Knowdive Research Group

UNIVERSITY OF TRENTO
Department of Information Engineering and Computer Science

DataScientia
Unitas per Varietatem

Knowledge Graph Engineering | Department of information engineering and computer science

Part 4.2
Phase 1 - Purpose Definition

Knowdive Research Group

UNIVERSITY OF TRENTO
Department of Information Engineering and Computer Science

DataScientia
Unitas per Varietatem

Knowledge Graph Engineering   Department of information engineering and computer science

Phase 1 - Purpose Definition
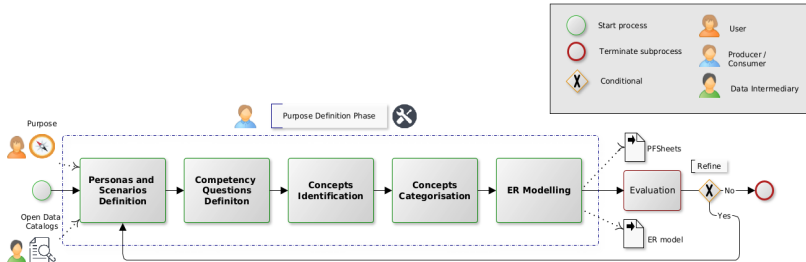
- **Input**: Purpose statement, data source list.
- **Objective**: Formalize the purpose, by extracting the functional requirements.
- **Output**: Purpose Formalization documents (PFSheets) and purpose ER model.

Phase 1 - Purpose Definition - Activities

- **Personas & Scenarios definition**: formalize the context and the actors involved in the project.

- **Competency Question definition**: state (informally) the requirements to be satisfied by the final KG.

- **Concepts Identification**: identify the information "entities" to be considered in the final KG.

- **Concept Categorization**: categorize the above entities based on their **Focus** respect to the Purpose to be satisfied.

- **ER modeling**: formalize the Purpose into an ER model.

Knowdive Research Group

UNIVERSITY OF TRENTO
Department of Information Engineering and Computer Science

DataScientia
Unitas per Varietatem

Knowledge Graph Engineering | Department of information engineering and computer science

Purpose Definition - Personas & Scenarios definition

- The goal of this activity is to **formalize the initial purpose statement** received as natural language sentence.

- Such a formalization follow an approach that aims at **extracting the specific information** from the initial Purpose.
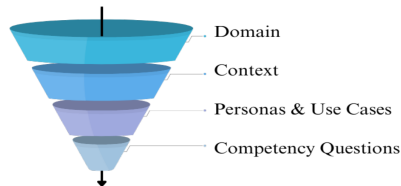

Figure 4: Funnel approach on the Purpose

Purpose Definition - Personas & Scenarios definition

- **Domain of interest**: It refers to the area of knowledge or field of study of interest [35]. Examples are the domains capturing knowledge about daily lives, such as music, tourism, and health, or geographical domains, like Trentino autonomous province.

- **Context**: The second level refers to the context description. The first prescriptive definition of Context referred to it as a location, identities of nearby people and objects, and changes to those objects [36]. More in details the context is defined over three main dimensions:

  - **Geographical boundaries**: Aspects that geographically constrain the problem.
  - **Temporal boundaries**: Aspects that constrain the problem in time.
  - **Domain boundaries**: Domain specific aspect constraining the problem.

---

[35] Fausto Giunchiglia and Biswanath Dutta, Dera: A faceted knowledge organization framework, 2011.

[36] Bill N Schilit and Marvin M Theimer, Disseminating active map information to mobile hosts, IEEE Network 8 (1994), no. 5, 22–32.

Knowdive Research Group

UNIVERSITY OF TRENTO
Department of Information Engineering and Computer Science

DataScientia
Unitas per Varietatem

Knowledge Graph Engineering   Department of information engineering and computer science

Purpose Definition - Personas & Scenarios definition

- **Personas & Scenarios**: user-centered subsets triggered by various subjects, *Personas*, and their real-world perceptions, called Use Cases or *Scenarios*. Personas generation is a widely heralded technique that provides semi-fictional subjects characterising the perception and needs of larger groups of end-users. Moreover, Use Cases are an essential complement to personas, ensuring a complete and good representation of end-users.

Knowdive Research Group

UNIVERSITY OF TRENTO
Department of Information Engineering and Computer Science

DataScientia
Unitas per Varietatem

Knowledge Graph Engineering   Department of information engineering and computer science

Purpose Definition - Competency Question definition

- The second activity add one more step in the formalization of the initial purpose, by **extracting the KG functional requirements**, from the output of the previous activity, shaping them as **Competency Questions (CQs)**.

- **Competency Questions**: a list of natural language questions. Each question defines a need (or query to the KG) that should be satisfied by the final KG. Each query refers to a Persona into a specific Scenario.

- **NOTE**: it is important to notice how the definition of CQs is crucial for the design of the final KG.
  - A **poor, set of CQs doesn't provide enough information** regarding which (information) entities needs to be modeled in the KG.
  - A set of CQs with **low heterogeneity, does not represent precisely all the possible (information) details** that the KG should be able to support.

