible=1;marginBottom=0;fillColor=#f8ecc;strokeColor=#B85450;" vertex="1">
          <mxGeometry x="520" y="720" width="160" height="180" as="geometry"/>
        </mxCell>
        <mxCell id="D_w_EhE-Y8vd0Tdokrf-4" parent="D_w_EhE-Y8vd0Tdokrf-3" style="outline:none;fontStyle:italic;childLayoutStackLayout;horizontalAlign1;startSizE=26;horizontalStack=0;resizeParentMax=0;resizeLast=0;collapsible=1;marginBottom=0;fillColor=#f8ecc;strokeColor=#B85450;" vertex="1">
          <mxGeometry x="520" y="830" width="160" height="180" as="geometry"/>
        </mxCell>
        <mxCell id="D_w_EhE-Y8vd0Tdokrf-5" parent="D_w_EhE-Y8vd0Tdokrf-4" style="outline:none;fontStyle:italic;childLayoutStackLayout;horizontalAlign1;startSizE=26;horizontalStack=0;resizeParentMax=0;resizeLast=0;collapsible=1;marginBottom=0;fillColor=#f8ecc;strokeColor=#B85450;" vertex="1">
          <mxGeometry x="520" y="940" width="160" height="180" as="geometry"/>
        </mxCell>
        <mxCell id="D_w_EhE-Y8vd0Tdokrf-6" parent="D_w_EhE-Y8vd0Tdokrf-5" style="outline:none;fontStyle:italic;childLayoutStackLayout;horizontalAlign1;startSizE=26;horizontalStack=0;resizeParentMax=0;resizeLast=0;collapsible=1;marginBottom=0;fillColor=#f8ecc;strokeColor=#B85450;" vertex="1">
          <mxGeometry x="520" y="1050" width="160" height="180" as="geometry"/>
        </mxCell>
        <mxCell id="D_w_EhE-Y8vd0Tdokrf-7" parent="D_w_EhE-Y8vd0Tdokrf-6" style="outline:none;fontStyle:italic;childLayoutStackLayout;horizontalAlign1;startSizE=26;horizontalStack=0;resizeParentMax=0;resizeLast=0;collapsible=1;marginBottom=0;fillColor=#f8ecc;strokeColor=#B85450;" vertex="1">
          <mxGeometry x="520" y="1160" width="160" height="180" as="geometry"/>
        </mxCell>
        <mxCell id="D_w_EhE-Y8vd0Tdokrf-8" parent="D_w_EhE-Y8vd0Tdokrf-7" style="outline:none;fontStyle:italic;childLayoutStackLayout;horizontalAlign1;startSizE=26;horizontalStack=0;resizeParentMax=0;resizeLast=0;collapsible=1;marginBottom=0;fillColor=#f8ecc;strokeColor=#B85450;" vertex="1">
          <mxGeometry x="520" y="1270" width="160" height="180" as="geometry"/>
        </mxCell>
        <mxCell id="D_w_EhE-Y8vd0Tdokrf-9" parent="D_w_EhE-Y8vd0Tdokrf-8" style="outline:none;fontStyle:italic;childLayoutStackLayout;horizontalAlign1;startSizE=26;horizontalStack=0;resizeParentMax=0;resizeLast=0;collapsible=1;marginBottom=0;fillColor=#f8ecc;strokeColor=#B85450;" vertex="1">
          <mxGeometry x="520" y="1380" width="160" height="180" as="geometry"/>
        </mxCell>
        <mxCell id="D_w_EhE-Y8vd0Tdokrf-10" parent="D_w_EhE-Y8vd0Tdokrf-9" style="outline:none;fontStyle:italic;childLayoutStackLayout;horizontalAlign1;startSizE=26;horizontalStack=0;resizeParentMax=0;resizeLast=0;collapsible=1;marginBottom=0;fillColor=#f8ecc;strokeColor=#B85450;" vertex="1">
          <mxGeometry x="520" y="1490" width="160" height="180" as="geometry"/>
        </mxCell>
        <mxCell id="D_w_EhE-Y8vd0Tdokrf-11" parent="D_w_EhE-Y8vd0Tdokrf-10" style="outline:none;fontStyle:italic;childLayoutStackLayout;horizontalAlign1;startSizE=26;horizontalStack=0;resizeParentMax=0;resizeLast=0;collapsible=1;marginBottom=0;fillColor=#f8ecc;strokeColor=#B85450;" vertex="1">
          <mxGeometry x="520" y="1600" width="160" height="180" as="geometry"/>
        </mxCell>
        <mxCell id="D_w_EhE-Y8vd0Tdokrf-12" parent="D_w_EhE-Y8vd0Tdokrf-11" style="outline:none;fontStyle:italic;childLayoutStackLayout;horizontalAlign1;startSizE=26;horizontalStack=0;resizeParentMax=0;resizeLast=0;collapsible=1;marginBottom=0;fillColor=#f8ecc;strokeColor=#B85450;" vertex="1">
          <mxGeometry x="520" y="1710" width="160" height="180" as="geometry"/>
        </mxCell>
      </root>
    </mxGraphModel>
  </diagram>
</mxfile>

```

Figure 29: Extract of teleontology graph *xml*

7.5 Teleontology Implementation in Protege

Having defined both the teleology in Protege (in the Informal modeling phase) as well as the ontology (in the sub-section about Ontology), it was proceeded with importing these two *owl* files and creating the class hierarchy within it. Now the resulting *owl* output file contains all the e-types, data properties, and object properties needed for the Formal Modeling phase completion.

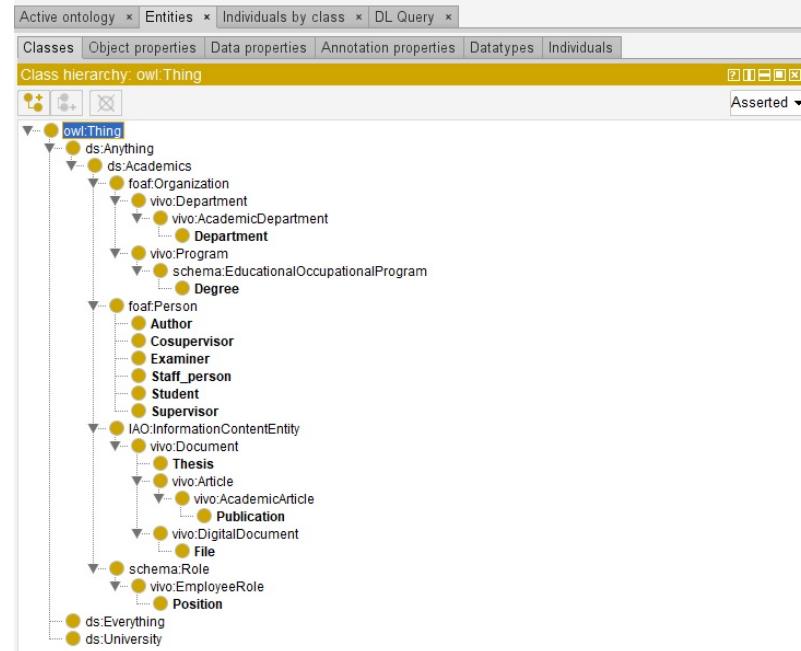
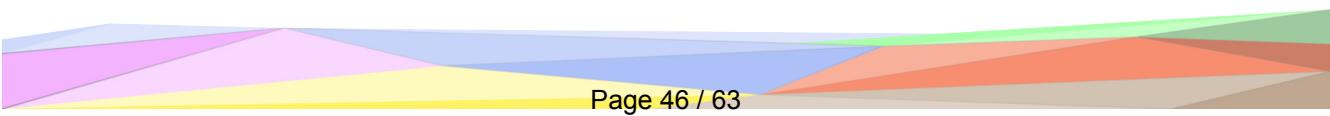


Figure 30: Entities in Teleontology in Protege



Later, due to some issues about imports, a new teleontology OWL file was constructed from scratch, containing all the previously defined entities, data properties, and object properties, linked accordingly.

