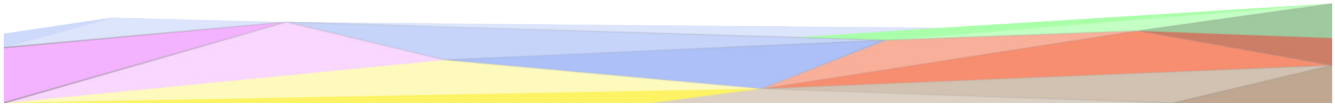
UNIVERSITY
OF TRENTO - Italy

Dipartimento di Ingegneria e Scienza dell'Informazione

– KnowDive Group –

International Digital University

Document Data:	Reference Persons:
February 13, 2025	Riccardo Germenia, Azamat Giniiatullin, Gaia Pizzuti



Index:

Revision History:

Revision	Date	Author	Description of Changes
0.1	October 25, 2024	All	Document created
0.2	October 26, 2024	Gaia Pizzuti	Added introduction
0.3	October 27, 2024	Riccardo Germenia, Gaia Pizzuti	Added personas and scenarios
0.4	October, 28, 2024	Riccardo Germenia, Gaia Pizzuti	Added competency questions
0.5	October 29, 2024	Azamat Giniiatullin	Added ER model
0.6	October 30, 2024	All	Phase 1 last changes
1.0	November 8, 2024	Gaia Pizzuti	Source Identification
1.1	November 10, 2024	All	Dataset Collection
1.2	November 10, 2024	All	Dataset cleaning
1.3	November 11, 2024	Riccardo Germenia	Dataset Standardization
1.4	November 12, 2024	All	Phase 1 revision
2.0	November 21, 2024	Gaia Pizzuti	Concept identification spreadsheet
2.1	November 22, 2024	Riccardo Germenia	Dataset filtering
2.2	November 23, 2024	Gaia Pizzuti, Riccardo Germenia	Phase 2 revision
3.0	November 27, 2024	Gaia Pizzuti	Started Protègè
3.1	November 29, 2024	Gaia Pizzuti, Riccardo Germenia	Protègè OWL file
3.2	December 3, 2024	Riccardo Germenia	Dataset Alignment and Cleanings
3.3	December 3, 2024	Azamat Giniiatullin	Teleology visualization
3.3	December 4, 2024	Gaia Pizzuti	Phase 3 revision
4.0	January 29, 2025	All	Entity matching
4.1	January 30, 2025	Azamat Giniyatullin	Entity identification
4.2	January 31, 2025	Riccardo Germenia	Data mapping
4.3	January 31, 2025	Gaia Pizzuti	Phase 4 revision
4.4	Febrary 1, 2025	Gaia Pizzuti	Initialization evaluation phase
4.5	February 2, 2025	Riccardo Germenia	KG exploitation
4.6	February 3, 2025	Azamat Giniiatullin	Metadata definition

1 Introduction

Nowadays, universities play an important role in promoting international research collaborations. In particular, the University of Trento and the National University of Mongolia have launched numerous collaborative research initiatives, covering various fields such as computer science and data science. However, as these collaborations grow, so does the need for a structured system that efficiently organizes and provides access to information on current and past collaborations.

The goal of this project is to design and implement a Knowledge Graph (KG) that contains collaborative research activities, publications, and related resources between the two universities. The KG will serve as a useful resource for researchers, administrators, and students, allowing them to easily access information about existing research partnerships, discover potential collaborators, and track project progress.
The KG will provide structured, queryable data to support a range of academic and administrative needs. From finding researchers with specific expertise to monitoring the progress of funded projects, the KG will provide users with critical information, ultimately promoting more effective collaboration between the University of Trento and the National University of Mongolia.

To guide the development of this KG, we will apply the iTelos methodology, a structured approach that focuses on identifying, modeling, and organizing knowledge elements to improve usability. The iTelos methodology is well suited for this project, as it provides a clear framework to define the purpose of the KG, specify competency questions, and build a robust ontology suited to the needs of academic institutions and research organizations.
A key aspect of this project is the reusability of the data. The Knowledge Graph will be designed to support data integration with other systems and facilitate long-term reusability. This approach will ensure that information on research collaborations, projects, and publications remains accessible for future applications, such as integration into other institutional KGs.

The report describes:

- **Section 2**: The purpose of the project and the domain of interest.
- **Section 3**: The data sources and resources that were identified and collected for the Knowledge Graph.
- **Section 4**: The terminology and formal language chosen for the Knowledge Graph.
- **Section 5**: The core concepts, relationships, and rules that govern the Knowledge Graph

are defined.

- **Section 6**: A detailed description of each entity within the Knowledge Graph.
- **Section 7**: How will the Knowledge Graph be evaluated to ensure that it meets the defined competency questions and user requirements.
- **Section 8**: The metadata standards and conventions adopted for the Knowledge Graph.
- **Section 9**: Unresolved challenges or questions identified during the development of the Knowledge Graph.

2 Purpose Definition

The Purpose Definition section documents the activities and outcomes of the first phase of the iTelos methodology. This phase is essential to capture the diverse elements of the project, identify key concepts, and establish a structured approach to meet the objectives of the project. The following is a detailed description of each component within this phase.

2.1 Informal Purpose

The primary purpose of this project is to develop a Knowledge Graph that effectively organizes and provides access to information on research collaborations between the University of Trento and the National University of Mongolia.

The Knowledge Graph will serve multiple objectives:

- **Supporting research collaboration**: by collecting detailed information on collaborative research projects, published articles, and researcher profiles from both universities, KG will facilitate easy discovery of potential collaborators and research opportunities.
- **Enabling project and publication tracking**: KG will provide a centralized repository of information on research outputs, including publications, conference proceedings, and project milestones.
- **Improving accessibility of university resources**: information on facilities, campuses, and laboratories at both universities will be incorporated into the KG to help in logistic planning.
- **Promoting data reusability**: a core objective of the Knowledge Graph is to ensure data reusability within and beyond universities.

- **Enhancing decision-making for funding and institutional planning**: By organizing data on funding sources, project outcomes, and research impact, the Knowledge Graph will support administrative decision-making processes.

2.2 Domain of Interest (DoI)

The Domain of Interest for this Knowledge Graph includes:

- **Geographical scope**: focused on the University of Trento in Italy and the National University of Mongolia, covering relevant campuses, facilities, and research centers at each institution.
- **Temporal scope**: the KG will include historical, current and anticipated future research collaborations, spanning over the last decade with a focus on recent and ongoing research from the past five years.
- **Main features**; the KG includes collaborative projects, publications, researcher profiles, facility information, and metadata on funding and research outputs. The DoI supports a wide range of research fields, notably in STEM, with specific focus areas aligning with institutional expertise, such as artificial intelligence and data science.

2.3 Scenarios definition

The following scenarios illustrate key interactions within the Knowledge Graph, demonstrating how it addresses various user needs:

1. Explore possible research collaboration on AI ethics with researchers from the National University of Mongolia.
2. Planning a research visit to the University of Trento for a collaborative project: find information about the facilities and resources available at the University of Trento.
3. Find a candidate with the right experience for a new postdoctoral project aimed at developing predictive models for climate-related crop yield patterns using machine learning.
4. Looking for ongoing collaborative projects in cross-linguistic AI between the two universities.
5. Discover how many publications every research group at the National University of Mongolia produced, the type of publication, and the group members who authored them.
6. Tracking the progress of several collaborative projects between the University of Trento and the National University of Mongolia.

7. Ensure that a new grant proposal complies with the requirements of both universities and includes all the necessary documentation.

A researcher wants to expand her network by including people who work on similar projects in different universities. On the other hand, a senior lecturer is more focused on the faculties of the universities since there is always the possibility of becoming a visiting professor. Moreover, he is interested in finding new PhD students, who after defending their thesis, may pursue a postdoctoral researcher career to help him develop his research ideas.
A Ph.D. student focuses her research on a handful of topics. This means that her interest is to keep up with state-of-the-art publications and datasets. In addition, collaborating on several projects increases the chances of publication, which provides additional accomplishments toward the goal of obtaining a doctoral degree.

2.4 Personas

Each persona is defined based on the diverse contexts and usage needs of the Knowledge Graph.

Name	Age	Description
Dr. Maria Rossi	28	Dr. Maria Rossi is a researcher in Computer Science at the University of Trento. She has a PhD in Artificial Intelligence. She published several papers on Machine Learning and collaborated with international researchers to co-author papers and apply for research grants. Her goal is to discover collaboration opportunities with researchers from the National University of Mongolia, track past and ongoing collaborations between both universities and access relevant papers and project information easily.
Prof. Bat Erdene	41	Prof. Bat Erdene is a senior lecturer in Data Science at the National University of Mongolia. He has over ten years of experience in data analysis and software development. He frequently applies for collaborative projects between Mongolia and international universities so he aims to find partners at the University of Trento for potential projects, access past and ongoing project information between the two universities and learn about joint conferences, workshops and seminars.
Lisa Bianchi	25	Lisa Bianchi is a PhD student at University of Trento working on a thesis related to cross-linguistic processing in AI, interested in collaborating with international institutions. Her goal is to find supervisors or collaborators from both universities, access datasets or ongoing research projects and understand the locations and facilities for research visits.
Dr. Enkhtuul Tserendorj	47	Dr. Enkhtuul Tserendorj is the research administrator at the National University of Mongolia. She holds a PhD in Educational Research, works in the administration department overseeing research projects and collaborations. Her goal is to track and manage research collaborations between the National Univeristy of Mongolia and international institutions like the University of Trento, facilitate the communication between research teams and ensure that collaborative projects properly documented and comply with both universities' guidelines.

While creating the personas, we tried to consider the different roles present in a research institution i.e. university. We started by considering a researcher who wants to collaborate with people from another university. To provide more interesting scenarios, we also added a professor who wants to take part in different projects to search for new ideas and is interested in finding students to propose a research grant. The professor was also needed to create some sort of hierarchy between the personas. These two personas can only provide competency questions related to the projects, so we decided to add a Ph.D. student who searches for both people (supervisors and collaborators) and information (datasets). Lastly, we opted for a manager: this person needs to know the status, both in terms of money and progress, of the projects.

