Part 3.2
iTelos data reuse processes

iTelos Reuse Processes - Data Flow

Low quality and low reusability of data make the reuse of data time-consuming and costly. Thus enhancing the cost of the communication between *Producer* and *Consumer*.

To reduce such costs, the idea is to introduce a new agent as mediator between producer and consumer, having the objective of handling data to make them reusable and interoperable:

- **The Data Intermediary**

Data Intermediary

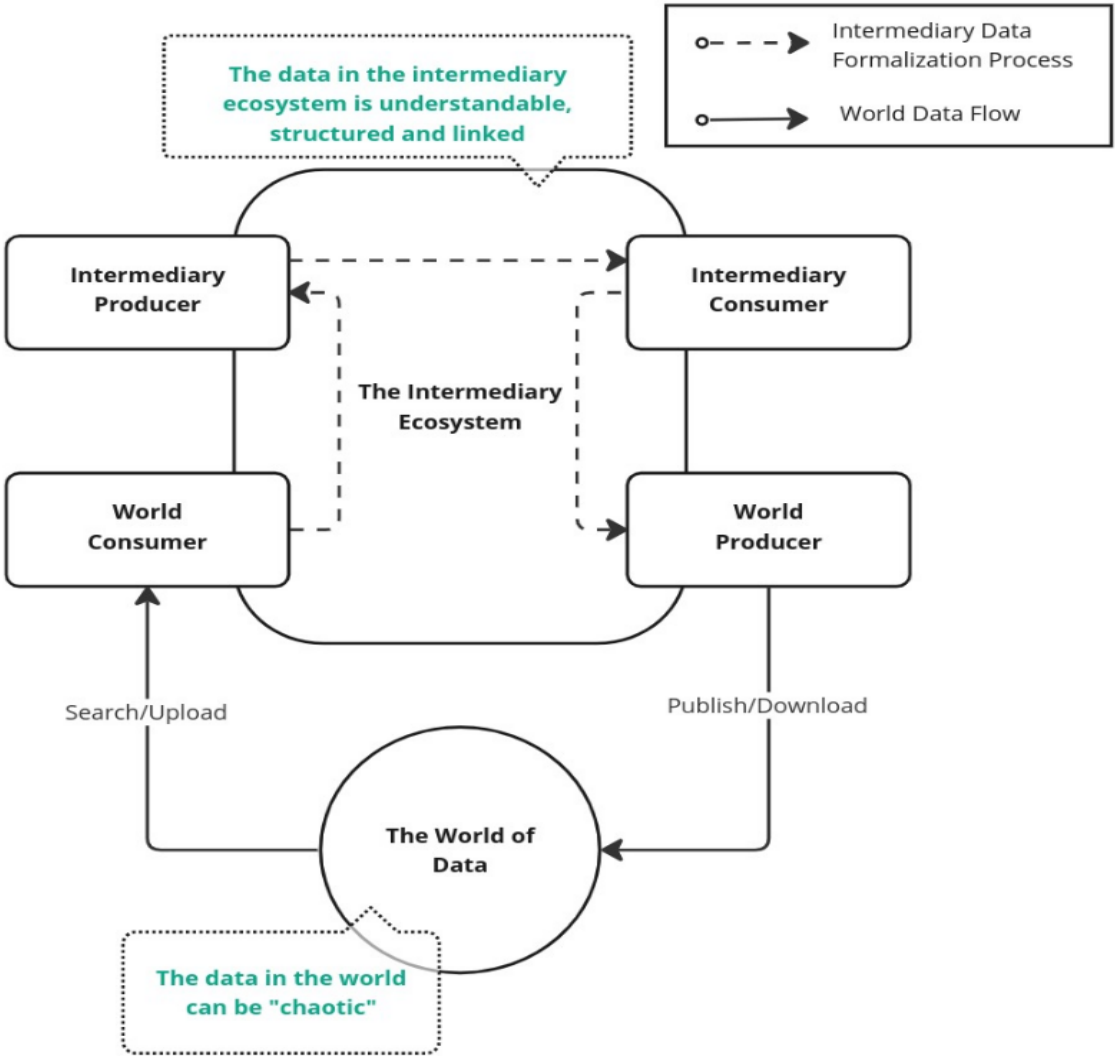
- **Definition**: A data intermediary is a *human agent* who supervises a structured *process* for the production of *language* specific resources, implemented into a dedicated *environment*.
  - Thus, DI := [language + process + environment].
- **Objective**: A data intermediary aims at:
  - collecting data from the (disordered) heterogeneous world of data;
  - cleaning and formatting following well-known standards;
  - increasing the reusability of such data;
  - generating purpose-specific reusable data to support specific application and services;
  - sharing high quality and reusable data to enhance homogeneity into the heterogeneous world of data.

Data Intermediary

To achieve its objectives, the data intermediary **agent** is internally divided in:

- **Intermediary Producer**: it collects and transforms low quality data into high quality and reusable resources.
- **Intermediary Consumer**: based on a specific purpose, it composes the high quality resources, produced by the intermediary producer, to support purpose-specific application and data services.