Again a metadata file was prepared using the Shapeness software to complement activities from the project with access to the file here.

```

1@prefix : <https://www.epos-eu.org/epos-dcat-ap#> .□
2@<https://www.epos-eu.org/epos-dcat-ap#Dataset/8b12c36b-078a-4cff-a940-e9e79311495b>
3  rdf:type          dcat:Dataset ;
4  dct:description   "Teleontology built for the project Digital University These Publications and Staff" ;
5  dct:identifier    "https://knowdive.disi.unitn.it/etypex^xsd:anyURI" ;
6  dct:issued        "2022-12-03T14:08:03Z^^xsd:dateTime" ;
7  dct:language      "en" ;
8  dct:title         "Teleontology" ;
9  dcat:contactPoint <https://www.epos-eu.org/epos-dcat-ap#ContactPoint/fe6c2697-cd86-43a6-a7b9-ec3ac915d4ec> ;
10 dcat:distribution <https://www.epos-eu.org/epos-dcat-ap#Distribution/5e814875-c050-462e-8573-8e4ea4210dab> ;
11 dcat:keyword       "knowledge graph" , "properties" , "language annotation" , "entities" , "teleontology" ;
12 dcat:theme         <https://www.epos-eu.org/epos-dcat-ap#Concept/29296612-5793-4666-bcd8-21b88dd27859> .
13@<https://www.epos-eu.org/epos-dcat-ap#Distribution/5e814875-c050-462e-8573-8e4ea4210dab>
14  rdf:type          dcat:Distribution ;
15  dct:description   "Teleontology for the project Digital University Theses, Publications and Staff, built by combining the previously defined teleology and the identified ontology. It is annotated according to the language annotation" ;
16  dct:format        "owl" ;
17  dct:identifier    "https://knowdive.disi.unitn.it/etypex^xsd:anyURI" ;
18  dct:issued        "2022-12-03T13:23:14Z^^xsd:dateTime" ;
19  dct:language      "en" ;
20  dcat:accessURL   "https://github.com/atanasova16/TrentinoDUnitTPS/blob/main/Teleologies/Formal%20Modeling/TrentinoDUnitTPS-5v0-annotatedETG.owl" .
21@<https://www.epos-eu.org/epos-dcat-ap#ContactPoint/fe6c2697-cd86-43a6-a7b9-ec3ac915d4ec>
22  rdf:type          schema>ContactPoint ;
23  schema:availableLanguage "it" , "en" ;
24  schema:contactType "Student" ;
25  schema:email       "milena.atanasova@studenti.unitn.it" .
26@<https://www.epos-eu.org/epos-dcat-ap#ConceptScheme/3900322c-b407-4f92-bd8d-3c190702f523>
27  rdf:type          skos:ConceptScheme ;
28  dct:description   "Theses, Publications and stuff for a Digital University" ;
29  dct:title         "Digital University" .
30@<https://www.epos-eu.org/epos-dcat-ap#Concept/29296612-5793-4666-bcd8-21b88dd27859>
31  rdf:type          skos:Concept ;
32  skos:definition  "Digital University ontology, defining the entities, object and data properties" ;
33  skos:inScheme     <https://www.epos-eu.org/epos-dcat-ap#ConceptScheme/3900322c-b407-4f92-bd8d-3c190702f523> ;
34  skos:prefLabel    "Teleontology" .

```

Figure 31: Metadata for teleontology owl

7.6 Language Annotation

The next step in the Formal Modeling phase concerns the language annotation. The main resource is the Universal Knowledge Core (UKC), which is a "multilingual, high quality, large-scale, machine-readable, and diversity-aware lexical resource" ³. The principle is to maintain a clear distinction between the language(s) used to describe the world as it is perceived and what is being described, i.e., the world itself. This system has several core concepts which describe it.

The Concept Core is the UKC representation of the world, it makes nodes language independent concepts. Synsets are lexicalized concepts, which consist of set of synonymous words. The Language Core (LC) in the UKC stores the words, senses, synsets, glosses and examples for all the languages supported by the UKC. In the LC each synset is univocally associated with one language and, within that language, with at least one word. Each synset is linked to one and only one concept.

With the UKC, it becomes easier to provide a more formal way of expressing the informal terms of the teleontology, by representing each informal concept with an identifier from UKC. Further, when an ontology is introduced, it is always searched to align the existing concepts from it with their equivalent in UKC (or adding new ones in UKC if they do not exist). Moreover, each concept could potentially be rendered multilingually which would help in adapting the knowledge in other languages and making use of it in different countries for similar purposes.

³<http://ukc.disi.unitn.it/>

The procedure is that each term from an ontology is searched in the Knowledge Base. If there is a synonymous match, then the term is mapped to this concept. Otherwise, via the annotator tool is declared a new concept.

Such a procedure allows for multiple knowledge resources to be mapped to a single big background knowledge.

7.7 Language Annotation Implementation

The software used for this project (KOS) provides an annotator tool with a graphical interface in order to facilitate the alignment of word/concept meanings.

The issues encountered after uploading the teleontology concern some of the terms. After the exploration of the possible descriptions, the concepts were mapped to their respective meaning, and if a synset about the meaning in mind was not present, it was added, thus potentially enlarging the lexicon resource. Below are shown the issues and how they were mapped.

These meanings were given following the way of the thought about what these terms describe in the ontology.⁴. Afterwards, the new owl file has the annotation as a GID after each entity name. Below there are two photos of how the object and data properties look like when annotated.



Figure 32: Annotated OP



Figure 33: Annotated DP

After completion of formal modeling, next follows the data integration part, which leads to the construction of a complete knowledge graph.

⁴As a disclaimer: for the purposes of using the knowledge graph construction application KOS, some of the object and data properties were renamed to begin with "has" as is stated in the application guidelines, in order to be able to input the teleontology model.