2.5 Competency Questions

The competency questions were crafted based on the scenarios and personas, ensuring that the Knowledge Graph is designed to answer relevant queries for each user role.
4

2.6 ER Model Definition

The final step of the purpose definition involves designing the Entity-Relationship (ER) model. The ER model provides a structural foundation for the Knowledge Graph, capturing the key entities and relationships required to represent research collaborations and the interactions among academic participants at the University of Trento and the National University of Mongolia.
The primary entities within this model include **Research, Researcher, Student, Professor, University, Organization** and **Group**. Entities are categorized into common, core, and contextual entities to provide clarity on their roles within the Knowledge Graph.

2.6.1 Common entities

Common entities are those that are fundamental in various academic contexts and disciplines. This includes:

2.6.1.1 Research It represents individual research projects carried out collaboratively between the two institutions. Each Research entity is characterized by attributes such as title, type, year, language, author, publication type and status. To capture the collaborative nature of research activities, research entities are connected to university through the conducted_in relationship, to researcher through the has_researcher relationship, to student through the has_student relationship and to professor through the has_professor relationship.

Number	Personas	Field	Question
CQ1	Maria Rossi, Enkhtuul Tserendorj	Research Collaboration	Which research projects have been conducted in collaboration between the University of Trento and the National University of Mongolia?
CQ2	Maria Rossi, Bat Erdene	Research Collaboration	Who are the researchers from both universities that have worked together on a specific project or paper?
CQ3	Maria Rossi	Research Collaboration	What are the research areas that have seen the most collaboration between the two universities?
CQ4	Bat Erdene	Research Collaboration	How many collaborative research papers have been published by researchers from both universities over the last five years?
CQ5	Bat Erdene, Enkhtuul Tserendorj	Research Collaboration	Which ongoing projects between the two universities are receiving external funding?
CQ6	Maria Rossi, Lisa Bianchi	Researcher Information	Which researchers from the University of Trento specialize in artificial intelligence and have collaborated with counterparts at the National University of Mongolia?
CQ7	Lisa Bianchi	Researcher Information	Which researchers from the National University of Mongolia have experience in data science and have co-authored papers with researchers from the University of Trento?
CQ8	Lisa Bianchi	Researcher Information	What is the academic profile (publications, projects) of a particular researcher from either university?
CQ9	Bat Erdene, Enkhtuul Tserendorj	Facilities and Locations	What are the main research facilities available at both the University of Trento and the National University of Mongolia for collaborative research?
CQ10	Enkhtuul Tserendorj	Project and Publication Tracking	What is the status of an ongoing research project between the two universities, and who are the main contributors?
CQ11	Enkhtuul Tserendorj	Project and Publication Tracking	What are the deliverables and deadlines for specific ongoing projects between the two universities?
CQ12	Bat Erdene	Student and Educational Collaboration	Which students are pursuing a PhD in Data Science?
CQ13	Lisa Bianchi	Student and Educational Collaboration	Which PhD students from the University of Trento have been involved in collaborative research with the National University of Mongolia?

2.6.1.2 Researcher Represents individuals actively engaged in research, with attributes such as name, surname, role, and languages. Each researcher has a connection with university through the works_in relationship and to group through the participate_in relationship.

scenarios	personas	CQs	entities	properties	focus
1, 2, 4, 5, 6, 7	Dr. Maria Rossi, Prof. Bat Erdene, Lisa Bianchi, Dr. Enkhtuul Tserendorj	1, 2, 3, 4, 5, 8, 10, 11, 13	Research	group_IDs, topic, university, title, link, publication type, status	Common
1, 3, 4, 5, 7	Dr. Maria Rossi, Prof. Bat Erdene, Lisa Bianchi, Dr. Enkhtuul Tserendorj	2, 3, 4, 5, 8, 10, 13	Researcher	name, surname, language, university, organization, role	Common
1, 2, 3, 4, 5, 6, 7	Dr. Maria Rossi, Prof. Bat Erdene, Lisa Bianchi, Dr. Enkhtuul Tserendorj	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13	University	name, organization, buildings	Common
1, 3, 4	Dr. Maria Rossi, Prof. Bat Erdene, Lisa Bianchi	2, 5, 8, 12, 13	Student	name, surname, language, university, organization, degree	Core
1, 4, 5	Dr. Maria Rossi, Lisa Bianchi, Dr. Enkhtuul Tserendorj	2, 4, 6, 7, 8, 12, 13	Professor	name, surname, language, university, organization, courses	Core
1, 2, 3, 4, 5, 7	Dr. Maria Rossi, Prof. Bat Erdene, Lisa Bianchi, Dr. Enkhtuul Tserendorj	6, 7, 9	Organization	university, name, type, address	Core
1, 2, 4, 5, 6, 7	Dr. Maria Rossi, Prof. Bat Erdene, Lisa Bianchi, Dr. Enkhtuul Tserendorj	1, 2, 3, 5, 9, 10, 11	Group	author, roles, publications	Contextual

2.6.2 Core entities

Core entities are the primary components that drive the research collaboration framework. These include:

2.6.2.1 Student It captures the involvement of students in research activities, characterized by attributes such as name, username, language, and degrees. Students are linked to the university through the study_in relationship and to the group through the participate_in relationship.

2.6.2.2 Professor It captures the participation of professors in research activities, characterized by name, surname, language, and courses. Professors are linked to the university through the

study_in relationship and to the group through the participate_in relationship.

2.6.2.3 University It includes both the University of Trento and the National University of Mongolia, is represented by attributes such as name. Universities are also connected to researchers, students, professors, and organizations.

2.6.2.4 Organization It represents distinct academic divisions within the universities involved in the Knowledge Graph, such as the University of Trento and the National University of Mongolia. It is characterized by name, type, address and is linked to the university through the located_in relationship.

2.6.3 Contextual entities

Contextual entities provide additional layers of information that enrich the understanding of the research environment.

2.6.3.1 Group It represents the researchers and their role in the project. Each group is linked to professor, student, and researchers.

3 Information Gathering

The Information Gathering phase focuses on identifying, collecting, and preparing relevant data sources that will serve as inputs for the construction of the Knowledge Graph. In this project, the datasets provided by the Data Scientia Foundation [6, 5] serve as primary sources. This section provides an overview of these data sources, detailing their content, type, origin, and level of diversity awareness. In addition, it outlines the data collection, cleaning, and standardization activities performed to ensure consistency and quality in the Knowledge Graph.

3.1 Sources Identification

The Sources Identification process aims to systematically locate and identify the datasets that align with the Knowledge Graph's objectives and support the competency questions (CQs). Following the iTelos methodology [7], this phase emphasizes the selection of data sources that are relevant and capable of contributing high-quality information to answer the Knowledge Graph CQs. Each source is evaluated based on its content, type, origin, and awareness of diversity, with an emphasis on resources that align with the iTelos methodology.

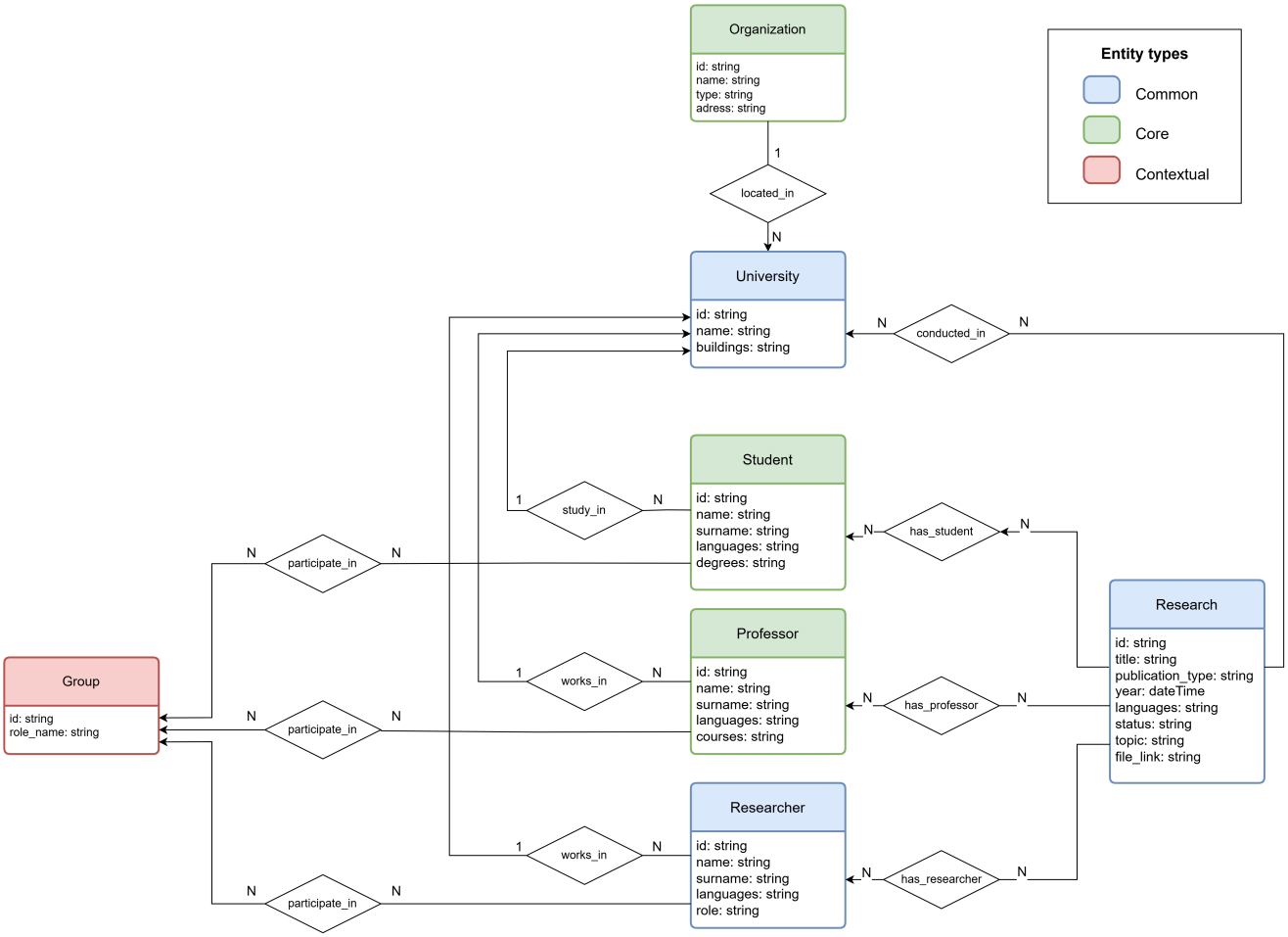


Figure 1: ER model

3.1.1 Dataset selection

The identification of data sources begins with a clear focus on the Knowledge Graph's Purpose. This involves selecting data that accurately represents key entities such as research projects, academic staff, departments, organizational units, and relevant publications. The dataset selection process also prioritizes data sources that capture collaboration between the University of Trento (UNITN) and the National University of Mongolia (NUM), as well as datasets that can improve understanding of these universities' organizational structures.

