Part 3
The Solution - iTelos

1 Part 0 - Course Organization

2 Part 1 - The Reuse Problem

3 Part 2 - State of the Art

4 Part 3 - The Solution iTelos

5 Part 4 - The iTelos Methodology

Part 3.1
EML data representation language

1. EML data representation language
2. iTelos data reuse processes
3. Distributed Stratified Data Mesh

Entity Modeling Language (EML)

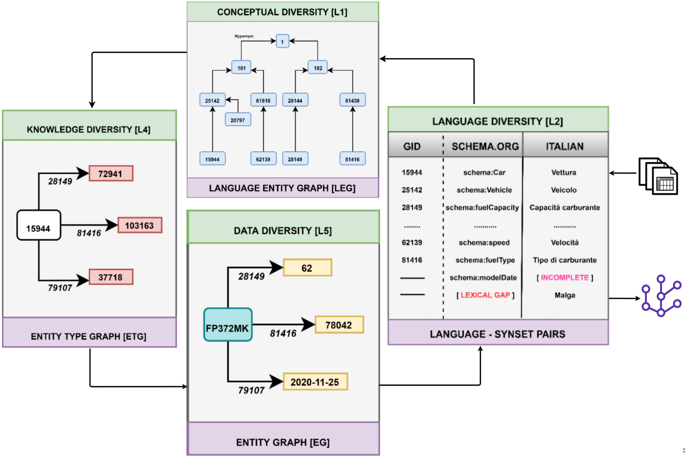
- EML is a data representation language able to **perceive and represent the data heterogeneity**.
- EML aims at **homogenising the heterogeneity of data by providing unique representations** for all four levels of heterogeneity.
  - EML-Sc homogeneity in Source heterogeneity
  - EML-F homogeneity in Format heterogeneity
  - EML-S homogeneity in Structure heterogeneity
  - EML-M homogeneity in Meaning heterogeneity

EML - Stratified Approach

- EML is designed over a **stratified information approach** in which the information is always composed by three layers of resources.
- Such layers are those already discussed as types of reusable resources
  - (L) Language (and concepts) resources
  - (K) Knowledge resources
  - (D) Data values resources

EML - Stratified Approach

- The stratified approach defines **how the information is a represented as a composition of different resources of different layers**.



GID	SCHEMA.ORG	ITALIAN
15944	schema:Car	Vettura
25142	schema:Vehicle	Veicolo
28149	schema:fuelCapacity	Capacità carburante
....	....	....
62139	schema:speed	Velocità
81416	schema:fuelType	Tipo di carburante
——	schema:modelDate	[INCOMPLETE]
——	[LEXICAL GAP]	Malga

Entity Modeling Language (EML)

- Therefore, EML handles the data heterogeneity by considering two orthogonal dimensions: **the type of heterogeneity, and the information layers**.
- This means that EML defines homogeneity for:
  - EML-Sc(L), EML-Sc(K), EML-Sc(D)
  - EML-F(L), EML-F(K), EML-F(D)
  - EML-S(L), EML-S(K), EML-S(D)
  - EML-M(L), EML-M(K), EML-M(D)
- Let's discover more concretely how EML defines all the above data homogeneity components.

EML-Sc

- As already discussed "*Source heterogeneity refers to the divers possible sources of information from which information resources can be collected*".
    - The heterogeneity is caused by the different interpretations of the information that each source has, over the same target reality.
- EML-Sc aims at homogenizing such heterogeneity, by **selecting reliable and standardized sources**, thus **limiting noisy interpretation of the target reality** respect to a specific purpose.
- Concretely EML-Sc is the **set of information sources** from which the project resources can be collected.
    - Websites
    - Data catalogs
    - Databases and Knowledge bases
- The set of sources defined by the EML-Sc is applied over Language, Knowledge and Data resources, thus defining in specific the sub-sets: **EML-Sc(L), EML-Sc(K), EML-Sc(D)**.