Cased Word Lemma	Concept UK ID	Description
supervised_by	112436	watch and direct
cosupervised_by	112436	watch and direct
examined_by	104182	examine someone's knowledge of something
written_by	108775	produce a literary work
takes	102927	take on a certain form, attribute, or aspect
offered_by	-1	provide possibility to subscribe to a program/course/event
operating_as	112129	perform a function
represented_as	104885	give expression to
affiliated_to	43546	a subsidiary or subordinate organization that is affiliated with another organization
located_in	111901	assign a location to
part_of	31976	one of the portions into which something is regarded as divided and which together constitute a whole
Publication	-2	an abstract, article or paper in a journal or electronic repository
Cosupervisor	-3	someone who is a secondary supervisor
File	-4	an electronic representation of a document
Staff_person	45612	personnel who assist their superior in carrying out an assigned task
has_department	43988	a specialized division of a large organization
has_cosupervisor	57473	one who supervises or has charge and direction of
has_location	779	a determination of the place where something is
has_affiliation_to	43546	a subsidiary or subordinate organization that is affiliated with another organization
has_examiner	53882	someone who administers a test to determine your qualifications
has_representation	104395	express indirectly by an image, form, or model; be a symbol
has_author	58221	writes (books or stories or articles or the like) professionally (for pay)
has_supervisor	57473	one who supervises or has charge and direction of
has_operating_as	112129	perform a function
has_position	3017	a job in an organization
has_student_id	57408	a learner who is enrolled in an educational institution
has_open_access	111484	obtain or retrieve from a storage device; as of information on a computer
has_author_surname	34003	the name used to identify the members of a family (as distinguished from each member's given name)
has_department_director	-5	someone who is the main responsible person for a department
has_degree_id	35898	an award conferred by a college or university signifying that the recipient has satisfactorily completed a course of study
has_thesis_degree	35789	an integrated course of academic studies
has_supervisor_surname	34003	the name used to identify the members of a family (as distinguished from each member's given name)
has_student_name	2	a language unit by which a person or thing is known
has_cosupervisor_name	2	a language unit by which a person or thing is known
has_name	2	a language unit by which a person or thing is known
has_author_name	2	a language unit by which a person or thing is known

Cased Word Lemma	Concept UK ID	Description
has_title	34061	a general or descriptive heading for a section of a written work
has_supervisor_name	2	a language unit by which a person or thing is known
has_version	31829	a category of things distinguished by some common characteristic or quality
has_citation	36248	a short note recognizing a source of information or of a quoted passage
has_department_name	43988	a specialized division of a large organization
has_thesis_year	80974	a period of time containing 365 (or 366) days
has_author_id	58221	writes (books or stories or articles or the like) professionally (for pay)
has_license	-6	kind of permissions that a certain document has
has_phone	34494	the number is used in calling a particular telephone
has_format	-7	the electronic format of a file
has_student_surname	34003	the name used to identify the members of a family (as distinguished from each member's given name)
has_thesis_id	34396	a treatise advancing a new point of view resulting from research; usually a requirement for an advanced academic degree
has_degree_name	35898	an award conferred by a college or university signifying that the recipient has satisfactorily completed a course of study
has_surname	34003	the name used to identify the members of a family (as distinguished from each member's given name)
has_cosupervisor_surname	34003	the name used to identify the members of a family (as distinguished from each member's given name)
has_language	33843	a system of words used to name things in a particular discipline
has_pub_title	34061	a general or descriptive heading for a section of a written work
has_department_website	34126	a computer connected to the internet that maintains a series of web pages on the World Wide Web
has_degree_type	35898	an award conferred by a college or university signifying that the recipient has satisfactorily completed a course of study
has_person_id	118	a human being
has_department_id	43988	a specialized division of a large organization
has_examiner_name	2	a language unit by which a person or thing is known
has_pub_type	31834	a subdivision of a particular kind of thing
has_pub_id	39085	a symbol that establishes the identity of the one bearing it
has_examiner_id	39085	a symbol that establishes the identity of the one bearing it
has_position_id	39085	a symbol that establishes the identity of the one bearing it
has_cosupervisor_id	39085	a symbol that establishes the identity of the one bearing it
has_supervisor_id	39085	a symbol that establishes the identity of the one bearing it
has_position_situated_department	46413	physical position in relation to the surroundings
has_file_link	-8	a reference to access a website
has_pub_year	80974	a period of time containing 365 (or 366) days
has_position_role	3735	the actions and activities assigned to or required or expected of a person or group
has_file_name	2	a language unit by which a person or thing is known
has_examiner_surname	34003	the name used to identify the members of a family (as distinguished from each member's given name)
has_department_address	45803	the place where a person or organization can be found or communicated with

8 KGC

The final phase of the creation of a knowledge graph following the iTelos methodology consists of merging the knowledge and data layers into a single resource, which can be exploited by the users. The previous results are composed here.

The crucial step is to map the formatted datasets to the teleontology, thus adopting a representation of all what is included in the datasets.

- **Input:** the final KG's ETG, and the set of formal data resources.
- **Output:** the final KG

There are three main steps of this last phase: Identity problem, Entity matching and KG evaluation.

8.1 Entity Identification. Semantic heterogeneity

This problem concerns the presence of various representations of the same entities from the real world within the data resources. This issue occurs on two levels: first there is the need for identification within a single dataset, and then, within different datasets in order to be able to join different data by this identification.

In this project at hand for some of the resources, there exist already predefined identifiers. Those are all the datasets concerning people: Students, Supervisors, Authors etc. They are unique codes assigned for each person. For those who had missing values for the field, they were added manually (see section 5). If some of the people are also part of the staff dataset, the same identifier is used. Already existing identifier codes have also the departments of the University. All those act as URIs (Uniform Resource Identifiers).

However, when an identifier is not available, as is in the case of the other datasets, the entities can be identified by its properties, i.e. by composing an Identifying set (IS) of attributes, enough to uniquely identify an entity.

For files, an IS was initially identified by the link, as they are unique for each file. However, after trying to deal with this in Karma, it was noted that there are problems using the link as a URI, and that is why the file name with removed punctuation and added an indexing integer was used instead. About positions, the identifying set was created as a combination of the role and the department to which it is affiliated, and afterwards these were aligned with the staff dataset where each position is mapped to the particular person using the same newly created identifying attribute. For theses, the title of each work is unique for each thesis, so it is enough to identify it. However, as the title would actually be a data property of an entity of type Thesis, a separate attribute "id" is created.

Instead, for publications, the title is not enough to serve as an identifying set, because there exist duplicate titles in this case. Moreover, not only this, but even if combining title with type, year and language, still there are duplicated samples. Therefore, for this case, an identifying set will be the title, the year and the authors. In this way at further development, the entities could be identified like this, and could potentially benefit from adding new attributes from other data

sources concerning this entity (more on this on the next subsection). However, for the purpose of the knowledge graph for not making the *id* so long, a new attribute "pub_id" was created which is unique for each single entity of this type following the structure "*pub{index}*".

Below is a summary table of how the identifiers are for each of the entity types.

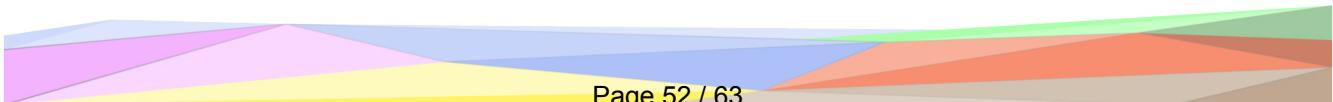
Staff Person	identifier string present already; if missing, added manually as namesurname_{index}	String
Author	identifier string present already; if missing, added manually as namesurname_{index}	String
Student	identifier string present already; if missing, added manually as namesurname_{index}	String
Supervisor	identifier string present already; if missing, added manually as namesurname_{index}	String
Cosupervisor	identifier string present already; if missing, added manually as namesurname_{index}	String
Examiner	identifier string present already; if missing, added manually as namesurname_{index}	String
Publication	IS was identified as one containing the title, year and name and surname of author; However, for brevity, an id is identified as: pub and an integer	String
File	filename and integer	String
Thesis	title is enough to serve as an IS; for brevity instead, thesis and an integer is used as an id	String
Deparment	identifier string present already	String
Degree	degreename without punctuation, lowercase and stripped white spaces	String
Position	Role and department id	String

8.2 Entity Matching

The next issue concerning the data integration part is the problem of matching entities, because in different datasets they can be represented with different properties (on schema level) or property values (on data layer). In the former case, it is actually useful, while in the later, there is a conflict, so the best way to cope with that is to look at metadata considering the authors, provenance and recency, or eventually to have a personal contact if possible to verify important issues.

Since still at the initial inception phase of this project, the datasets were divided in such a way as for one dataset to represent only one entity, at least from this point of view, it is more useful to have done that. In this way for this aptricular knowledge graph, there should not be duplicates in the entities, apart from the staff persons who could be present as supervisors/co-supervisors/examiners/students. But in this case, they would enrich in properties with possibly telephone number, and the properties about position, which is linking them to additional information. All this shows the power of links in the knowledge graph.