3.1.2 Dataset evaluation

Each potential data source is evaluated based on its capacity to support the specific CQs that define the Knowledge Graph's intended functionalities. Only datasets that provide meaningful answers to these questions are retained, ensuring all data sources are directly relevant to the project's purpose.

3.2 Dataset collection

The Dataset Collection process involves systematically gathering, organizing, and verifying the datasets identified as relevant during the Sources Identification phase. This phase follows iTelos principles by focusing on data quality, completeness, and compatibility with the Knowledge Graph's requirements.
Data is collected from both primary and supplementary sources, with each dataset carefully organized to retain its relevance and usability for the Knowledge Graph. Sources include:

- **UNITN's LiveData Platform**: This platform provides structured datasets on UNITN's staff, research publications, departments, and collaborations. These datasets are retrieved in formats compatible with the Knowledge Graph, supporting integration with minimal preprocessing [6].
  In particular, LiveDataUNITN offers the following datasets: courses, organizations, research papers and staff. For each dataset we then find the corresponding KG.
- **UniTN's website**: In the "webapps" subdomain of the UniTN's website it is possible to query an api that will provide all the information about the staff including their email.
- **NUM's LiveData Platform**: NUM's datasets, including staff information, organizational units, and research-related details, are accessed and structured to reflect the internal academic and research organization of the university [5].
  In particular, LiveDataNUM offers the following datasets: conferences, courses, curriculum

& programs, journals, organization, research papers, projects, rooms and staff. For each dataset we then find the corresponding KG.

Below we leave a description of the dataset that we decided to use.

3.2.0.1 UNITN Staff Source

This is a cleaned and formatted dataset, created by the University of Trento (UNITN). The Staff dataset provides comprehensive information on staff members at the University of Trento, including their roles, departments, and areas of expertise [6]. This dataset will be used to find data regarding professors and, more generally, researchers at the University of Trento. The dataset is organized as a JSON object where each entry represents a member of the staff. The field included in the data of each member of the staff are id, nome, cognome, telefono, cun, ssd and posizioni that includes ruolo, nomeStruttura and idStruttura.
The most relevant fields are:

- **nome**: the first name of the staff member, stored as a string.
- **cognome**: the last name of the staff member, also stored as a string.
- **telefono**: a list that stores one or more phone numbers for the staff member.
- **ssd**: a code for the scientific-disciplinary sector of the staff member. This field is crucial for linking staff to particular academic disciplines or research fields.
- **posizioni**: a list of positions held by the staff member, represented as a list of objects. Each position includes three main attributes:
  - **ruolo**: the role or position title of the staff member within the university (e.g., ”Collaboratore di ricerca”, ”Docente”).
  - **nomeStruttura**: the name of the department, institute, or structure where the staff member holds this position.

This dataset will also be integrated with the one we created called ”people_emails.json”. This dataset consists of a json file which contains as the key the full name of the person and as the value their email. We decided to add this information because otherwise there would be no easy way to contact one person from working at UniTN, the phone number was missing in most cases.

3.2.0.2 Research paper Source

This is a set of cleaned and formatted datasets, created by the University of Trento (UNITN), that includes information on research papers published by UNITN research staff. A different dataset is provided for each publication year, starting from 2016 to 2023 [6]. This dataset will be used to find data regarding the various research projects carried

out by the University of Trento in the past or currently underway. Each dataset is structured to capture detailed information on academic publications, including their essential metadata, authorship, citations, and any associated files. Each entry contains the following fields: titolo, tipo, anno, lingua, and autori (which is composed by nome, cognome, and id).
The most relevant fields are:

- **titolo**: this field stores the title of the paper, serving as the primary identifier of the publication's content and focus.
- **tipo**: the type of publication, stored as a string, specifies the format or category of the paper (e.g. "articolo su rivista").
- **autori**: this field is a list of objects, each representing an author of the paper. It includes:
  - **nome**: the first name of the author, stored as a string.
  - **cognome**: the last name of the author, also stored as a string.
  - **id**: a unique identifier for each author, provided as a string.

3.2.0.3 UNITN Organization Source

The Organization Dataset is a curated and cleaned dataset developed by the University of Trento (UNITN), containing detailed information about the internal organizational structure of the university [6]. This dataset will be used to find data regarding the various buildings that make up the university in order to be able to answer questions regarding the locations of the University of Trento. Each entry is composed by the following fields: id, tipo, sottoTipo, nome, descrizione, indirizzo, telefono, sitoWeb, pathStruttura.
The most relevant ones are:

- **id**: a unique identifier for each organizational unit, formatted as a string.
- **tipo**: this field captures the main category of the organization, represented as a string.
- **nome**: this string field contains the official name of the organizational unit.
- **indirizzo**: the physical address of the unit, stored as a string, which typically includes street information, city, and postal code.
- **email**: similar to the telefono field, this list holds any email addresses associated with the unit.
- **sitoWeb**: this field is a list of URLs to the unit's website(s) or online presence, enhancing digital accessibility.

3.2.0.4 Journals Source

The Research Dataset is a cleaned and formatted dataset created by the National University of Mongolia (NUM) to catalog information about research journals relevant to NUM's publication standards [5]. This dataset will be used to find data regarding the various research projects carried out by the National University of Mongolia and in particular those that were later published. The fields are structured as follows: journal_id, journal_title, ISSN, publisher, start_date, URL, has_impact_factor and impact_factor.
The most relevant ones are:

- **journal_id**: a unique identifier for each journal entry, stored as an integer.
- **journal_title**: he title of the journal, represented as a string.
- **ISSN**: he International Standard Serial Number of the journal, formatted as a string (e.g., "1234-5678"). This unique identifier allows precise identification of the journal, especially useful in bibliographic and citation contexts.
- **URL**: the web link to the journal's homepage or information page, represented as a string.

3.2.0.5 NUM Organizations source

The Organization Dataset is a cleaned and formatted dataset created by the National University of Mongolia (NUM) [5], providing structured information about NUM's organizational units. This dataset will be used to find data regarding the various buildings that make up the university in order to be able to answer questions regarding the locations of the National University of Mongolia. Each record contains the following fields: Structural unit number, name of structural unit, abbreviation for structural unit, unit number of the structure, the name of the unit to which the structure belongs. The most important ones are:

- **buttsiin_negjiin_dugaar**: (structural unit number) a unique identifier for each organizational unit, stored as a string.
- **buttsiin_negjiin_ner**: (name of structural unit) the full name of the organizational unit in Mongolian. This field captures the official name of each organizational unit, useful for accurate identification within university records.

3.2.0.6 Paper journal Source

The Research Paper Database is a cleaned and structured dataset curated by the National University of Mongolia (NUM) to catalog detailed information on research papers published by NUM researchers [5]. This dataset will be also used to find data regarding the various research projects carried out by the National University of Mongolia and in particular those that were later published. Each record contains the following fields: research_id, research_title, abstract_mn, page_number, publication_date, journal_id, volume_number, issue_number, URL, DOI, research_type_id, research_type, journal_ISSN,

journal_title.
The most important ones are:

- **research_id**: a unique identifier for each research paper, stored as an integer.
- **research_title**: the title of the research paper, provided as a string.
- **publication_date**: the date of publication, formatted as a string.
- **URL**: a URL to the online version of the research paper.
- **research_type**: it denote the specific area of study of the research.
- **journal_ISSN**: the International Standard Serial Number (ISSN) of the journal where the research was published.

3.2.0.7 Paper conference source The Research Conference Paper Dataset is a structured and curated data set created by the National University of Mongolia (NUM) that catalogs research papers presented by NUM researchers at academic conferences [5]. This dataset will be used to find data regarding the various research projects carried out by the National University of Mongolia and in particular those that were later presented during a conference. Each record contains the following fields: research_id, research_title, abstract_mn, conference_id, page_number, participation_date, country, volume_number, research_type_id, research_type.
The most important ones are:

- **research_id**: a unique identifier for each research paper, stored as an integer.
- **research_title**: the title of the research paper, provided as a string.
- **participation_date**: the date of participation, formatted as a string.
- **research_type**: it denote the specific area of study of the research.

3.2.0.8 Paper author source The Paper Author Dataset is a structured dataset created by the National University of Mongolia (NUM) that details the authors of research papers published by NUM researchers [5]. This dataset will be used to find data regarding the authors of each research projects carried out by the National University of Mongolia. Each record contains the following fields: research_id, order, person_id, first_name and last_name.
The most relevant ones are:

- **research_id**: a unique identifier for each research paper, stored as an integer.
- **first_name**: the given name of the author, stored as a string in Mongolian.
- **last_name**: the author's family name, stored as a string in Mongolian.

