UNIVERSITY
OF TRENTO - Italy

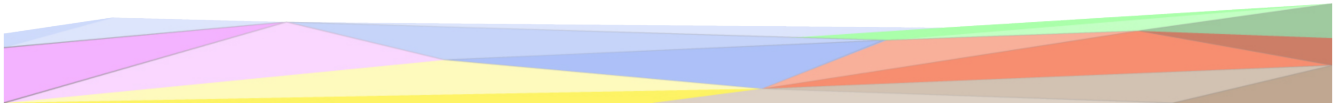
Dipartimento di Ingegneria e Scienza dell'Informazione

– KnowDive Group –

Knowdive Research Products

Phase 1+2

Document Data:	Reference Persons:
November 13, 2024	Aloisi Deborah, Giaretta Leonardo



Index:

Revision History:

Revision	Date	Author	Description of Changes
0.1	16.10.2024	Both	Document created
1.0	21.10.2024	Deborah Aloisi	Added Introduction
2.0	23.10.2024	Leonardo Giaretta	Added Domain of Interest
2.1	25.10.2024	Deborah Aloisi	Added context
2.2	26.10.2024	Both	Added Scenarios and Personas
2.3	28.10.2024	Both	Added Competency Question
2.4	28.10.2024	Both	Defined entities
2.5	29.10.2024	Leonardo Giaretta	Associate entities with CQ
2.6	30.10.2024	Both	Added ER
3.0	7.11.2024	Deborah Aloisi	Added part of ontologies and language datasets
3.1	8.11.2024	Deborah Aloisi	Added description and collected LiveDataUNITN dataset
3.2	9.11.2024	Leonardo Giaretta	Added Openmap data
3.3	10.11.2024	Leonardo Giaretta	Added Site scrapped data
3.4	11.11.2024	Both	Added other ontologies
3.5	12.11.2024	Deborah Aloisi	Added Site scrapped data
3.6	12.11.2024	Both	Added language dataset
3.7	12.11.2024	Both	Data standardization
3.8	12.11.2024	Both	KRG-Language dataset completed
3.9	13.11.2024	Both	Done data cleaning

1 Introduction

Data has become an essential part of our lives, but without proper organization and management, it loses its value.

This project has the objective of organizing data about members, alumni, papers and other products of the Knowdive research group in a Knowledge Graph that makes it simpler to query and find information about it.

This will be accomplished using the iTelos methodology, a structured approach designed to streamline the Knowledge Graph Engineering (KGE) process and minimize development effort. The methodology is intended to assist users in addressing the challenges that arise when building purpose-specific knowledge graphs (KGs). It also facilitates the creation of reusable resources, promoting circular data reuse and thus reducing the effort required to generate new resources.

In the iTelos methodology, re-usability is a central principle in KGE, with project documentation playing a critical role in enhancing the re-usability of the resources developed throughout the project. This report aims to document both the process and methodology used, explaining the rationale behind key decisions and how resources were leveraged to achieve the project's objectives. By doing so, the documentation will serve as a valuable resource for external readers, potentially enabling them to reuse the project's outputs for other purposes.

The report is structured as follows:

- Section 2: Definition of the project's purpose and its domain of interest.
- Section 3: Definition of the Data Source

2 Purpose Definition

2.1 Informal Purpose

The final Knowledge Graph (KG) can be utilized as a general-purpose service to assist users in discovering research products (such as papers, reports, and other published outcomes) produced by the Knowdive research group at the University of Trento (DISI). This functionality can be articulated as a user request in the following form:

”A service which helps the users to query and know about all the products and the people that are working or were working in the Knowdive research group at the University of Trento”

2.2 Domain of Interest (DoI)

- Personal context : We expressed our Personal context Purpose from the point of view of a member of the Knowdive research group, as they are the category of users which have a strictly personal interest in this project. ”I want a KG which collects the results of my research and available data from the projects I’m taking part into”
- Reference context: We expressed our Reference Context purpose in a dual way, in order to emphatise two different aspects of this project: the geographical aspect and the research aspect. ”I want a KG which represents the distribution across the globe of Knowdive’s members and projects” ”I want a KG which represents the research products of the Knowdive research products”
- Personal-Reference context Through the Personal-Reference context purpose we were able to express the point of view of a non-member user. ”I want a KG which allows me to access the data regarding the Knowdive group members, their research products and their projects by filtering through various criteria.”
- Reference-Personal context Through the Reference-Personal context purpose we expressed the broader point of view that takes into account both member and non-member users. ”I want a KG which integrates the data about members, research product and projects of the Knowdive Research Group, allowing users to consult data on various scientific topics and members of the group to organize their work results.”