In terms of provenance of the data sources here, they come from the Open Data Trentino portal, and should also be in line with the FAIR principles. Furthermore, they are also quite recent, so they should be trustable.



8.3 Implementation using the data mapping tool Karma

The Karma data linker allows to efficiently and seamlessly manage the URI-s and the entity matching problem. For the implementation of this part, once the software was open, the teleontology `owl` file is imported, and afterwards, dataset by dataset as well, and linked. When there are some connections (Object properties), in this case a URI of the other class is identified and an outgoing link from the main entity to the range entity is added representing this relationship. Here are visually how these look for two of the E-types.

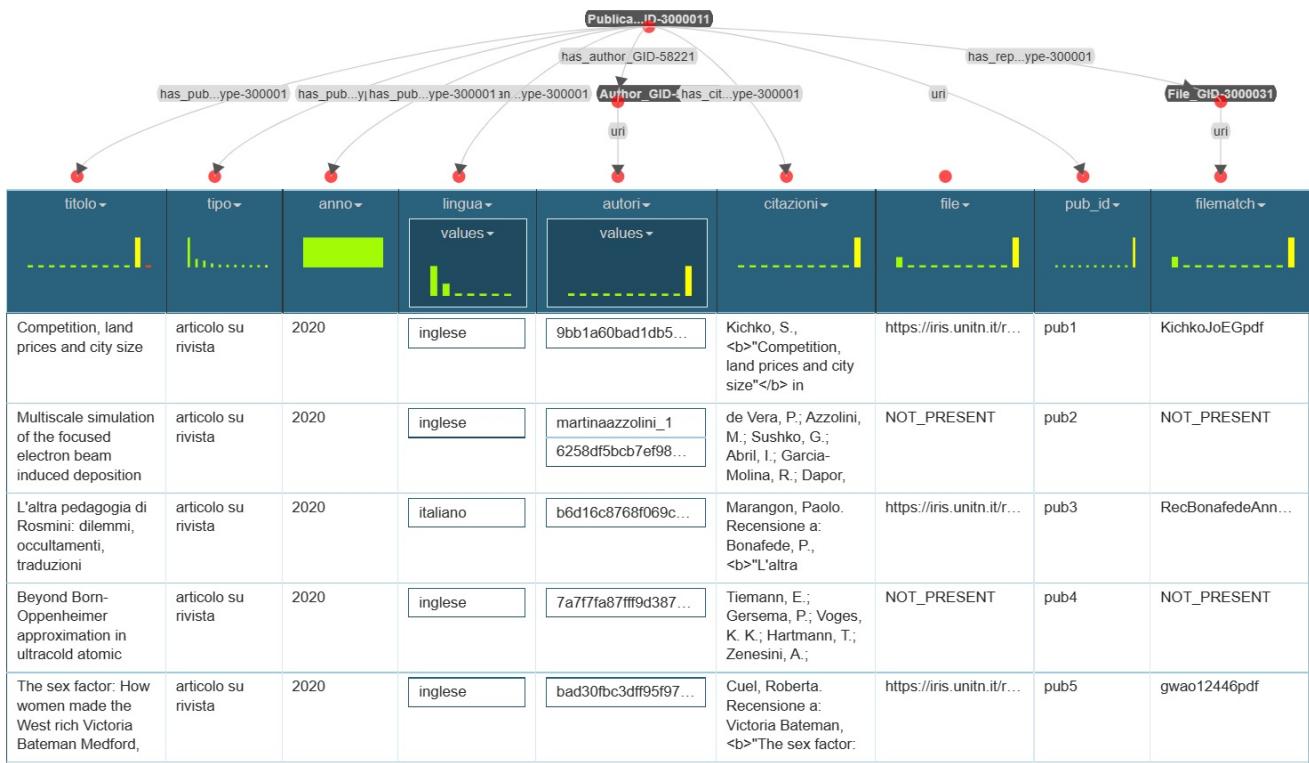


Figure 34: Publication data mapping with Karma

Here, the E-type Publication has its data properties linked directly to the columns, as well as an its own URI. The column `filermatch` contains the URI of Files, so its semantic type is set as this one. Then the outgoing link from class Publication is set to class File and the object property "has_representation_as" is added. The same for `has_author` connection with the Author entity with its uri, that is linking to the Authors dataset.

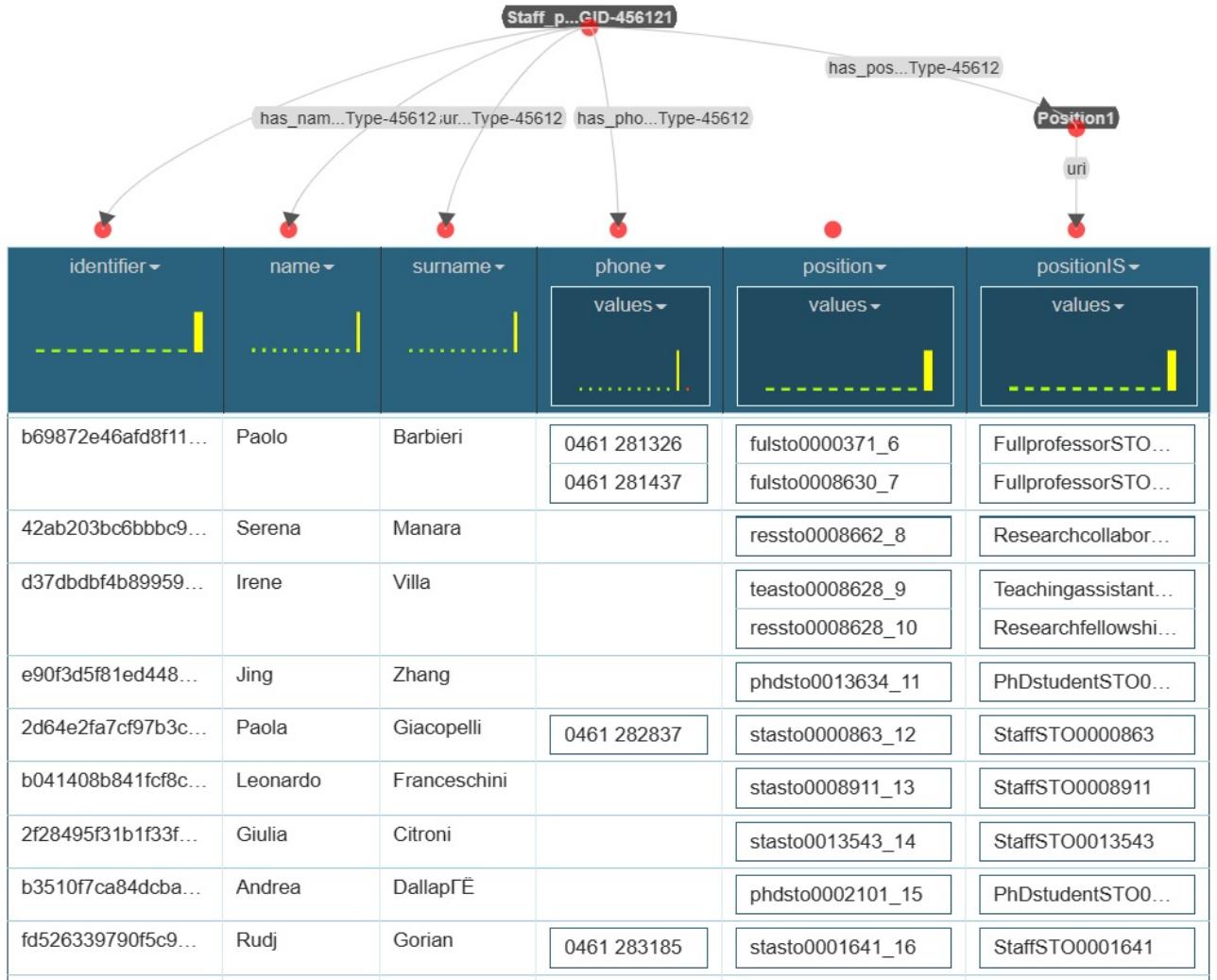


Figure 35: Staff data mapping with Karma

In this case, the mapping is shown for the entity Staff person. It has the *URI* identifier, then the data properties, but also since a staff person has a position, there is an outgoing link to that entity which in itself has a *URI* linked to its unique identifying set identified and described earlier. In this way, since these identifiers are the ones in the Positions dataset, the linkage is done seamlessly, and there are no repetitions in the positions dataset.