3.2.0.9 Project overview source

The Project Dataset is a curated and structured dataset created by the National University of Mongolia (NUM) to catalog research projects initiated or participated in by NUM researchers [5]. This dataset will be used to find addictional information about research projects carried out by the National University of Mongolia. Each entry in this dataset provides detailed information about project_id, project_code, project_name, project_name_eng, abstract_mn, abstract_eng, start_date, end_date, sponsor_id, sponsor_name, performer_id, performer_name, funding_type_id, funding_type, research_field_id, research_field, type_id, project_type, project_state_name.
The most important ones are:

- **project_id**: a unique identifier for each research project, represented as an integer.
- **project_name_eng**: the translated name of the project.
- **performer_name**: the name of the department or unit responsible for performing the project, represented as an string.
- **research_field**: a string identifying the research discipline.
- **project_type**: the type of research project.
- **project_state_name**: a string indicating the current status of the project.

3.2.0.10 Project member source

The Project Members Dataset is a structured dataset created by the National University of Mongolia (NUM) that details the personnel involved in various research projects, including their roles, identifiers, and affiliations [5]. This dataset will be used to find data regarding the group members who collaborated in carrying out a research for the National University of Mongolia. Each entry in the dataset provides the following informations: project_id, project_code, position, member_position, researcher_id, first_name, last_name, sisi_id, person_id.
The most relevant ones are:

- **project_id**: a unique identifier for the research project, represented as an integer.
- **member_position**: a string that describes the researcher's role or employment type within the project (e.g. "Core staff").
- **first_name**: the given name of the researcher, stored as a string in Mongolian.
- **last_name**: the researcher's family name, stored as a string in Mongolian.

3.2.0.11 Project keywords source

The Project Keywords Dataset is a structured dataset created by the National University of Mongolia (NUM) that catalogs keywords associated with various research projects [5]. This dataset will be used to find data regarding the keywords of each research projects carried out by the National University of Mongolia. They are usefull to find research relevant to the topic of interest. Each entry in this dataset links a specific keyword to a project using the following fields: project_id, project_code, keyword_id, keyword.
The most relevant ones are:

- **project_id**: a unique identifier for the research project, represented as an integer.
- **keyword**: the keyword associated with the project, stored as a string.

3.2.0.12 Project finance source

The Project Finance Dataset is a structured dataset created by the National University of Mongolia (NUM) that provides financial information on research projects, specifically the total cost associated with each project [5]. This dataset will be used to find data regarding the finantial status of each research projects carried out by the National University of Mongolia. They are usefull to understand the projects' fundings. Each entry contain the following fields: project_id, project_code, total_cost. The most relevant ones are:

- **project_id**: a unique identifier for the research project, represented as an integer.
- **total_cost**: an integer representing the total funding or budget allocated to the project.

3.2.0.13 NUM Staff Source

The Staff Dataset is a structured, cleaned dataset developed by the National University of Mongolia (NUM), containing comprehensive information on NUM staff members [5]. This dataset will be used to find data regarding professors and, more generally, researchers at the National University of Mongolia. The fields are organized as follow: teacher's personal number, teacher's last name, teacher's name, the name of the relevant budget unit, the unit number of the relevant budget, position, acquired education, stood up, email address, office address, extension phone.
The most important ones are:

- **bagsh_ajiltny_khuviin_dugaar**: (teacher's personal number) the unique identifier for each member of the staff, stored as a string.
- **bagsh_ajiltny_ovog**: (teacher's last name) the last name of the staff member in Mongolian.
- **bagsh_ajiltny_ner**: (teacher's first name) the first name of the staff member.
- **khariyaalakh_buttsiin_negjiin_ner**: (the name of the relevant budget unit) the name of the organizational unit or department to which the staff member belongs.

- **alban_tushaal**: (position) the job title or position of the staff member.
- **imeil_khayag**: (email address) the email addresses associated with the staff member.
- **alban_uruunii_khayag**: (office address) the office location of the staff member.

3.2.1 Diversity-Awareness and data quality

Following iTelos principles [7], the data sources are assessed for diversity-awareness and quality standards. Data from well-structured repositories (e.g., UNITN's Live Data platform and NUM's internal databases) is prioritized, as these sources ensure high-quality information with standardized metadata, reducing the likelihood of biased or incomplete records.

3.2.2 Downloading the data

The first step to download the data was to gather the links of all the aforementioned datasets. After that we put them in a JSON file, called "sources.json" which can be found in the github repository [1]. Inside the JSON file the links are organized in two categories: UniTN and NUM. The actual download is performed by the "downloader.py" script. This script creates a folder for each category found in the sources file and for each URL takes the file name (we assume that the file name is the text after the last "/" character), performs a get request and saves the content in a JSON file using the name found before. The script also prints the links used in the form of hypertext using markdown syntax so that they can be easily added to the "sources.md" file which is also found in the repository. The CSV files are manually downloaded and added to both the repository and the markdown sources file. Lastly we used the UniTN API [2] to gather the emails of the staff members. This step was performed using a script to query the API and store the "nominativo" (name and surname) and the "email" fields obtained from the queries in a JSON file.

3.3 Datasets cleaning

The Data Cleaning phase is crucial in ensuring that the data collected align with the goals of the Knowledge Graph and are free of noise - data that are irrelevant or insufficient to satisfy the purpose of the Knowledge Graph. This phase in the iTelos methodology [7] involves systematically refining datasets to improve their accuracy, relevance, and consistency. In our case this step mostly consisted of removing the unnecessary fields and the empty ones. The list of fields that we deemed not useful can be found in Table 1.
The actual cleaning was performed by manually applying some regular expressions using the

“find and replace” function of the IDE, the list of used regexes can be found in the “cleaning_regexes.md” file [3].

Dataset name	Field name	University
mindprod	file.licenza file.formato file.versione	UniTN

Table 1: List of removed fields from the used datasets

3.4 Datasets standardization

The Data Standardization phase is essential for ensuring that data from diverse sources aligns consistently within the Knowledge Graph. Following the iTelos methodology [7], this phase focuses on harmonizing formats, naming conventions, and values across datasets to enable seamless integration, support interoperability, and enhance the quality of query responses.
All the datasets that we considered were either in JSON or CSV format which are considered “high-quality” according to the iTelos methodology. Moreover the data they contained was already well structured so no additional work was needed from us to perform the standardization of the datasets.

3.5 Language resources

A language resource is a file containing the information about dataset names and fields in both the original language and in english. This is especially useful if the data comes from different countries (e.g. Mongolia and Italy). For our language resources we decided to start from the ones provided by the Datascientia Foundation [4, 8]. After checking them we found out that for the UniTN datasets the language resource was up to date on the other hand the one referring to the NUM needed some additional work.
In general, no changes to the file formats were necessary since, as already mentioned, these were already in the same format (i.e. CSV) and internally they were all in JSON format. The main issue we encountered while working on these datasets is that sometimes the keys of the dictionaries are transliterated from Cyrillic to Latin characters. To translate them with a dedicated tool (e.g. Google Translate) it is necessary to first transliterate the word back to Cyrillic characters. This operation was performed by using the CyrTranslit [9] python library, this library helps the user transliterating words from Latin characters to different Cyrillic languages, including Mongolian, and vice-versa. In Table 2, it is possible to find a list of the updates performed on the language resource.
Lastly there are some elements which do not appear in the datasets and do not have a mapping

inside the tables which we decided not to remove because they may become useful in a later phase.

Dataset name	Field name in dataset	Changes
bagsh-ajilchidiin-ners	alban_uruunii_dugaar dotuur_utas	Not in the dataset (kept in the language source) Added to the language source
course	khariyaalakh_tenkhim_dugaar zorilgo_angli hicheeliin_tuluwluguunii_helber	Added to the language source Added to the language source Added to the language source
hicheeliin-huvaari	khicheeliin_dugaar	Added to the language source
research-interests	batalgaajuulsan_suraltsagchiin_too -	Added to the language source Dataset not found (kept in the language source)
journal	-	New name v-journal
paper-authors	-	New name v-authors
conference	-	New name v-conference

Table 2: Differences between the original language resource and the updated one for NUM's datasets

3.6 Purpose definition update

During the initial stages of the project, as the second phase progressed, it became evident that some adjustments were needed in the Purpose Definition Section to ensure smooth and accurate progress.

3.6.1 Scenarios

We found that the scenarios developed in the first phase were overly specific, limiting the model's flexibility to accommodate broader use cases.
We generalized the scenarios, by removing references to personas, to allow for a wider range of queries and applications, ensuring that they better support the purpose of the Knowledge Graph as we move into the second phase.

3.6.2 ER model

We noted that the initial ER model did not fully capture the required relationships and structures within the Knowledge Graph, which required adjustments to better represent key entities and their interactions. In particular, the ER model was more geared toward use for database implementation than knowledge graph implementation. This is because assigned to the various entities, we inserted the ids of the entities connected to them marked as foreign key. Another important aspect was the lack of relationships between the various entities.
As a consequence, we removed foreign keys from entity attributes, while keeping the id of the single entities, lastly we inserted the relationships between entities that are connected. The

relationships can be found in the appropriate rhombuses, and for each relationship we added the cardinality of the relationship (e.g. 1-n, 1-1, etc.).
Finally, we added colors to the various entities so that we could distinguish them between common, core, and contextual.

4 Language Definition

The Language Definition phase represents the third step in the iTelos methodology, focusing on the creation of a formalized domain-specific language to represent the information included in the Knowledge Graph (KG). This phase ensures that the concepts, properties, and entities of the KG are consistently defined, allowing interoperability and semantic precision.
The objective of this phase is to formalize the concepts and resources required for KG. The input for this phase includes the following.

- The purpose formalization sheet developed in the first phase.
- The ER model constructed during the initial stages.
- The resource set, which comprises the data resources collected and processed in the second phase.

The outcome is a set of language resources that include formal concept definitions and a set of filtered resources aligned with these definitions.

4.1 Concept identification

The Concept Identification activity formalizes the concepts required for KG, ensuring that they are uniquely defined and semantically precise.

4.1.1 Concept selection

Concepts are identified from the ER model, purpose formalization sheet, and collected resources. The identified concepts were then entered into the excel file “Concept spreadsheet”.

4.1.1.1 ETypes We inserted all the Etypes (University, Research, Researcher, Organization, Student, Professor and Group). For each Etype we inserted also the attributes and the relationships.
Take for example the student Etype: we have entered all its attributes i.e., name, surname, languages and the degree it is pursuing, and its relationships i.e., the verb study that connects the student entity to the university entity.