In term of context we define:

- Geographical boundaries: We express the spatial scope with the coordinate of the University of Trento, more precisely with the Department of Information and Computer Science

(DISI) where the Knowdive research group is officially located. The coordinate are:

Department	Latitude	Longitude	Altitude (m)
Dipartimento di Ingegneria e Scienza dell'Informazione	46.0668994	11.1481093	370

- Temporal boundaries: The scope is from the starting of the Knowdive research group in the 2006 to the current year, since a potential user could be both interested in the archived data and current and updated information.
- Domain boundaries: The project concerns the Knowdive Research Group and in particular its member, alumni, papers and other works produced.

2.3 Purpose Formalization

2.3.1 Scenarios definition

1. A Master's student is searching inspiration for the topic of their thesis. They want to see the publications of the Knowdive group so that they can write to a professor that works on their interested topic.
2. A company wants to employ a person with reference to the Knowdive research group, so they want to check if the credentials are correct and if he/she aligns with the position they are offering.
3. A University wants to invite member of the Knowdive research group for a seminar and are searching which member would be the best to contact.
4. A student that will move to another university wants to know if in his/her new university there is a member of the Knowdive group and if the research interests of such a member align with his/hers.
5. An Institution (such as a ranking Institution) wants to measure the growth of the Knowdive group through its research output and/or the current amount of projects its undertaking.
6. A Researcher of the Knowdive group wants to have a neatly organized aggregation of the research products made by him/her and his/hers colleagues.

2.3.2 Personas

Name	Age	Profession	Description
Carlotta	24	Student	Carlotta is approaching the end of her Master's degree, but without a thesis she can't graduate. She doesn't have a precise idea of what she want to study so she is searching inspiration on between other thesis. In doing so she ended on the Knowdive research group page and wants to see what they are studying
Gianpaolo	47	HR personnel	Gianpaolo is one of the HR personnel at company X in Trentino. His company wants to develop a project correlated to a Knowledge graph but they don't have any employee that can take care of it. So his company instructed him to search for a possible candidate to hire for the project and the first idea that has come to his mind is to search for figures connected to the local University.
Mabel	35	Secretary of University	Mabel was tasked by a professor at her university to contact the Knowdive Research group to invite a member of the group to host a seminar. She wants to find who is the best person to contact for the task
Franco	23	Student	Franco is a Master's student with a personal interest in the Human Machine Symbiosis topic. He was planning to try and propose a Project course on this topic, but he doesn't know which professor may be interested. By talking with other students he discovers the existence of the Knowdive group and sets out to find a member that matches his interest.
Anna	22	Student	Anna will be soon finishing her Bachelor's degree and will be moving to another university. Having just developed a passion for Artificial Human Cognition and knowing that the Knowdive group does research on that topic, she searches for connections between the group and her future University.
Valerio	41	Institution X employee	As a part of a future analysis on the research environment in Trentino, Valerio has been tasked by the Institution X to collect data regarding the growth of the various research groups and organization of the Trentino territory. He is currently collecting data on Knowdive and needs the most precise data possible about the amount of members of the group through time, their projects and the amount of research published.
Arianna	37	Knowdive member	Arianna always liked tidy and clear aggregation of data so she want to see all the products done by her research group divided by colleague, categories and so on.