Similar procedures were implemented for the other datasets as well, and the output *ttl* files collected and put in the project repository here.

8.4 KG's evaluation

Finally, after the KG has been constructed following all the steps, it is useful to evaluate its quality and its ability to provide use considering the initial purpose. And since all the previous steps were explained in detail, the decisions can be traced and in case there are some issues, it would be easier to identify them and consider what could be changed to work better.

There are 2 main goals of the iTelos methodology:

- 1. How much the final KG is able to satisfy the Competency Queries? (e-types similar to queries; connectivity of the graph)
- 2. How much reusable is the final KG? (comparison with reference ontologies)

There are some metrics to consider. Coverage is the most useful one: measuring how much a portion of knowledge (etypes and properties) is covered by a KG. Connectivity instead is on a data level, evaluating how entities and properties are connected among them.

These issues may be examined once the complete knowledge graph is constructed.

8.5 GraphDB Implementation

A tool which was used to build the Graph database with the identified teleontology mapped with the datasets, is Graph DB, where all the output *ttl* files from Karma were uploaded in a repository which represents the database. The tool allows to explore the graph visually, and in terms of other characteristics.

A visual view of a part of the graph can be seen below (in the GraphDB). The red nodes are the E-types, and arrows represent either data or object properties between instances. In terms of the connectivity metric, it can be explored how entities are connected among each other, and to which E-types they belong. It provides an interactive interface so entities can be exploited/expanded/collapsed.

Visual graph ⓘ

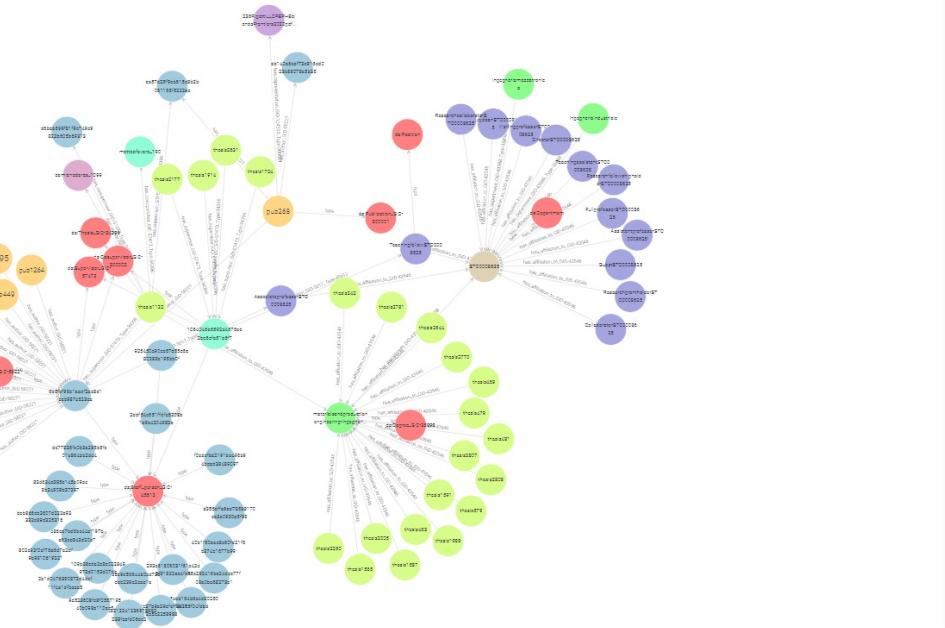


Figure 36: Part of the Graph Visual

In terms of coverage, the most important entities for the purpose are surely present, the ones which are not included are actually stated in the section about Informal modeling. In terms of properties, values that could be useful but are not present, are some contact like emails for persons, as well as probably some more information about degree programs.

Below are the class relationships. In terms of connectivity, Thesis and Publication have the highest number of links outgoing to other classes, as they are connected to other entity types.

Class relationships ?

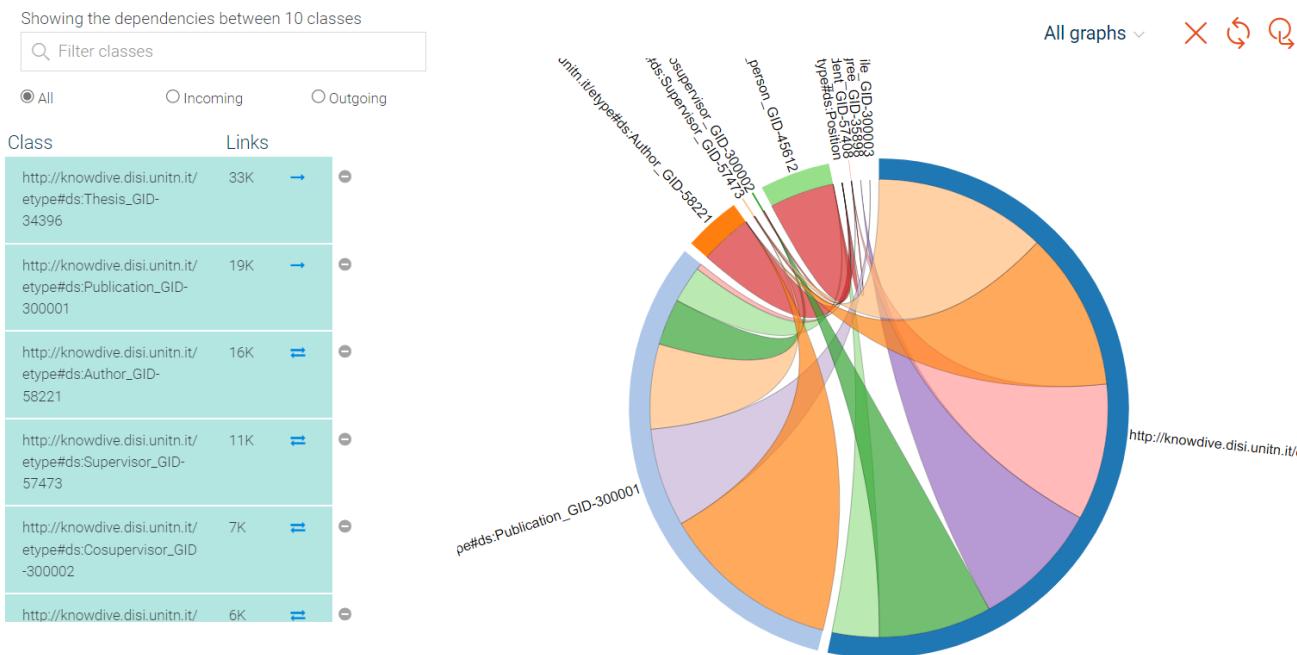


Figure 37: Connections

Then, in the next section, this graph is explored by using some queries simulating some questions, which gives a general idea of what can be answered from the achieved output and that can be used to evaluate the final result.