Knowdive
Research Group

UNIVERSITY
OF TRENTO
Department of Information
Engineering and Computer Science

DataScientia
Unitas per Varietatem

Knowledge Graph Engineering   Department of information engineering and computer science

Phase 1 - Purpose Definition - Example

- Here an example of Purpose definition phase applied to a concrete iTelos project.

- Let's check how the above activities have been implemented.

Knowdive Research Group

UNIVERSITY OF TRENTO
Department of Information Engineering and Computer Science

DataScientia
Unitas per Varietatem

Knowledge Graph Engineering · Department of information engineering and computer science

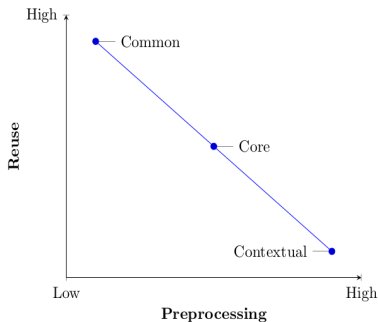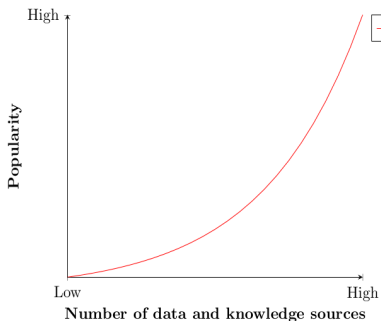# Purpose Definition - Concepts Identification

- Having the CQs listed down, the next activity aims at **extracting the concepts identifying the information entities** (and their properties) to be modeled in the KG.

- To this end a dedicated spreadsheet is adopted, called Purpose Formalization sheet (PFsheet).

- By following the **Middle-out approach**, this activity has to be done considering both the purpose (knowledge layer) and the data sources available (Data layer)

	A	B	C	D	E	F	G
	Scenarios	**Personas**	**Competency Questions**	**Entities**	**Properties**	**Focus**	**Popularity**
	1	2	2,3	Car	color, number_of_wheels, plag	Core	Common
	1	1	4	Pharmacy	address, name	Contextaual	Core
	...	...	...	...	...	...	...

Purpose Definition - Concept Categorization

- In the above figure, the PFsheet shows in the last two columns a categorization over:

  - the **Focus**, which defines how much an entity is "important" respect to the main purpose;

  - the **Popularity**, which defines how much an entity is **reused** in already existing data (considering the input information sources).

- Both Focus and Popularity, for each entity, can have three value:

  - **Common**: (Focus) general entities for the purpose considered. (Popularity) the entity is largely available in existing resources.

  - **Core**: (Focus) specific entities for the purpose considered. (Popularity) the entity is available in existing resources but not so common.

  - **Contextual**: (Focus) very specific entities for the purpose considered. (Popularity) the entity is not available in existing resources.

- By following the **Middle-out approach**, this activity has to be done considering both the purpose (knowledge layer) and the data sources available (Data layer)

Purpose Definition - Concept Categorization

Knowdive Research Group

UNIVERSITY OF TRENTO
Department of Information Engineering and Computer Science

DataScientia
Unitas per Varietatem

Knowledge Graph Engineering   Department of information engineering and computer science

Purpose Definition - ER modeling

- The next activity, in the first iTelos phase, aims at producing a **formal representation of the initial purpose**, shaped as ER model.

    - the **ER model** is the first (graphical) version of the final KG structure (or knowledge layer).

Knowdive Research Group

UNIVERSITY OF TRENTO
Department of Information Engineering and Computer Science

DataScientia
Unitas per Varietatem

Knowledge Graph Engineering | Department of information engineering and computer science

ER Model - Definition

- An **Entity–Relationship (ER) Model** describes interrelated things of interest in a specific domain of knowledge.

- It is composed of **classes** / **entity types** (etypes) (which classify the things of interest, i.e. **entities**) and specifies **relationships** that can exist between entities (instances of those entity types).

- The ER model is, thus, an **abstract data model** that defines a data or information structure which can be implemented in a data/knowledge base.

- It is usually drawn in a graphical form as **boxes (classes)** that are connected by **lines (relationships)** which express the associations and dependencies between entities.

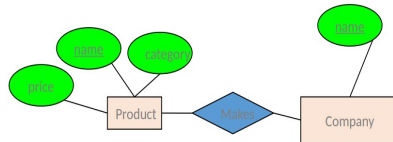- An ER model is the **informal foundation** for the specification of domain-specific teleologies.

ER Model - Notations

Following are the main components and its symbols in ER Diagrams:

- **Rectangles**: This Entity Relationship Diagram symbol represents entity types.

- **Ellipses** : This symbol represents attributes.

- **Diamonds**: This symbol represents relationship types.

- **Lines**: It links attributes to entity types and entity types with other relationship types.

- **Identifying attributes** are underlined.