Table 1: Personas

2.3.3 Competency Question (CQs)

Persona	ID	Question
Carlotta	1.1	I'd like to know all the thesis written in collaboration with the Knowdive Research group

Carlotta	1.2	I'd want to see all Master thesis written in collaboration with the Knowdive group
Carlotta	1.3	I'm curious if the people that wrote their thesis in collaboration with the Knowdive group have any postdoctorate work.
Carlotta	1.4	Give me the topics a member of the KRG is studying
Carlotta	1.5	I've finally decided on a topic for my thesis, but I still don't know which member of the KRG to contact. I would need to have a list of all members that work on my chosen topic with the possibility to filter through role and location.
Gianpaolo	2.1	Give me all the current member of the KRG
Gianpaolo	2.2	Give me all the ex-member and their position when they left
Gianpaolo	2.3	Give me only the members that are researchers or professors
Gianpaolo	2.4	Let me search if a specific person has been part of the KRG and get their position when they left and a list of their research products and of the projects they participated in.
Mabel	3.1	Give me a list of the research topics of the KRG.
Mabel	3.2	I want to be able to find all the members of the KRG that are either a professor or a researcher that works or has worked on a given topic.
Mabel	3.3	I want to know who is the coordinator for the seminar activities
Mabel	3.4	I need to get the contact information for a specific member of the KRG.
Mabel	3.5	The person I thought of is unavailable, but I could contact one of his collaborators. I need to find all the members that have worked with him/her on the chosen topic.
Franco	4.1	I want to see all research products produced by the KRG on a specific topic.
Franco	4.2	I want to see each KRG member that has worked on a research product.
Franco	4.3	I know a professor that is part of the KRG, I want to see on which topics he/she worked on during the last years.
Franco	4.4	Are there any members that have worked with the professor I know that have also worked on the topic I prefer?
Anna	5.1	I'd like to know which KRG members are located in my future university.
Anna	5.2	I'd like a list of all non-members of the KRG that have collaborated with the group on a given topic.
Anna	5.3	I'd like a list of all ex-members of the KRG that have worked on a given topic.
Anna	5.4	I would need the location of an ex-member of KRG before he left the group.
Anna	5.5	I would like to know which courses the members of the KRG teach
Anna	5.6	I would like to know who teach the Studies in Human Behavior corse
Valerio	6.1	I need to track how many members the KRG had through the years and who these members are.
Valerio	6.2	I need to track the amount of research products produced by the KRG through the years and fetch the appropriate identifying data about each of those products.

Valerio	6.3	I need to be able to track in which years a current or past member of the KRG has been part of the group, how many and which research products he/she produced during that time.
Valerio	6.4	I need to track the amount of research products on a given topic that has been produced on a given topic and fetch the appropriate identifying data about each of those products.
Valerio	6.5	I need to track the status and duration of each project in which the KRG has partecipated.
Valerio	6.6	For each project, i need to extract all the Universities and Institutions that have partnered with the KRG.
Arianna	7.1	I'd like to have a list of all the research products of the KRG, ordered by date.
Arianna	7.2	I'd like to know, for each project in which the KRG has taken a part in, which members have worked on it.
Arianna	7.3	For each member of the KRG, I would like to have a list of his/her research products.
Arianna	7.4	I want to know in which years a member of the KRG was active the most
Arianna	7.5	I want to know which topic is more researched based on related works created

Table 2: CQs

CQ ID	Common entities	Core Entities	Contextual entities
1.1		Research Product	
1.2		Research Product	
1.3	Person	Research Product	
1.4	Person	Member, Research Topic	
1.5	University	Member, Research Topic	
2.1		Member	
2.2	Alumni		
2.3	Member		
2.4	Member, Research Product, Project		
3.1		Research Topic	
3.2	Member, Research Topic		
3.3	Member		
3.4	Member		
3.5		Member, Research Topic	
4.1		Research Product, Research Topic	
4.2		Member, Research Product	
4.3		Member, Research Topic	
4.4		Member, Research Topic	
5.1	University	Member	
5.2	Person	Research Topic, Research Product	
5.3		Alumni, Research Topic	

5.4	University	Alumni	
5.5	University	Member	Course
5.6	University	Member	Course
6.1		Member, Alumni	
6.2		Research Product	
6.3		Member, Alumni, Research Product	
6.4		Research Product, Research Topic	
6.5		Project	
6.6	University	Project	
7.1		Research Product	
7.2		Project, Member	
7.3		Member, Research Product	
7.4		Member, Project, Research Product	Course
7.5		Research Topic, Research Product	

Table 3: Entities related to CQs

2.3.4 Concepts identification

In a knowledge graph, entities represent real-world objects and are instances of Entity Types (ETypes), which define the types of real-world objects. This section categorizes ETypes based on their popularity categories: Common, Core, and Contextual.