9 Outcome Exploitation

After completion of all phases of the methodology, the final result has to be exploited in order to see how it performs and whether the result is able to add value for users.

The best way to perform analysis of whether the purpose could be solved here is to translate some kinds of CQs to SPARQL statements and execute the queries in the graph database.

Some examples shown below:

- Courses of study offered at UniTn?

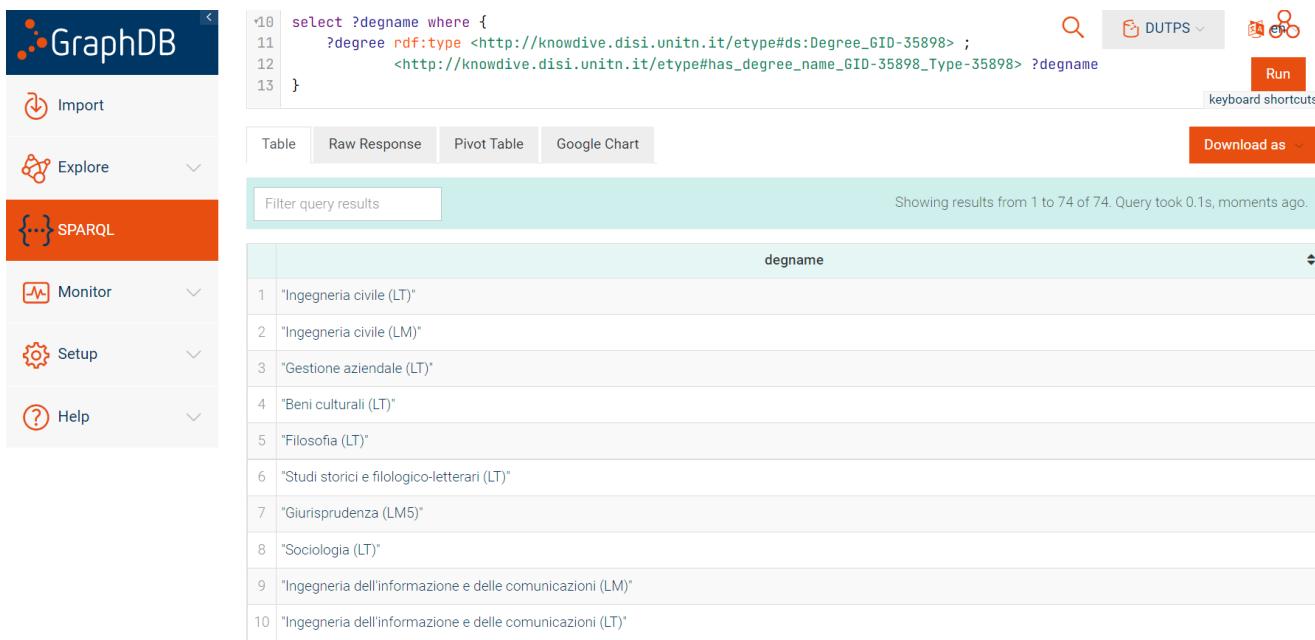


Figure 38: Degree Programs

- Which degree programs are the most attended? Translated to number of theses by degree⁵.

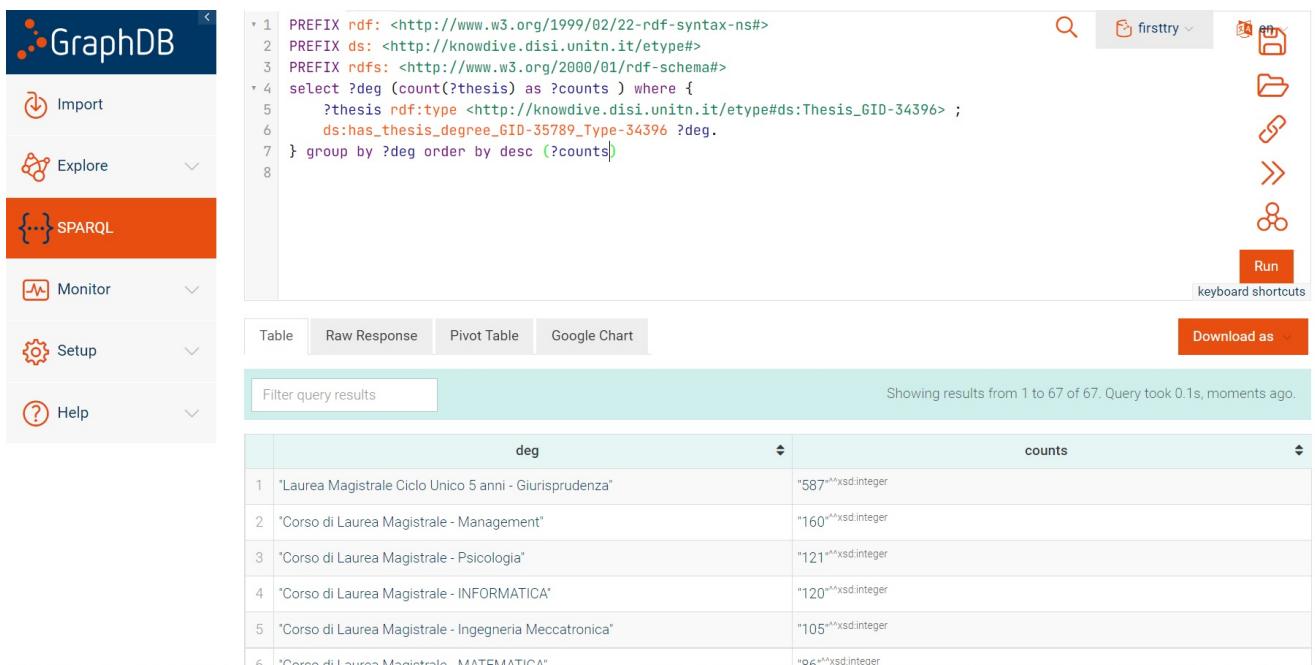


Figure 39: Theses by courses of study

Furthermore, if interested, the theses can be grouped by department instead of degree.

⁵Theses and publications included are only for years 2020 to 2022



- Important figures in different fields. Identified it as how many publications these people have and from which department do they come from. Any contact point of them?

63 #Counts of pubs by authors who are also staff person having positions affiliated to departments
 64 SELECT (count(?title) as ?countpubs) ?name ?surname ?role ?namedep ?phone
 65 WHERE {
 66 ?pub a <http://knowdive.disi.unitn.it/etype#ds:Publication_GID-300001> ;
 67 <http://knowdive.disi.unitn.it/etype#has_pub_title_GID-34061_Type-300001> ?title ;
 68 <http://knowdive.disi.unitn.it/etype#has_representation_GID-104395_Type-300001> ?file;
 69 <http://knowdive.disi.unitn.it/etype#has_author_GID-58221> ?entity .
 70 { ?entity rdf:type <http://knowdive.disi.unitn.it/etype#ds:Author_GID-58221> }
 71 UNION { ?entity rdf:type <http://knowdive.disi.unitn.it/etype#ds:Staff_person_GID-45612> } .
 72
 73 ?entity <http://knowdive.disi.unitn.it/etype#has_position_GID-3017_Type-45612> ?position ;
 74 <http://knowdive.disi.unitn.it/etype#has_name_GID-2_Type-45612> ?name;
 75 <http://knowdive.disi.unitn.it/etype#has_surname_GID-34003_Type-45612> ?surname;

DUTPS       

Table Raw Response Pivot Table Google Chart Download as 

Filter query results  Showing results from 1 to 445 of 445. Query took 0.2s, on 2022-12-13 at 00:55.