4.1.1.2 Specific data values

We have also included other keywords and properties related to those previously entered such as additional entity-related verbs or terms that may come in handy in queries or filtering.
Taking up the previous example of the student entity, we have included the verbs to enroll and to attend so that they can be used later. Wanting to give a further example, we added the term `publication` since it is related to the the `publication type` attribute. Having added the term publication we also added the verb `to publish`.
The purpose of this operation is to add possible values found within the dataset so that the project can be best described. In addition, by adding these values we can avoid possible ambiguities between two or more terms since we are going to specify the meaning we intend to give them in the project. For example, after defining the concept of `Structure` we defined the various types of structure we can find within a university such as `Unit`, `Directorate`, `Rectorship`.

4.1.2 UKC alignment

The Universal Knowledge Core (UKC) is utilized to check whether the selected concepts are already defined: if a concept exists in UKC, its formal definition is adopted. Otherwise, it is formally defined and may later be added to the UKC for reuse in future projects.
In our project we utilized 115 concepts in total. In particular 103 concepts were already defined in the UKC while the other 12 were defined by our group and can be recognized by the ConceptId that will be in the format “KGE24-12-n” where n is a number that is incremented by one for each word that is defined.

4.1.3 Language resource creation

The formal definitions of all concepts are consolidated into a language resource file that can be found in the GitHub directory. This file includes:

- ConceptId: a unique identifier for each concept.
- Word-En: the English word for the concept.
- Gloss-En: the English description (gloss) of its meaning in the domain context.
- Word-It: the Italian word for the concept.
- Gloss-It: the Italina description (gloss) of its meaning in the domain context.
- Word-Mng: the Mongolian word for the concept.
- Gloss-Mng: the Mongolian description (gloss) of its meaning in the domain context.

ConceptID	Word-En	Gloss-En	Word-It	Word-Mng
UKC-110305	University	an institution for higher learning with teaching and research facilities constituting a graduate school and professional schools that award master's degrees and doctorates and an undergraduate division that awards bachelor's degrees	Università	их сургууль
UKC-43416	Organization	the persons (or committees or departments etc.) who make up a body for the purpose of administering something	Organizzazione	байгууллага
UKC-51492	Professor	someone who is a member of the faculty at a college or university	Professore	профессор
UKC-52021	Student	a learner who is enrolled in an educational institution	Studente	оюутан
UKC-52193	Researcher	a scientist who devotes himself to doing research	Ricercatore	судлаач
UKC-31159	Research	a search for knowledge	Ricerca	судалгаа
UKC-59	Group	any number of entities (members) considered as a unit	Gruppo	бүлэг

Table 3: Excerpt from the Concept Spreadsheet file containing concepts for the seven entities in the ER model.

4.2 Dataset filtering

Dataset Filtering aligns the data resources with the formally defined concepts to ensure consistency in the KG. The primary goal of dataset filtering is to align the collected datasets with the purpose-specific vocabulary and schema established in the Language Definition phase. This ensures semantic consistency, improves data quality and the usability of the Knowledge Graph for complex queries and analyses.
Since this process requires the removal of many elements from the datasets which have different types, we decided to use a python script to filter all the unnecessary data. The script (which can be found in the Github repository, inside the "Phase 3" material) recursively searches through the dictionaries and lists of the datasets and removes all the entries that match a blacklist.

4.2.1 Alignment with concepts

Each dataset is systematically reviewed to ensure that its entities, properties, and relationships correspond to the formally defined concepts in the Knowledge Graph. This step is performed manually to ensure that each key is converted to the corresponding concept.
Attributes, entity types, and relationships not covered by the formal schema are either removed

or mapped to existing concepts when appropriate.
In particular we removed:

- `extension phone` in the NUM Staff dataset.
- `page number` in NUM Paper Journal dataset.
- `page number` and `country` in NUM Paper Conference dataset.
- `sottotipo` (sub-type) in UNITN Organization dataset.

Dataset	University	Removed attribute
Staff	NUM	Extension phone
Paper Journal	NUM	Page number
Paper Conference	NUM	Page number
Organization	UNITN	Sub-type

Table 4: Table of the attributes that were removed from the datasets after the Concept identification.

4.2.2 Heterogeneity resolution

This process resolves any inconsistencies between collected datasets by aligning them with the purpose-specific vocabulary.
In particular, we used the concepts defined before to align all the keys in the datasets to the same format.

For example, we translated the Mongolian key ”the number of the unit to which the structure belong” (note that this is the English translation of the key) with the key ”structure_unit_number.

Mongolian key	Translation	New Key
bagsh_ajiltny_khuviin_dugaar	teacher’s personal number	teacher_id
bagsh_ajiltny_ovog	teacher’s last name	teacher_surname
bagsh_ajiltny_ner	teacher’s name	teacher_name
khariyaalakh_buttsiin_negjiin_ner	the name of the structural unit to which it belongs	teacher_structure_name
khariyaalakh_buttsiin_negjiin_dugaar	the structural unit number to which it belongs	teacher_structure_id
alban_tushaal	position	role
ezemshsen_bolovsrol	acquired education	degree
tuluv	status	status
imeil_khayag	email address	email
alban_uruunii_khayag	office address	office_address

Table 5: Excerpt of the translated simplified key from Mongolian to English in the Staff dataset.

4.3 Information Gathering update

During the language definition phase, we identified the need to revisit the results of the information collection phase to improve both the internal coherence of the project and the understand-

Italian Key	Translation	New Key
nome	Name	name
cognome	last name	surname
posizioni	positions	statuses
ruolo	role	role
nomeStruttura	structure's name	structure_name
idStruttura	structure's id	structure_id

Table 6: Excerpt of the translated simplified key from Italian to English in the Staff dataset.

ability of the report. Although the datasets collected during the initial phase provided a solid foundation, some adjustments were necessary to improve the readability of the documentation. The objective of this changes is to improve the reader's understanding of the content and the various phases of the project.

4.3.1 Source Identification

In the previous version of the report the sources considered for Source Identification were listed without explaining the specific datasets used, their relevance, or how they were accessed and downloaded.
The report was updated to specify the sources used, including their origin, purpose and method of access.

4.3.2 Dataset standardization

The Dataset Standardization activity lacked explicit mention of the transformations performed on the collected data.
The last version of the report didn't provide enough clarity on the datasets being of high quality, we were able to download them in the same format (i.e. JSON or CSV format) as they were provided and the data inside them did not require additional work.

4.3.3 Repository Organization

The repository did not follow the provided template structure, making it less intuitive for readers to locate resources and processes within the project's files.
The repository was reorganized to follow the template provided in the project slides. We added the key folders to store the material considered for each phase (i.e. "Phase 1- Purpose Formalization", "Phase 2 - Information Gathering", etc), the Documentation folder to store the project report document and the final presentation of the project, the Evaluation folder containing all the material produced during the evaluation of the project results and the Metadata folder to collect

the definition of all the metadata defined for the different inputs and outputs along all the project phases.

5 Knowledge Definition

The Knowledge Definition phase is the fourth step in the iTelos methodology, where the conceptual and structural representation of the Knowledge Graph is formalized. This phase transforms the output of the earlier phases into a coherent ontology and schema, ensuring alignment with the domain's semantics and enabling seamless integration of datasets.
This phase was carried out in two main steps: Modeling the Teleology and Modeling the Teleontology. These distinct activities ensured that the Knowledge Graph design met the project-specific requirements while also aligning with reusable knowledge standards.

5.1 Teleology modeling

The Purpose Teleology represents a simplified, domain-specific schema designed to directly address the Competency Questions (CQs) defined earlier in the project. This step focused on identifying and modeling only the entities, relationships, and attributes relevant to the project's specific objectives.

5.1.1 Schema formalization

In this part we translate the ER model into a formal ontology written in OWL (Web Ontology Language) using Protègè. We decided to utilize the ontology made available by LiveData UNITN [6] as a starting point and modify it according to the specific needs of our project.

5.1.1.1 Entity Types Each entity type from the ER model was formalized as an OWL class. The following modifications to the original ontology were made:

- The class `Person` was renamed to `Staff`, and three subclasses were introduced: `Researcher`, `Student`, and `Professor`. This reflects our domain, where the primary entity type is a researcher, who can be either a professor or any type of student.
- The class `University` was added, and the preexisting classes `Organization` and `Location` were restructured as its subclasses.
- The class `Creative Work` was renamed to `Research` to introduce a more specific concept. `Research` includes various types, such as dissertations, patents, and publications.

- The class `Group` was added.
- The class `Metadata` and its subclasses (`Duration`, `NL String`, `S String`, and `Moment`) were removed, as they were not relevant to our purpose.
- The class `Computer File` was removed.
- The class `Position` was removed, as this concept is now represented as a property.

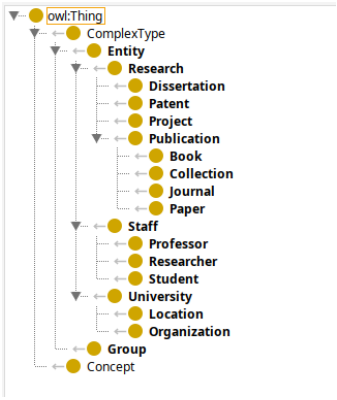


Figure 2: Etypes on Protègè

5.1.1.2 Data Properties For each entity, attributes were defined under the `Data Properties` section. The following modifications were made:

- Several properties were removed, including `ADA id`, `defense date`, `fax number`, and `frequency`, as they were not relevant to our purpose. Additionally, `elevation`, `latitude`, and `longitude` were removed for being overly specific.
- The property `full name` was added to complement existing name-related properties (`first name`, `middle name`, and `last name`).
- All occurrences of `Creative Work` in property names were updated to `Research` to reflect the revised terminology.