Knowdive
Research Group

UNIVERSITY
OF TRENTO
Department of Information
Engineering and Computer Science

DataScientia
Unitas per Varietatem

Knowledge Graph Engineering   Department of information engineering and computer science

ER Model - A Simple Example

ER Model & Diagrams

ER is a *visual syntax* for DB design which is **precise enough** for technical points, but **abstracted enough** for non-technical people.

Reference: M. Hahsler. DS1300: The ER Model.

ER Model - Entities and Entity Types

- **Entities** & **entity types** are the primitive units of the ER model

  - <u>Entities</u> are the individual objects (instances), which are members of entity types
  - <u>Entity type</u> are the *classes* or *types* of objects in our model
  - Example: Person is an entity type while Michael is an entity.
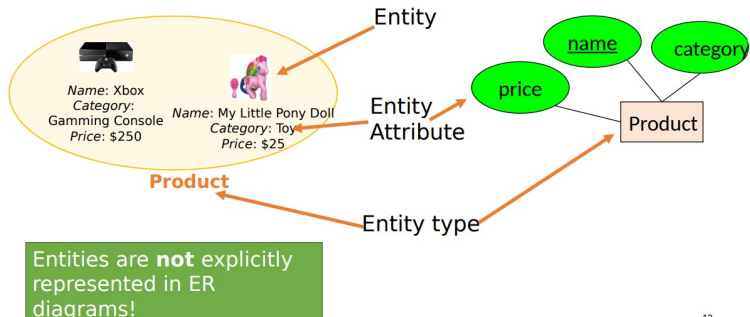 - *We **use entity types** in ER models*

Product

Person

Reference: M. Hahsler. DS1300: The ER Model.

Knowdive
Research Group

UNIVERSITY
OF TRENTO
Department of Information
Engineering and Computer Science

DataScientia
Unitas per Varietatem

Knowledge Graph Engineering   Department of information engineering and computer science

ER Model - Entities vs. Entity Types

Example:

Entity

Entity
Attribute

Name: Xbox
Category:
Gamming Console
Price: $250

Name: My Little Pony Doll
Category: Toy
Price: $25

Product

name

category

price

Product

Entity type

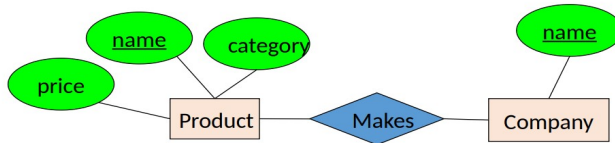Entities are **not** explicitly
represented in ER
diagrams!

Reference: M. Hahsler. DS1300: The ER Model.

ER Model - Relationships

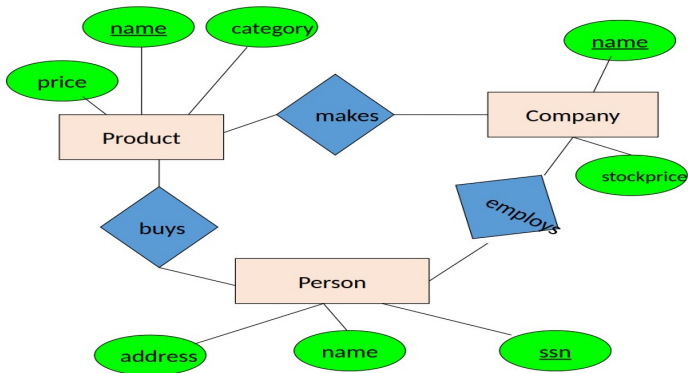- A **relationship type** is between two entity types

> **How to read a relationship in both directions:**
> 1. A product is made by a company
> 2. A company makes a product

Reference: M. Hahsler. DS1300: The ER Model.

ER Model - Example

Reference: M. Hahsler. DS1300: The ER Model.

Phase 1 - Purpose Definition - Example

- Here an example of Purpose definition phase applied to a concrete iTelos project.

- Let's check how the above last activities have been implemented.

Phase 1 - Purpose Definition - Tools, templates & output

- What we need to implement the Purpose Definition phase?

- **Tools**:
    - **Document Writer**: to write the project report with Domain, Context, Personas, Scenarios and CQs.
    - **Spreadsheet editor**: to fill the PFsheet for Concept Identification and Categorisation activities.
    - **yEd**: to model the ER (ot any other diagram editor).

- **Templates**:
    - Project Report Template
    - PFsheet template

Knowdive Research Group

UNIVERSITY OF TRENTO
Department of Information Engineering and Computer Science

DataScientia
Unitas per Varietatem

Knowledge Graph Engineering   Department of information engineering and computer science

Phase 1 - Purpose Definition - Producer & Consumer

- As already discussed, the methodology can be applied both as a data producer and consumer.
    - **What is the difference in this phase?**

- **The producer** aims at creating resources, thus it's purpose is to enhance the availability of such resources, in one (or more) specific domain and contexts.
    - Notice how such resources will be applied fo specific purposes, even if the producer did not consider them (or they have been considered in a more generic way) during the resource production.

- **The consumer** already has a specific purpose to satisfy, thus all the activities are leaded by such purpose.

- **NOTE**: The formalization of the purpose (general or specific) in the first phase of iTelos, is crucial to define the output of the project, both on producer and consumer side.