	countpubs	name	surname	role	namedep	phone
1	"148"^^xsd:integer	"Roberto"	"luppa"	"Associate professor"	"Dipartimento di Fisica"	"0461 281997"
2	"140"^^xsd:integer	"Ester"	"Ricci"	"Assistant professor"	"Dipartimento di Fisica"	"0461 282031"
3	"50"^^xsd:integer	"Lorenzo"	"Bruzzone"	"Full professor"	"Dipartimento di Ingegneria e Scienza dell'Informazione"	"0461 282056"
4	"30"^^xsd:integer	"Eugenio"	"Aprea"	"Associate professor"	"Dipartimento di Biologia Cellulare, Computazionale e Integrata - CIBIO"	"0461615750"

Figure 40: Most publications by authors

The query is not full on the image. As some publications have more authors, some are in collaborations. This result can be filtered further by for example considering only one department and see the persons in it. For instance, by adding at the end: *FILTER(?namedep = "Dipartimento di Ingegneria e Scienza dell'Informazione")*, we would achieve only the ones from DISI, which would answer one of the competency questions.

- Which are the study programs in Sociology? Where is the faculty located? Study programs with information about departments. Filtered by department

PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
select ?namedeg ?namedep ?director ?address ?website where {
?deg a <http://knowdive.disi.unitn.it/etype#ds:Degree_GID-35898> ;
<http://knowdive.disi.unitn.it/etype#has_degree_name_GID-35898_Type-35898> ?namedeg.
?deg <http://knowdive.disi.unitn.it/etype#has_department_GID-43988_Type-35898> ?dep .
?dep a <http://knowdive.disi.unitn.it/etype#ds:Department> ;
<http://knowdive.disi.unitn.it/etype#has_department_name_GID-43988> ?namedep ;
<http://knowdive.disi.unitn.it/etype#has_department_director_GID-300004> ?director ;
<http://knowdive.disi.unitn.it/etype#has_department_address_GID-45803> ?address ;
<http://knowdive.disi.unitn.it/etype#has_department_website_GID-34126> ?website .
FILTER(?namedep = "Dipartimento di Sociologia e Ricerca Sociale") .
}

Run

keyboard shortcuts

Table Raw Response Pivot Table Google Chart Download as

Filter query results Showing results from 1 to 9 of 9. Query took 0.1s, moments ago.

	namedeg	namedep	director	address	website
1	"Sociologia (LT)"	"Dipartimento di Sociologia e Ri cerca Sociale"	"Giuseppe Sciortino"	"Via Giuseppe Verdi, 26, 38122 T rento"	"http://www.sociologia.unitn.it"
2	"Sociology and social research - sociologia e ricerca sociale (LM)"	"Dipartimento di Sociologia e Ri cerca Sociale"	"Giuseppe Sciortino"	"Via Giuseppe Verdi, 26, 38122 T rento"	"http://www.sociologia.unitn.it"
3	"Gestione delle organizzazioni e del territorio (LM)"	"Dipartimento di Sociologia e Ri cerca Sociale"	"Giuseppe Sciortino"	"Via Giuseppe Verdi, 26, 38122 T rento"	"http://www.sociologia.unitn.it"

Figure 41: Sociology degree programs

Again, can be applied to other departments as well.

- Who are the ones responsible for Biology theses?

namethesis	namedeg	sname	stsurname	supervisorname	supervisorsurname	cosupervisorname
GENOME-WIDE ASSOCIATION STUDY OF SMOKING BEHAVIOUR PHENOTYPES IN INDIVIDUALS OF AFRICAN ANCESTRY	Biologia quantitativa e computazionale (LM)	NOEMI NICOLE	PIGA	Alessandro	Romanel	CHIARA
TUCANA AND KAMER: TWO NEWLY DEVELOPED R-SHINY BIOINFORMATICS ONLINE TOOLS FOR THE INFERENCE OF CO-EXPRESSION NETWORKS AND SURVIVAL ANALYSIS	Biologia quantitativa e computazionale (LM)	LUCA	TRIBOLI	Alessandro	Romanel	FEDERICO MANUEL
A NETWORK-BASED APPROACH TO IDENTIFY DRUG REPURPOSING CANDIDATES FOR ALZHEIMER'S DISEASE	Biologia quantitativa e computazionale (LM)	FEDERICA	MARIOTTI	Enrico	Domenici	SILVIA
DIMENSIONALITY REDUCTION OF THE SINGLE CELL RNA-SEQ DATASET 'TABULA MURIS'	Biologia quantitativa e computazionale (LM)	PAOLO	GRIGIS	Enrico	Blanzieri	Toma
SAMPLING EQUILIBRIUM STATES OF FREE ENERGY SURFACE USING BOLTZMANN GENERATORS	Biologia quantitativa e computazionale (LM)	JOAN	NIGORRA BARCELLO	Gianluca	Lattanzi	Roberto
STRUCTURAL AND MOLECULAR INSIGHTS ON TMC1 PERMEATION PATHWAYS	Biologia quantitativa e computazionale (LM)	DAVIDE	ZAMBONI	Gianluca	Lattanzi	DANIELE
"THE TRANSCRIPTOMICS OF EXTRACELLULAR VESICLES: A PRECISION MEDICINE APPROACH FOR CANCER"	Biologia quantitativa e computazionale (LM)	FEDERICO	VANNUCCINI	Francesca	Demichelis	Yari
DEVELOPMENT AND VALIDATION OF QUANTITATIVE STRUCTURE-ACTIVITY RELATIONSHIP (QSAR) MODELS AS AN INDICATION OF POTENTIAL ENDOCRINE DISRUPTING EFFECTS	Biologia quantitativa e computazionale (LM)	LUCA	D'ALESSANDRO	Graziano	Lolli	ALESSANDRA
LARGE-SCALE METAGENOMIC ANALYSIS OF A PREVIOUSLY UNKNOWN FAMILY OF PREVALENT INTESTINAL BACTERIA ACROSS POPULATION	Biologia quantitativa e computazionale (LM)	GERMANA	BALDI	Nicola	Segata	Francesco

Figure 42: Outcome of query about persons related to theses from Biology department

- Who are the professors at the Psychology department? Is there a way to reach them (any contact information)?

```

83 SELECT DISTINCT ?name ?surname ?phone WHERE{
84   ?staff a <http://knowdive.disi.unitn.it/etype#ds:Staff_person_GID-45612> ;
85     <http://knowdive.disi.unitn.it/etype#has_name_GID-2_Type-45612> ?name ;
86     <http://knowdive.disi.unitn.it/etype#has_surname_GID-34003_Type-45612> ?surname ;
87     <http://knowdive.disi.unitn.it/etype#has_phone_GID-34494_Type-45612> ?phone ;
88     <http://knowdive.disi.unitn.it/etype#has_position_GID-3017_Type-45612> ?position .
89   ?position a <http://knowdive.disi.unitn.it/etype#ds:Position> ;
90     <http://knowdive.disi.unitn.it/etype#has_affiliation_to_GID-43546> ?dep ;
91     <http://knowdive.disi.unitn.it/etype#has_position_role_GID-3735> ?role .
92   ?dep a <http://knowdive.disi.unitn.it/etype#ds:Department> ;
93   <http://knowdive.disi.unitn.it/etype#has_department_name_GID-43988> ?namedep .