5.1.1.3 Object Properties For each entity, relationships were defined under the `Object Properties` section. The following modifications were made:

- Several new relationships were added, including `has description`, `has contact info` (which includes `has phone number` and `has email`), `has full name`, and `has degree`.
- Some relationships, such as `has start` and `has end`, were removed as they were not relevant to our project.

action	property
removed	ADA id, defense date, elevation, fax number, frequency, function, latitude, longitude, maximum, minimum, MIUR category, photo, start, end, subtype of organization, veracity, social security number
added	Full name
renamed	Year of Research instead of Year of Creative Work and Type of Research instead of Type of Creative Work

Table 7: List of data properties that were added and removed in the ontology.

action	property
removed	has start, has end, time
added to File	has link
added to Language	has language
added to has location	has address
added to has research type	Dissertation, Patent, Publication
added to Publication	Book, Collection, Paper

Table 8: List of object properties that were added and removed in the ontology.

5.1.1.4 Hierarchy Parent-child relationships were modeled using class hierarchies. For example, `Publication` was defined as a parent class with subclasses such as `Periodical`, `Paper`, `Book` and `Collection`. [1]

class	subclass
University	Location, Organization
Event	Conference, Course, Project
Staff	**Researcher, Professor, Student**
Research	Dissertation, Patent, Publication
Publication	Book, Collection, Paper, Periodical

Table 9: List of classes with their relative subclasses.

5.1.2 ER model revision

During the process of formalizing the ontology in Protégé, we identified gaps in the initial ER model. These gaps highlighted the need for additional entities to accurately represent the domain and support the datasets and Competency Questions. As a result, the ER model was updated to incorporate these new entities, ensuring a more precise Knowledge Graph.

[1] In Table 9 you can find in bold the classes that we introduced.

Figure 3: Data properties on Protègè

5.2 Teleontology modeling

The Teleontology represents the alignment of the teleology with existing reusable knowledge resources, such as the LiveKnowledge catalog. This ensures that the Knowledge Graph design is not only purpose-specific but also semantically consistent with broader knowledge standards.

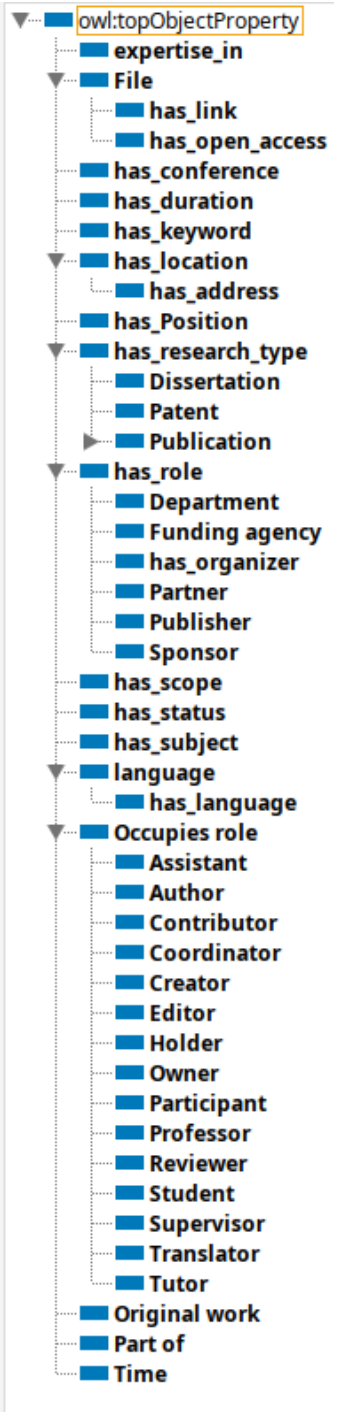


Figure 4: Object properties on Protègè

5.2.1 Dataset Alignment

The Dataset Alignment activity focuses on integrating the collected datasets into the formalized schema, ensuring consistency with the ontology and enabling seamless querying in the Knowl-

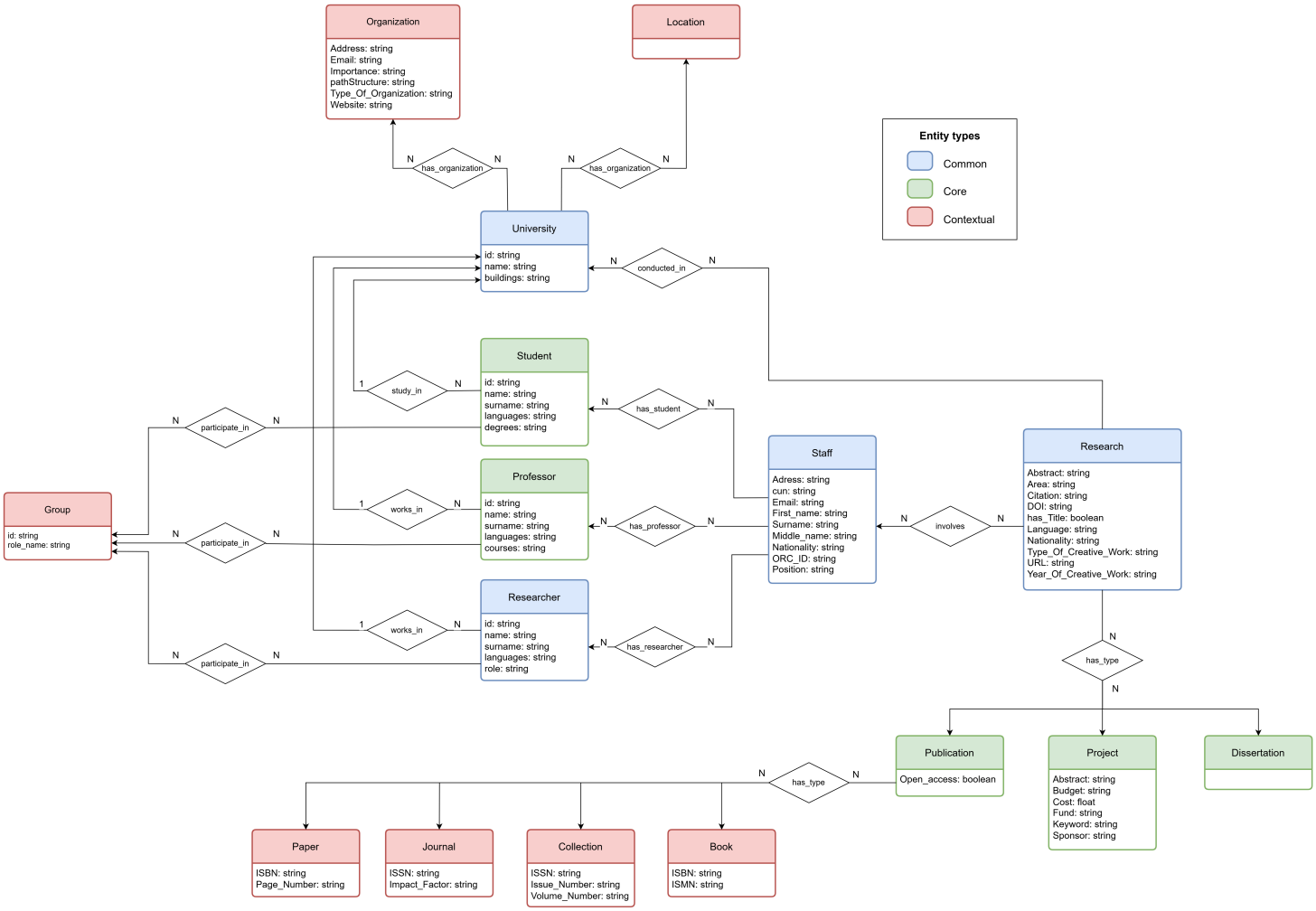


Figure 5: Updated ER model

edge Graph. Even if we aligned the datasets following the same schema, we decided to keep them split into two categories, one for each university (i.e., UniTN and NUM).

5.2.1.1 UniTN In the original “DU-UNITN-people.json” file the emails were missing, so during the second phase we created a new dataset “people_emails.json”. During this phase we decided to merge the two datasets and add the emails in the “DU-UNITN-people.json” one. Moreover we decided to remove the dictionaries surrounding the actual data (it contained information about the language and the number of entries). This allows for easier access to the actual data of the datasets since now the list of entries is the actual content of the dictionary. Lastly we fixed a typo in the “DU-UNITN-organizations.json” dataset and changed “publication_type” to “organization_type”.

5.2.1.2 NUM The changes on the NUM dataset's, organized by dataset, can be found in the following list:

- bagsh-ajilchdiin-ners.json: renamed "teacher_*" instances with "professor_*", renamed "office_address" to address and removed the field "degree";
- baiguullagiin-butets.json: renamed "structure_unit_number" to "structure_path";
- paper-conference.json: renamed "research_type" to "research_area" and moved the "link" (URL) field inside a "file" dictionary;
- paper-journals.json: renamed "research_type" to "research_area";
- project_finance.json: renamed "total_cost" to budget;
- project_member-sisi: removed "member_position";
- v-authors: removed "order".

5.3 Language Definition update

During the transition from the Language Definition phase to the Knowledge Definition phase, we identified several missing concepts that were necessary to accurately represent the domain and fully align the datasets with the Knowledge Graph schema. These gaps became apparent during the formalization of the schema and dataset alignment processes. To address them, we expanded the Concept Spreadsheet by adding 32 new concepts.
The original concept list did not account for several essential entities and attributes required to represent the complexity of the research collaboration domain and support the competency questions. Specifically:

- **Key entities**: concepts such as Location, Event, Dissertation, and Patent were missing, which are crucial for modeling specific aspects of academic and research activities.
- **Roles and Contributor**s: additional roles like Coordinator, Translator, and Supervisor were absent, limiting the ability to capture relationships involving contributors beyond basic authorship.
- **Attributes and Identifiers**: important descriptive attributes, including ISBN, ORCID, and Identifier, were missing, which are critical to uniquely identify publications and researchers.

