3 Language Definition

Until now, our purpose formulation has been carried out without a clearly defined language resource. This presents a significant challenge, as the words or concept structures used to represent the Etypes and their properties can be interpreted differently by users. Furthermore, some words are polysemous, meaning they have multiple meanings, which makes them ambiguous. Therefore, it is crucial to address this linguistic diversity by associating each concept with a formal definition. This can be effectively achieved using the Universal Knowledge Core (UKC), a high-quality, diversity-aware database. Following the iTelos methodology, the concepts to be identified should initially represent the Etypes, object properties, and data properties. These elements were already established during Phase 1, resulting in the creation of a CSV file that serves as a language dataset. This dataset contains formally defined concepts tailored to our purpose-specific domain.

Three different tables are presented below: the first contains the entity types, the second the properties, and the third the relationships. For each concept, an ID has been assigned. Most of these concepts are found in the UKC ontologies. At the same time, for more biological terms, such as Microbiome or Relative Species Abundance, it was possible to locate them in other ontologies available on BioPortal, with the specific link provided. However, some concepts could not be found in any external ontology and have been identified using the code KGE-QCB1-number.

For the term 'Brinkman Index,' while some similar concepts related to smoking habits were found in biological ontologies (such as the number of cigarettes smoked per person), we decided to create a new ID. This is because the Brinkman Index is calculated specifically as the product of the number of cigarettes smoked per day and the number of years of smoking, which makes it distinct from other related concepts.

ConceptID	**Word-en**	**Gloss-en**
UKC-36	Person	A human being
KGE_QCB1-1	Risk factor	Something that makes a person more likely to get a particular disease or condition
OHMI_0000003	Microbiome	A biome that consists of a collection of microorganisms (i.e., microbiota) and the surrounding environment where the microorganisms reside
UKC-43176	Species	A taxonomic group whose members can interbreed (biology)
UKC-65892	Correlation	A reciprocal relation between two or more things
UKC-67961	Cancer	Any malignant growth or tumor caused by abnormal and uncontrolled cell division; may spread to other parts of the body through the lymphatic system or the blood stream
UKC-27611	Stage	A position on a scale of intensity, amount or quality

Table 5: Language concepts for e-types

ConceptID	Word-en	Gloss-en
UKC-681	has_Medical_diagnosis	Identification of a disease from its symptoms
KGE-QCB1-2	has_Species	A person has a taxonomic group whose members can interbreed.
KGE-QCB1-3	has_Interaction	A species correlates with a particular person
KGE-QCB1-4	has_Risk_Factor	A person is associated with risk factors.

Table 6: Language concepts for e-types relations (object properties)

ConceptID	Word-en	Gloss-en
UKC-44477	Taxonomy	A classification of organisms into groups based on similarities of structure or origin etc.
UKC-26728	Age	How long something has existed.
UKC-27174	Gender	The properties that distinguish organisms based on their reproductive roles.
UKC-45187	Country	The territory occupied by a nation.
http://purl.bioontology.org/ontology/MESH/D015992	BMI	An indicator of body density as determined by the relationship of BODY WEIGHT to BODY HEIGHT. BMI = weight (kg) / height squared (m^2).
KGE-QCB1-5	Brinkman Index	Is calculated from cigarettes per day times smoking years.
KGE-QCB1-6	Alcohol level	Is a measure of alcohol in the blood as a percentage.
UKC-2	Name	A language unit by which a person or thing is known.
UKC-66329	Medium	A state that is intermediate between extremes; a middle position.
UKC-80756	Low	Less than normal in degree or intensity or amount.
UKC-80747	High	Greater than normal in degree or intensity or amount.
http://purl.obolibrary.org/obo/OHMI_0000468	Relative Species Abundance	A quality of ecological community that refers to how common or rare a species is relative to other species in a defined location or community.

Table 7: Language concepts for e-types attributes (data properties)

Data Filtering

The second part of the language definition focuses on the data filtering process to ensure that the data layer resources match the identified concepts. In this case, no further filtering is needed because the resources are already well-aligned with the data layer.