94   FILTER(?namedep = "Dipartimento di Psicologia e Scienze Cognitive" &&
95   ?role = "Full professor" )
96 }
```



Run

Press Alt+Enter to keyboard shortcuts

Table	Raw Response	Pivot Table	Google Chart	Get HTML snippet to embed results on a web page												
Chart Config				Showing results from 1 to 13 of 13. Query took 0.1s, moments ago.												
<table border="1"> <thead> <tr> <th>name</th> <th>surname</th> <th>phone</th> </tr> </thead> <tbody> <tr> <td>Massimo</td> <td>Turatto</td> <td>0464 808673</td> </tr> <tr> <td>Paola</td> <td>Venuti</td> <td>0464 808451</td> </tr> <tr> <td>Jeroen Andre Filip</td> <td>Vaes</td> <td>0464 808412</td> </tr> </tbody> </table>					name	surname	phone	Massimo	Turatto	0464 808673	Paola	Venuti	0464 808451	Jeroen Andre Filip	Vaes	0464 808412
name	surname	phone														
Massimo	Turatto	0464 808673														
Paola	Venuti	0464 808451														
Jeroen Andre Filip	Vaes	0464 808412														

Figure 43: Full time professors at Psychology



- Publications in English which have access point

100 WHERE ?
101 ?pub a <http://knowdive.disi.unitn.it/etype#ds:Publication_GID-300001> ;
102 <http://knowdive.disi.unitn.it/etype#has_pub_title_GID-34061_Type-300001> ?title ;
103 <http://knowdive.disi.unitn.it/etype#has_representation_GID-104395_Type-300001> ?file;
104 <http://knowdive.disi.unitn.it/etype#has_author_GID-58221> ?author ;
105 <http://knowdive.disi.unitn.it/etype#has_pub_year_GID-80974_Type-300001> ?year ;
106 <http://knowdive.disi.unitn.it/etype#has_language_GID-33843_Type-300001> ?lang ;
107 <http://knowdive.disi.unitn.it/etype#has_pub_type_GID-31834_Type-300001> ?type.
108 ?file a <http://knowdive.disi.unitn.it/etype#ds:File_GID-300003> ;
109 <http://knowdive.disi.unitn.it/etype#has_file_link_GID-300007_Type-300003> ?filelink .
110 ?author a <http://knowdive.disi.unitn.it/etype#ds:Author_GID-58221> ;
111 <http://knowdive.disi.unitn.it/etype#has_author_name_GID-2_Type-58221> ?name ;
112 <http://knowdive.disi.unitn.it/etype#has_author_surname_GID-34003_Type-58221> ?surname.
113 FILTER(?lang = "inglese" && ?filelink != "NOT_PRESENT")

Run

Table Raw Response Pivot Table Google Chart Download as ↻ 1 2 ↗

Filter query results Showing results from 1 to 1,000 of 1,120. Query took 0.3s, minutes ago.

	title	year	type	filelink
1	"Competition, land prices and city size"	"2020"	"articolo su rivista"	"https://iris.unitn.it/retrieve/handle/11572/344136/556764/Kichko_JoEG.pdf"
2	"The sex factor: How women made the West rich Victoria Bateman Medford, M A: Polity Press, 2019, 248 pp., ISBN-13 978-1-50-95-2676-5, \$15.70; ISBN-13 978-1-50-95-2676-2 (pb), \$24.95"	"2020"	"articolo su rivista"	"https://iris.unitn.it/retrieve/handle/11572/257904/381199/gwao.12446.pdf"
3	"Gender, age, geographical area, food ne	"2020"	"articolo su rivista"	"https://iris.unitn.it/retrieve/handle/11572/344136/556764/Kichko_JoEG.pdf"

Figure 44: All publications with access link in English

- Publications matching a particular search substring

The screenshot shows a SPARQL query editor interface. At the top, there is a code editor with a series of numbered lines (116 to 129) containing SPARQL code. To the right of the code editor is a toolbar with various icons for file operations like save, open, and copy. Below the code editor are tabs for 'Table', 'Raw Response', 'Pivot Table', and 'Google Chart'. A red 'Run' button is located in the top right corner of the editor area. In the bottom right corner of the editor area, there is a 'keyboard shortcuts' link. Below the editor is a table with two rows of data. The table has columns for title, filelink, name, surname, and phone. The first row corresponds to query line 116 and the second to line 129. The table is titled 'Filter query results' and indicates it is showing results from 1 to 2 of 2. The query took 0.4s and was run moments ago.

	title	filelink	name	surname	phone
1	"Intelligenza artificiale e giustizia: potenzialità e rischi"	" https://iris.unitn.it/retrieve/handle/11572/278021/375179/20.7%20AI%20Pisa.pdf "	"Carlo"	"Casonato"	"0461 283876"
2	"Divergenti? Le strategie di Unione europea e Stati Uniti in materia di Intelligenza artificiale"	" https://iris.unitn.it/retrieve/handle/11572/269572/352976/03%20Chiti%20Marchetti%20[29-50].pdf "	"Barbara"	"Marchetti"	"0461 283857"

Figure 45: Publications about artificial intelligence

Those are only some examples of how the graph could be explored, there are numerous queries that could be written and executed. Some of those shown consider some of the competency questions written at the beginning of the document in the Purpose Formalization phase. Some have been written in *txt* files and put in the project repository here.

It seems the results are satisfying keeping in mind what has been the aim of the construction of this Knowledge graph. And apart from the Graph Database, there are all the other resources created during the different stages as artifacts produced from the iTelos methodology which can further be optimized at later stages.

10 Conclusions & Open Issues

In conclusion, the quality of the final outcome for this project has been constructive, including all the main information found and collected for the entities identified, and related through various links as seen in the visual representation.

Overall, the project seems to be able to satisfy the initial purpose.

In terms of scheduling, the project respected the deadlines. There have been some problems using the application KOS in particular in the Formal Modeling phase with the uploading of the Teleontology, but they were resolved by manually recreating the full *owl* file as mentioned before.

Of course, there are some open issues and implementations that are missing or which probably could be developed better. In fact, for certain steps, it was mostly followed a data-driven

approach, a.k.a. leading the solution based on the data and not so much on a more abstract basis, which is probably not the optimal every time.

Some issues that were not satisfied concern the problem of not having sufficient contact information for all persons. Only some of the staff personnel have a phone number, while the other information is not made available at least in these data sources. Since it is a personal information, this could be improved by adding this kind of data from authorized people who have the access and the right to explore it. This issue leaves another implication that as mentioned previously, the person entities do not have many data properties that could be exploited in terms of making the KG richer in information.

Other pieces of information that are missing, are the data about journals and conferences. In the current datasets there is no specific information about where the articles are published and/or presented, therefore such entities are not included in the final KG. However, it may be useful to consider also adding them in both the knowledge and data layer. There is a reference to that in the Informal modeling phase. A possible solution would be to either create a dataset of journals from online/offline libraries, or to find a way to scrape such kind of data from online sources (such as Oxford University Press, Sage Publishing, etc.). There is still another issue that usually publishings in these structures usually are paid and access to the full text is not allowed to everybody, but only to authorized persons. Universities do provide access to some, but not to all articles. So for those reasons, the journal was not added as an entity type, as even if some journals were added, they would have remained unconnected from the graph if there was no way to see which articles to which journals belong. Same is true for the conferences, as at least until now, no database of conferences that have taken place has been found.

Overall, the implementation of the project followed the steps and schedules that were identified, and managed to achieve the result described⁶.

⁶Metadata also for this project report is available in the repository of the project