Finally, we edited some minor errors in the drafting of the text. Specifically in the Paragraph 4.1.1.1 we changed "entity" to "Etype" and in Paragraph 4.1.1.2 we improved the description of the purpose of the operation, namely that of adding data values to the concept spreadsheet that

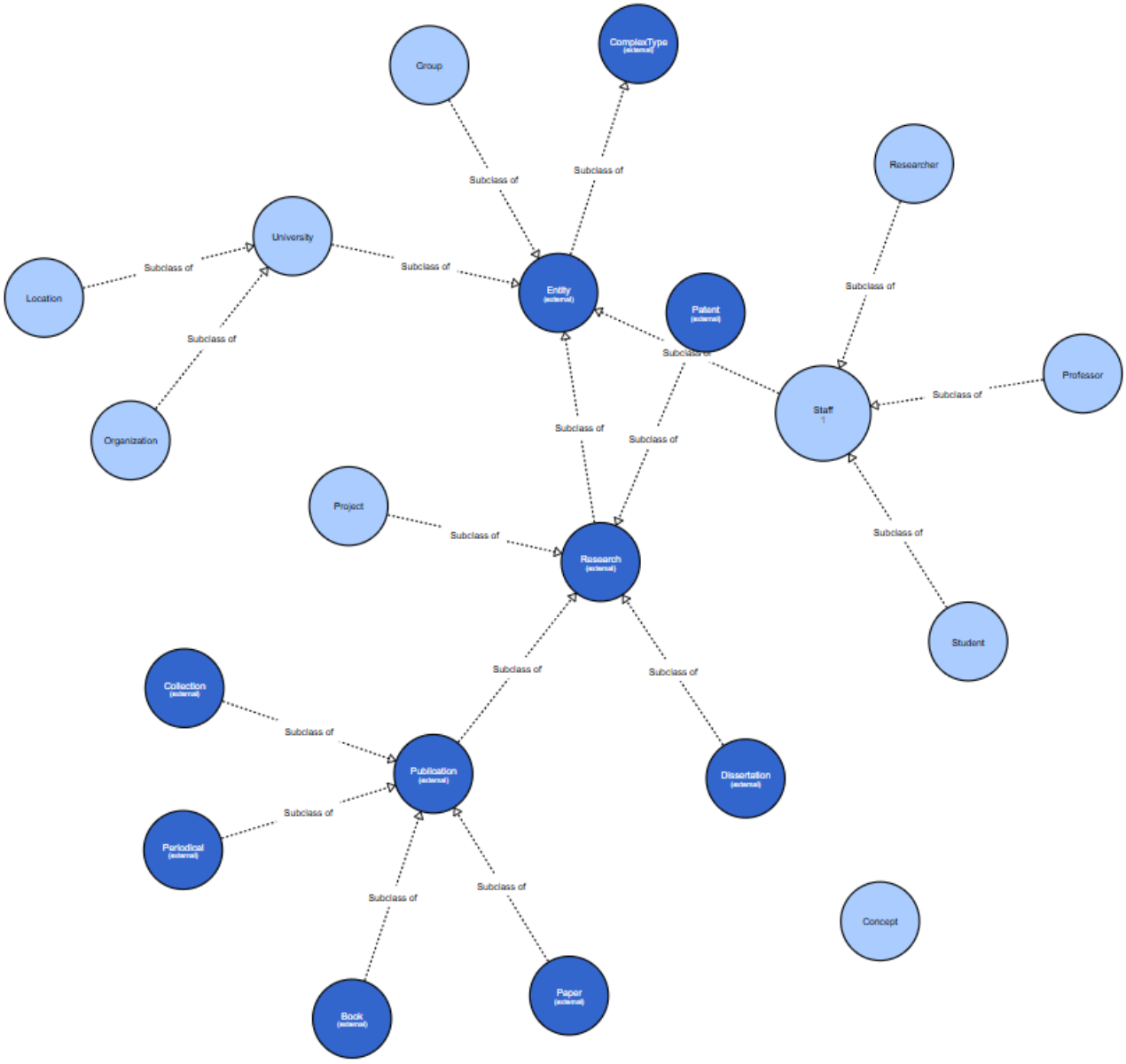


Figure 6: Visualization of the teleology created with WebVOWL

may come in handy for the project so that we can also disambiguate possible values that might otherwise be confused.

6 Entity Definition

In the final phase of the iTelos methodology, the focus shifts to merging the formal knowledge layer (teleontology) with the cleaned, aligned datasets into a unified Knowledge Graph. This

phase primarily ensures that every entity is uniquely identified and accurately mapped into the formal schema, thereby addressing the residual heterogeneity in data values.

6.1 Entity matching

Thanks to our middle-out approach adopted in the previous phases, the challenge of entity matching—that is, reconciling different representations of the same real-world entity across multiple datasets—has already been effectively resolved. By designing our teleontology with careful attention to the properties and relationships present in our datasets, we ensured that entities such as researchers, projects, and publications are consistently defined across all sources. With the entity matching problem largely addressed, our efforts in this phase focus on two critical activities: Entity Identification and Data Mapping.

6.2 Entity identification

Even though entity matching has been solved, it remains essential to formally assign unique identifiers to every entity in the Knowledge Graph. Its necessary to formally identify different entities. This process involves two key aspects:

- dentifying an entity within a single dataset – each entity must be uniquely defined within a specific dataset.
- Using a consistent type of identification – if the same entity is represented differently across multiple datasets, a standardized identification format should be adopted.

Each entity is defined by its properties. In some cases, well-structured datasets already contain a specific property designed to identify the entity it belongs to. This property is called an Identifier There are multiple types of identifiers, depending on how entities need to be recognized:

- URI (Uniform Resource Identifier) – a unique sequence of characters that identifies a logical or physical resource.
- URL (Uniform Resource Locator) – a URI that includes both the method of accessing the resource and its network location.
- URN (Uniform Resource Name) – a URI that identifies a resource by name within a specific namespace.

Identifiers play a crucial role in ensuring consistency and data integrity within knowledge graphs. Their standardized use enables efficient entity matching and prevents data duplication.

Entity	URI	Example	Note
Student	urn:student:id	urn:student:12345	
Professor	urn:professor:id	urn:professor:12345	
Researcher	urn:researcher:id	urn:researcher:12345	
Staff	urn:staff:orc_id	urn:staff:0000-0002-1825-0097	ORCID (Open Researcher and Contributor ID) is an international standard for identifying scientists.
Research Dissertation	urn:research:doi urn:dissertation:doi	urn:research:10.1000/xyz123 urn:dissertation:10.1000/dissertation567	DOI (Digital Object Identifier) is an international standard for identifying scientific publications.
University	urn:university:id	urn:university:NUM	
Publication	urn:publication:id	urn:university:NUM	
Project	urn:project:id	urn:project:H2020-001	
Organization	urn:organization:id	urn:organization:12367	
Paper Collection Book	urn:paper:isbn urn:collection:isbn urn:book:isbn	urn:paper:978-3-16-148410-0 urn:collection:978-3-16-148410-0 urn:book:978-0-12-345678-9	ISBN (International Standard Book Number) A unique book number used in the international cataloging system.
Journal	urn:journal:issn	urn:journal:1234-5678	ISSN (International Standard Serial Number) A unique number for periodicals (magazines, newspapers, scientific collections).

Table 10: Examples of identifiers used for different entities.

6.3 Entity mapping

The final step in this phase is Data Mapping, which involves aligning the data values from our cleaned datasets with the properties and structure defined in our teleontology. This activity is critical for integrating the data layer with our formal knowledge representation.
We perform this activity using Karma.
The example in Figure 7 shows the Staff dataset mapping the Staff entity of the teleontology while the Figure 8 shows the Research dataset mapping the Research entity of the teleontology.

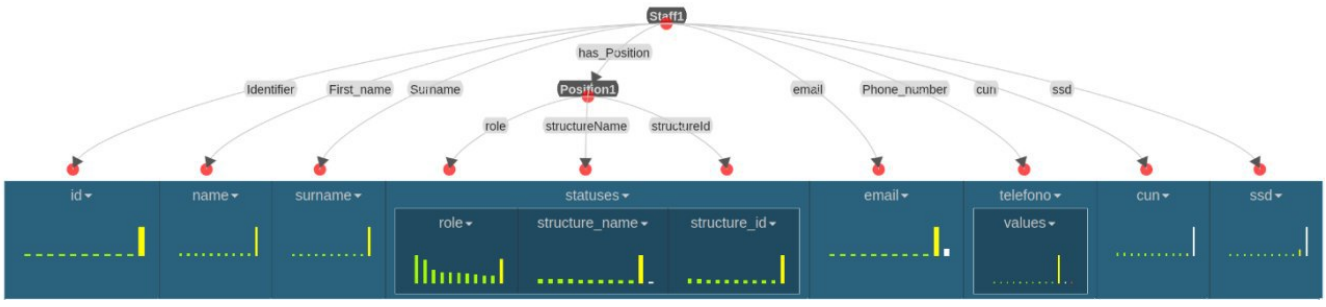


Figure 7: Staff in Karma Integration Tool

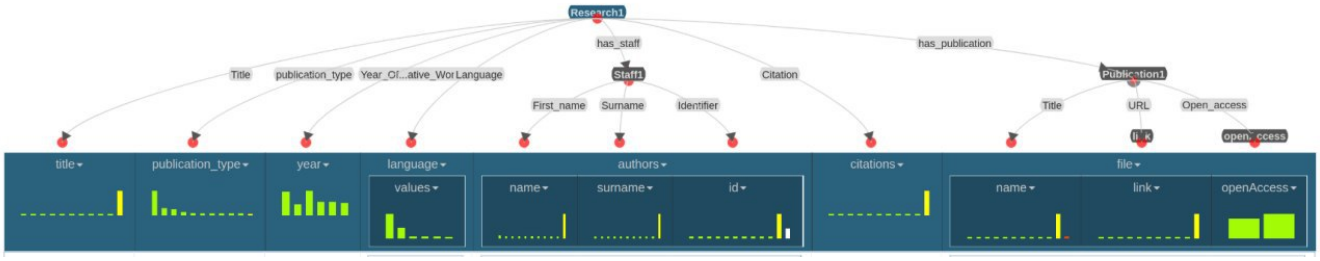


Figure 8: Research in Karma Integration Tool

7 Evaluation

In order to ensure that our final Knowledge Graph (KG) meets its intended purpose and is both reusable and interoperable, we conducted a thorough evaluation following the iTelos methodology. The evaluation was structured around two main objectives:

- purpose satisfaction: assessing how well the KG covers the domain knowledge required to answer our Competency Questions (CQs).
- reusability: determining how much of the KG can be reused by comparing our teleontology with established reference ontologies.