Common entities

Common entities consist of resources that contain information relevant across multiple contexts or domains of interest. While not directly tied to the user's Purpose, they are essential for supporting it within the knowledge graph.

For this project, the identified common entities are:

- University
- Person

Core entities

Core entities are those that contain essential information about the key aspects central to the Purpose, forming the foundational data necessary to construct the knowledge graph. These entities are harder to locate and less reusable than common entities.

For this project, the identified core entities are:

- Member
- Alumni
- Research Product (Thesis, Papers, Reports, etc...)

- Research Topic
- Project

Contextual entities

Contextual entities are those that encompass resources containing specific, often unique information directly related to the user's Purpose. These resources are designed to add distinctive value, setting the application apart from competitors. While core resources are essential for creating a functional application, contextual resources provide a competitive edge. Typically, they are not reusable and often need to be created from scratch.

For this project, the identified contextual entities are:

- Course

2.3.5 ER model definition

The entities previously defined are now used to construct the draft of a possible ER model. The model is going to be improved in the next phase during the information gathering.

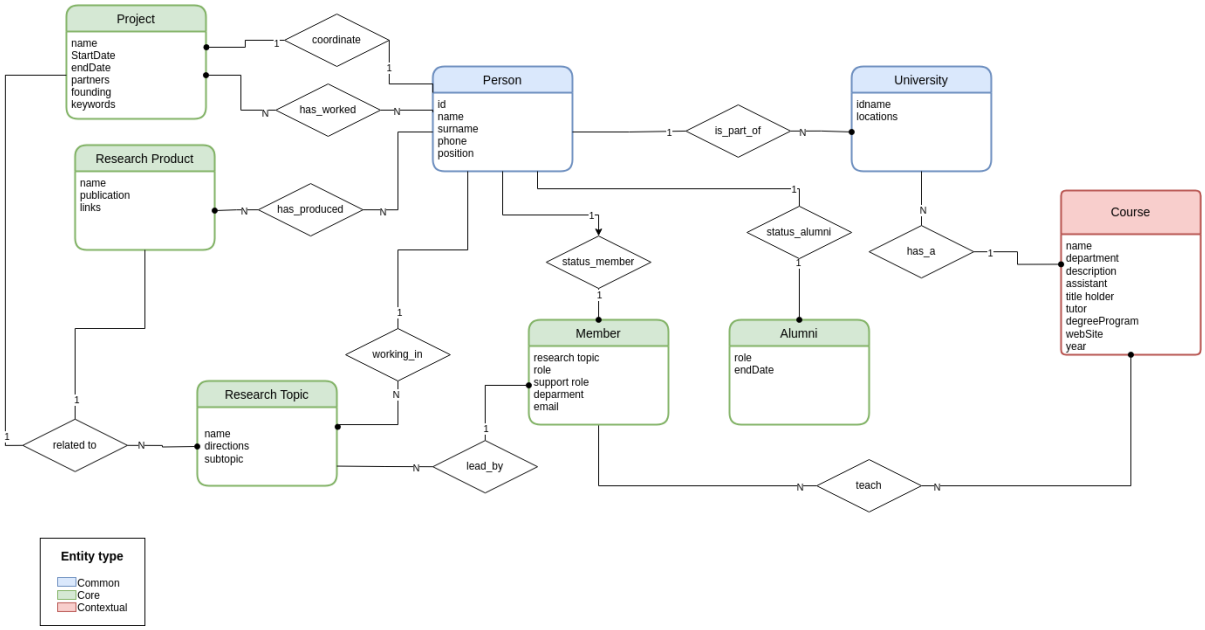


Figure 1: Entity-Relation diagram

3 Information Gathering

3.1 Sources identification

In this project there are three considered types of resources:

3.1.1 Data values datasets

For the purpose of the project we have considered numerous data sources, where some of these have to be scrapped.