EML-F

- *"Format heterogeneity refers to the divers possible data formats that can be employed to differently encode information"*.
- EML-F aims at homogenizing such heterogeneity for all the three types of information (Language, Knowledge and Data), by **defining which standardized and well-known formats have to be applied to represent the information to be reused**.
  - **EML-F(L)**: Excel, XML
  - **EML-F(K)**: RDF-OWL [26] [27]
  - **EML-F(D)**: JSON, Excel, CSV, RDF

[26] RDF documentation

[27] OWL documentation

EML-S

- *"Structure heterogeneity is conventionally understood as the existence of variance in the representation and description of the same target reality"*.
- EML-S aims at homogenizing such heterogeneity by **defining the structures of the information elements** for each types of information (Language, Knowledge and Data).
  - **EML-S(L)**: defines the *concept* structure
    [word, sense, synset, concept] - UKC structure
  - **EML-S(K)**: defines the *EType* structure
    [data and object properties, annotations]
  - **EML-S(D)**: defines the values *data types*
    [string, int, long, bool, ...]

EML-M

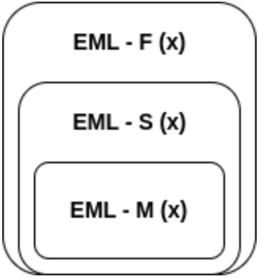
- *"Meaning Heterogeneity, is defined over the values of the information properties which can be used to identify a real world entity"*.
- EML-M aims at homogenizing such heterogeneity by **defining which values have to be adopted to identify real world entities, and how to shape such identifiers**.
- For each types of information, EML-M defines the required identifiers.
  - **EML-M(L)**: concept ID
  - **EML-M(K)**: EType ID
  - **EML-M(D)**: SURI/SURL

EML outcome

- The homogeneity introduced by EML over the resources to be reused, **reduces the cost to exploit thus resource in future** (reusability), due to:
    - the reduction of the heterogeneity,
    - the adoption of standards formats and information structures, and,
    - the adoption of identifiers.

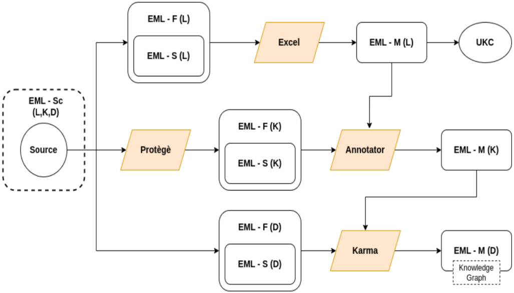
Progressive Ordered Encapsulation (POE)

- It is important to notice how the homogeneity introduced by EML is **progressively encapsulated starting from the information sources until the generation of EML-compliant reusable resources**.
  - This means that EML-M includes the homogeneity defined in EML-S, as well as, EML-S includes the homogeneity of EML-F, and so on.



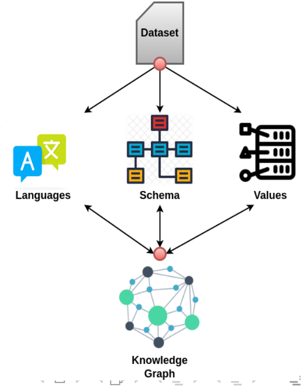
EML generation process

- The generation of EML-compliant resources is a responsibility of the **data reuse processes** that will be described in the next section.
- Nevertheless we can report here which is the EML generation process that will be considered in the top level processes, together with the tools allowing for the EML generation.



EML standard

- The different layers extracted (defined explicitly) from “conventional” datasets are **composed together** into a **Knowledge Graph** (KG).
- Why a Knowledge Graph ?
  - adaptability (different contexts)
  - scalability
  - its structure allows:
    - the information **layers composition**, and,
    - the **KG decomposition** into single layers resources.


Dataset
Languages
Schema
Values
Knowledge Graph