These objectives were evaluated through a combination of coverage and connectivity metrics, which we describe below.

7.1 Evaluation of the Knowledge layer

The knowledge layer of the KG consists of the formalized schema (teleontology) that we developed during the previous phases.
To evaluate this layer, we used coverage metrics to compare the concepts and properties in our teleontology against two targets.

7.1.1 Coverage of Competency Questions

We first extracted the set of etypes required to answer the Competency Questions.
The coverage is calculated as the ratio of the number of etypes and properties present in both teleontology and CQs to the total number of etypes and properties extracted from the CQs.

$$COV_E = \frac{|CQ_E \cap T_E|}{CQ_E} = \frac{7}{7} = 1 \tag{1}$$

We can observe that the number of entities that we can extract from the compentency question is much lower that the number of entities present in the teleontology, i.e. 21 entities, thus reviewing the coverage we observe that it is 100%, indicating complete coverage.

$$COV_p = \frac{|CQ_p \cap T_p|}{CQ_p s} = \frac{32}{32} = 1 \tag{2}$$

The same reasoning can be applied to the properties coverage. In particular, the number of entities extract from the competency questions are 32 while the ones extract from the teleontology are 46.

7.2 Coverage of Reference Ontology

Next, we compared our teleontology with established reference ontologies. Again, we computed the ratio of the intersecting etypes to those in the reference ontologies.

$$COV_E(RO_E) = \frac{|RO_E \cap T_E|}{RO_E} = \frac{15}{26} = 0.58 \tag{3}$$

In particular, as we can see from the equation, the number of entities extract from the reference ontologies are 26 while, as said before, the ones extracted by the teleontology are 46.

$$COV_E(RO_E) = \frac{|RO_E \cap T_E|}{RO_E} = \frac{32}{72} = 0.44 \tag{4}$$

The same reasoning can be applied to the properties coverage.
Together, these evaluations confirmed that our teleontology is both comprehensive for our domain and well-aligned with established knowledge representations.

8 Metadata Definition

Metadata is the structured information that describes, explains, and organizes the resources produced during the Knowledge Graph Engineering process. In our project, metadata plays a pivotal role in ensuring that all language, knowledge, and data resources are easily retrievable, understandable, and reusable by both producers and consumers of the Knowledge Graph.

8.1 Role and purpose

This section of the report compiles definitions for all metadata associated with the various resources generated throughout the entire process. The metadata outlined here pertains to both the final project deliverables and the intermediate outputs from each phase, including language,

schema, and standardized data source values. Defining metadata is essential for facilitating the sharing and dissemination of the produced resources through data catalogs. Therefore, this section also details where the metadata will be published to ensure proper distribution of the described resources. The structure of this section is designed to comprehensively describe the metadata related to all types of resources generated by the project.

- Project metadata description.
- Language resources metadata description.
- Knowledge resources metadata description.
- Data resources metadata description.

8.1.1 Project metadata

This layer captures information about the iTelos project itself. The Table 11 shows the relative fields.

Field	Description
project title	KGE 2024 - International Digital University
project URL	NR
keywords	kge, education, num, unitn, DU
project type	NR
description	Build a Knowledge Graph able to support application and services providing information about research collaboration (published paper, project or others) between the university of Trento and the national university of Mongolia, describing also the different location of the two university facilities. Development of this KG using iTelos methodology, a structured approach that focuses on identifying, modeling, and organizing knowledge elements to improve usability.
start date	20-10-2024
end date	14-02-2025
funding agency	University of Trento
input data sources	UNITN's LiveData Platform - UNITN's website - NUM's LiveData Platform
outputs	Language, knowledge definitions- elaborated data in both JSON format- data mapping models and RDF formatted data
coordinator	NR
observations	Although the datasets collected during the initial phase provided a solid foundation, some adjustments were necessary to improve the readability of the documentation.

Table 11: Project Information: KGE 2024 - International Digital University

8.1.2 Language metadata

This layer captures information about language resources. The Table 12 shows the relative fields.

Field	Description
data license	CC-BY-SA-4.0
data URL	Repository
keywords	Concepts, UKC
publisher	Riccardo Germenia, Azamat Giniyatullin and Gaia Pizzuti
creator	Riccardo Germenia, Azamat Giniyatullin and Gaia Pizzuti
owner	Riccardo Germenia, Azamat Giniyatullin and Gaia Pizzuti
language	English, Italian, Mongolian
level	N/A
size	INSERT SIZE OF XLSX
name	Concepts definition for KGE 2024 project – International Digital University
publication timestamp	INSERT TIMESTAMP
description	Concept definition for data resources using UKC, including UNITN's and NUM's LiveData Platforms.
version	Version 1.0
domain	N/A
file format	xlsx

Table 12: Dataset Information: KGE 2024 – Language Resource

8.1.3 Knowledge metadata

The following layer presents the knowledge metadata for our project. The Table 13 shows the relative fields for the Teleontology files while the Table 14 shows the relative fields for the Teleology files.

8.1.4 Data metadata

The following tables presents the metadata for the RDF models and datasets used in the KGE 2024 project. This metadata serves as a key reference for understanding the dataset's characteristics, licensing, and ownership. In particular Table 15 shows the relative fields for the RDF models and Table 16 shows the relative fields for the datasets.

9 Repository

To facilitate access to the project's source code and documentation, we created a public repository on GitHub. The repository contains all the code, resources, and scripts used to build and integrate the Knowledge Graph.

Field	Description
data license	CC-BY-SA-4.0
data URL	Teleontology
keywords	teleontology
publisher	Riccardo Germenia, Azamat Giniyatullin and Gaia Pizzuti
creator	Riccardo Germenia, Azamat Giniyatullin and Gaia Pizzuti
owner	Riccardo Germenia, Azamat Giniyatullin and Gaia Pizzuti
language	English
level	N/A
size	37.2 KB
name	teleontology
publication timestamp	INSERT TIMESTAMP
description	Contains the entities aligned with other publicly available knowledge resources.
version	version 1.0
domain	N/A
file format	OWL

Table 13: Dataset Information: KGE 2024 – Teleontology Files

Field	Description
data license	CC-BY-SA-4.0
data URL	Teleontology
keywords	teleology
publisher	Riccardo Germenia, Azamat Giniyatullin and Gaia Pizzuti
creator	Riccardo Germenia, Azamat Giniyatullin and Gaia Pizzuti
owner	Riccardo Germenia, Azamat Giniyatullin and Gaia Pizzuti
size	144.7 MB
name	teleontology
publication timestamp	INSERT TIMESTAMP
description	The Teleology file contains entity types, data properties, and object properties used in this project.
file format	ttl

Table 14: Dataset Information: KGE 2024 – Teleology Files

Link to the repository: International Digital University Repository

Field	Description
data license	CC-BY-SA-4.0
data URL	RDF model
keywords	RDF model, KG
publisher	Riccardo Germenia, Azamat Giniyatullin and Gaia Pizzuti
creator	Riccardo Germenia, Azamat Giniyatullin and Gaia Pizzuti
owner	Riccardo Germenia, Azamat Giniyatullin and Gaia Pizzuti
language	English
level	N/A
size	589 KB
name	Data mapping for KGE 2024 project – International Digital University
publication timestamp	INSERT TIMESTAMP
description	Karma models to map JSON computed datasets into RDF's.
version	version 1.0
domain	N/A
file format	ttl

Table 15: Dataset Information: KGE 2024 – RDF model

Field	Description
data license	CC-BY-SA-4.0
data URL	all_data.brf
keywords	RDF model, KG
publisher	Riccardo Germenia, Azamat Giniyatullin and Gaia Pizzuti
creator	Riccardo Germenia, Azamat Giniyatullin and Gaia Pizzuti
owner	Riccardo Germenia, Azamat Giniyatullin and Gaia Pizzuti
size	39.9 MB
name	Knowledge graph data for KGE 2024 project – International Digital University
publication timestamp	INSERT TIMESTAMP
description	Published RDF data for KGE 2024 project – International Digital University.
file format	brf

Table 16: Dataset Information: KGE 2024 – Data

References

[1] https://github.com/GaiaPizzuti/KGE-project/blob/cec75e3b3aa7131c0c526d3c1534a88f13ed8f3a/Phase%202%20-%20Information%20Gathering/Code%20Libraries/sources.json.

[2] https://webapps.unitn.it/api/du/v1/persona/search?text=&filter=&items=100.

[3] https://github.com/GaiaPizzuti/KGE-project/blob/cec75e3b3aa7131c0c526d3c1534a88f13ed8f3a/Phase%202%20-%20Information%20Gathering/Code%20Libraries/cleaning_regexes.md.

[4] Datascientia Foundation. *DU-UNITN Language Resources*. https://github.com/datascientiafoundation/LiveDataUNITN-DREP/blob/main/Language%20Resources/DU-UNITN%20Language.csv.

[5] Datascientia Foundation. *Live Data NUM*. https://datascientiafoundation.github.io/LiveDataNUM/. Accessed: 2024-11-13.

[6] Datascientia Foundation. *Live Data UNITN*. https://datascientiafoundation.github.io/LiveDataUNITN/. Accessed: 2024-11-13.

[7] F. Giunchiglia and S. Bocca. *The iTelos Methodology - Information Gathering*. 2024.

[8] KDE Milab. *DU-NUM Language Resources*. https://github.com/kde-milab/LiveDataDREP/blob/main/Language%20Resources/DU-NUM%20Language.csv. Accessed: 2024-11-13.

[9] Georges Labrèche. *CyrTranslit*. Version v1.1.1. A Python package for bi-directional transliteration of Cyrillic script to Latin script and vice versa. Supports transliteration for Bulgarian, Montenegrin, Macedonian, Mongolian, Russian, Serbian, Tajik, and Ukrainian. Mar. 2023. doi: 10.5281/zenodo.7734906. url: https://doi.org/10.5281/zenodo.7734906.