The one that were suitable are:

- LiveDataUNITN catalog
  - Courses
  - Research Product
  - Staff
- Knowdive site
  The Knowdive site has been scraped through various means to obtain crucial information, such as the Research Topics of the Knowdive group, which wouldn't have been available otherwise.
  - Research Topic
  - Courses
  - Research Product
  - Project
  - Staff
- OpenStreetMap
  The usage of OpenStreetMap as a data source allows us to put together information about the various universities, especially geographical information.
  The size of the complete dataset was the first difficulty we encountered and the solution has been to download data from this source only when needed.
  The second problem we came across has been the highly variance between the entries of the various universities, with certain universities having entries in OSM for their campuses, other for their structures and others having a single unified entry. This has been resolved during the cleaning of the dataset.
  A third problem has been the lack of data on the coordinates of the universities in most of

the dataset as it is downloadable from OSM. This kind of data has been manually integrated in the dataset during cleaning.

  - University

3.1.2 Knowledge datasets(ontologies)

The knowledge sources chosen for the project were selected based on which ones would be useful in our domain and for our purpose, particularly those containing information regarding the research activities and the organizational structure of an institution. The knowledge sources identified for this project are the following:

- Digital University Ontology, which models and represents the information regarding the research activities of the University of Trento (UNITN), as well as its staff.
- VIVO Core Ontology (VIVO), which is suitable for the academic and research domain.
- Academic Institution Internal Structure Ontology (AIISO) which provides classes and properties to describe the internal organizational structure of an academic institution.
- Semantic Web for Research Communities (SWRC) which is useful for modeling entities of research communities such as persons, organizations, publications (bibliographic metadata) and their relationship.

3.1.3 Language datasets

The language dataset we use for this project are:

- DU-UNITN Language: the dataset collect and describe the terms used into the Digital University data of the University of Trento.
- KRG-UNITN Language: the dataset is created for collecting and describing the term used into the data of the Knowdive Research Group.

3.2 Datasets collection

A part of the data were easily collected, in particular, all the data from the LiveDataNum catalog:

- Courses:LiveDataUNITN-courses
- Research paper: LiveDataUNITN-papers
- Staff: LiveDataUNITN-people

The Courses dataset consist of the following field for each course:

Field name	Description
nome	Name of the course
dipartimento	The academic structure that offers the course;(nome, id)
descrizione	Textual description of the course
docenti	List of professors who teach the course; (nome, cognome, id)
assistenti	List of assistants for the course; (nome, cognome, id)
titolari	List of title holders of the course; (nome, cognome, id)
tutor	List of tutors for the course; (nome, cognome, id)
corsoStudi	Degree program of the course
sitoWeb	Official (Esse3) webpage of the course

The Staff dataset consist of the following field for each person:

Field name	Description
id	Unique identifier of the person
nome	First name
cognome	Last name
telefono	Phone number (formatter as a list)
posizioni	A list of positions;(ruolo, nomeStruttura, idStruttura)

The Research paper dataset consist of the following field for each paper:

Field name	Description
titolo	Title of the paper
tipo	Type of the paper
anno	Year of publication
lingua	Language in which the paper is written (formatter as a list)
autori	A list of people;(nome, cognome, id)
citazioni	Text used for citation in other paper
file	File of the paper;(nome, link, licenza, formato, versione, openAccess)

The data format of these sources is semi-structured and in particular JSON format, so we decided to use the JSON format to all the datasets for maintaining homogeneity.

Then we scrapped the Knowdive Research Group Site for information on:

- Member: KRG-UNITN-Member
- Alumni: KRG-UNITN-Alumni
- Research topic: KRG-UNITN-Research-Topic
- Published papers and other products: KRG-UNITN-AHC-Products+KRG-UNITN-LD-Products+KRG-UNITN-KD-Products+KRG-UNITN-HMS-Products=KRG-UNITN-Products
- Projects: KRG-UNITN-Projects
- Courses: KRG-UNITN-Courses

The information obtained was saved and integrated into JSON format compatible with the previous datasets.

The KRG member dataset consist of the following fields for each person:

Field name	Description
id	Unique identifier of the person
name	First name
surname	Last name
role	Role related to the university
researchTopic	Topic in which they work
supportRole	Role related to the Knowdive research group
university	University they are affiliate with
department	Which department they are part of
email	Person's email

The KRG alumni dataset consist of the following fields for each person:

Field name	Description
id	Unique identifier of the person
name	First name
surname	Last name
role	Role related to the university they had
university	University they are affiliate with
endDate	Date in which they left the Knowdive Research Group

The KRG research topics dataset consists of the following fields for each topic:

Field name	Description
name	The name of the research topic
Lead by	Name of the leading member on the topic
Members	list of members that work on the topic
Direction	listed directions on the Knowdive site
sub-topics	listed sub-topics on the Knowdive site
projects	listed projects on the Knowdive site

The KRG products dataset consist of the following fields for each product:

Field name	Description
name	The name of the research product
publication	Where the product is available
authors	The authors of the product
links	A link to the product

The KRG project dataset consist of the following fields for each project:

Field name	Description
name	The name of the project
start Date	When the project started
end Date	When the project ended
coordinator	The Person or University coordinating the project
partners	The list of Universities or Organizations partecipating in the project
members Partecipating	The list of members of the KRG taking part in the project
related Topics	List of the KRG Research Topics related to the project
funding	Organization or Program funding the project
keywords	List of keywords regarding the project

The KRG courses dataset consist of the following fields for each course:

Field name	Description
name	The name of the course
teachers	The listed teachers for the course
year	The year the course
university	The university that hosted the course
webpage	The webpage of the course

Finally, we also collected the OSM data in the XML format. Due to the various problems with data from this source previously mentioned, this data has been processed on a case by case basis. We provide a representation of the final fields in the next section.

The Data dataset collected are enough to answer every CQs we have described in Table 2.

3.3 Datasets cleaning

Data cleaning focuses on removing irrelevant or unnecessary ”noise” from collected datasets, ensuring that the data is accurate, relevant, and useful. We doing so by identifying and removing entities or types within a dataset that have no use and no relevance to the analysis or purpose at hand. Eventually the quality of the dataset is improved, making it more focused and valuable for its intended use.

The data scrapped from the Knowdive website didn’t undergo a substantial cleaning procedure, as most of the work was already done during the scrapping process. Only a couple of the Research Topic dataset’s fields were removed after some consideration, namely the Direction and sub-topics fields as they were, more often than not, repetitive or unnecessary.

Instead the data that come from the LiveDataUNITN catalog and OpenStreetMap need to be removed of the surplus data.

For the LiveDataUNITN catalog we removed for each datasets:

- LiveDataUNITN-Courses: all the courses that are not teached by a member of the KRG and the attributes 'dipartimento', 'tutor' and 'corsoStudi'.
- LiveDataUNITN-Papers: all the paper not written by a member or alumni of the KRG and the attribute 'citazioni'.
- LiveDataUNITN-People: all the people that are not part of the KRG and the attributes 'ssn' and 'cun'.

For the OpenStreetMap data, we cleaned up all the fields used by OSM to connect and reference its various object, leaving only the actual data about the universities. We then added manually the geographical data where it was missing and restructured the data to allow for smoother use.
The dataset now contains a list of Universities identified by their names. For each university a location field has been created, containing a list of structures connected to the specific university. Each element of the locations is structured as follows:

Field name	Description
tag	a collection of data about the location, as available through OSM
_id	ID used by OSM
_timestamp	The timestamp of the last update. Used by OSM.
latitude/longitude	The latitude and longitude values

3.4 Datasets standardization

Data standardization is crucial for ensuring consistency and usability across various datasets, particularly when they come from diverse sources. Aligning different data formats and types, permit to reuse of the data making them more accessible and easier to manage. We have decided to aligning our data to the formats of common standards such as CSV, XML, TSV, JSON, RDF, TTL and OWL to facilitate smoother data integration.

The datasets from the LiveDataNum catalog once downloaded where named with the .txt extension despite being in the JSON format. Removing the extension was enough to align them to the standard.

The dataset obtained from OpenStreetMap has been converted from its original format (XML) to json. This operation has been performed before the actual cleaning.

The dataset created from scrapping the Knowdive website were created in JSON format from the start to align with the previous datasets.

The Knowledge datasets are in the common used format of OWL and TTL.

The Language dataset downloaded is in the CSV format so the dataset we created for the Knowdive Research Group is also in the CSV format for a easy integration.

4 Additional material

In this section we have collected the links and references to all additional materials related to our project.

Github: KGE-Project Repository