Data Intermediary Action



Data Intermediary

- Concretely the Data Intermediary is composed by:
  - The EML language,
  - The iTelos data reuse processes.
  - The Distributed Stratified Data Mesh.
- While the EML have been already described above, the below sections will be focused on the remaining two components.

iTelos Dat Reuse Processes

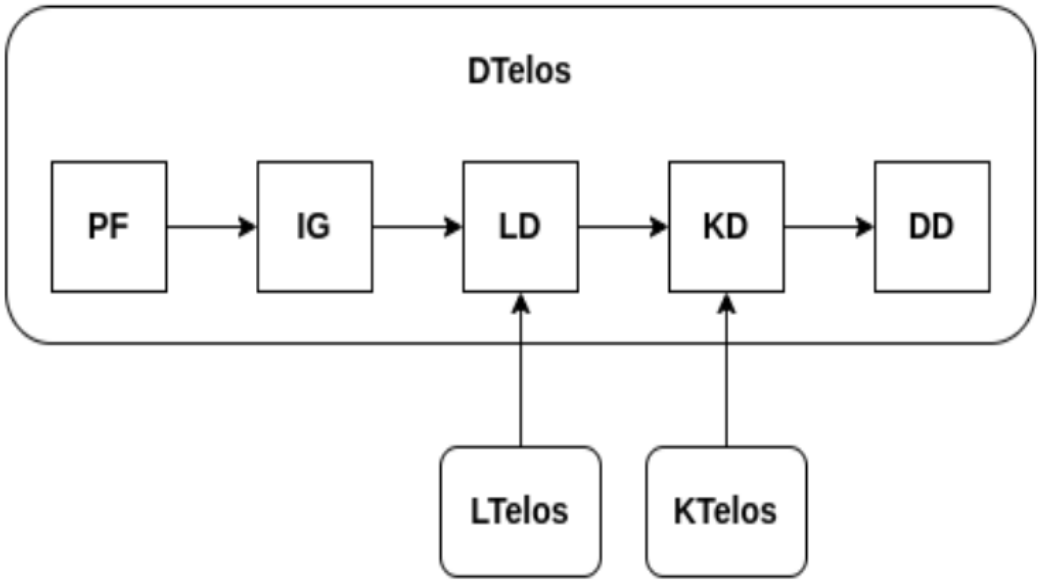
- **iTelos** is a **structured, phase-based methodology, implemented by three parallel process that share the same methodology structure**.
  - The three processes are dedicated to the three types of information respectively.
    - **LTelos**: generation of EML-M(L) resources
    - **KTelos**: generation of EML-M(K) resources
    - **DTelos**: generation of EML-M(D) resources

iTelos Dat Reuse Processes

- Due to the POE feature of EML, the progressive encapsulation is transferred over the iTelos process too.
  - For this reason **the DTelos process involves the (parallel) execution of KTelos, as well as, KTelos involves the execution of LTelos.**
- **Note**: For lack of time, the KGE course is focused on DTelos, assuming the other two processes executed in parallel.

iTelos Data Reuse Process - DTelos

- Below the DTelos process structure, **based on the iTelos methodology**, where the LTelos and KTelos processes are taken in input (see next slide for the details over each single process phase).



iTelos Methodology

iTelos [28] is a phase-based methodology (implemented by LTelos, KTelos and DTelos) that,

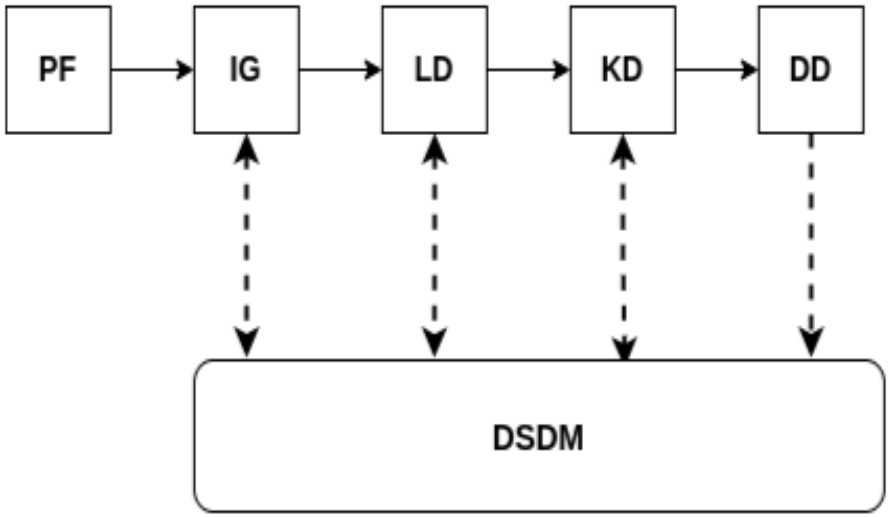
- takes in **input** a set of data resources, having an arbitrary level of quality.
- The methodology structure is the same for:
  - **Intermediary Data Producer (IDP)**: collecting and transforming existing resources into reusable (EML-compliant) resources.
  - **Intermediary Data Consumer (IDC)**: composing already produced EML-compliant resources.
- And produce as **output** high quality data, shaped as:
  - (IDP) different knowledge graphs (e.i., one for each input dataset);
  - (IDC) a single knowledge graph created by composition of existing KGs and/or lower quality resources.

[28] From greek "**telos**" which means purpose

Part 3.3
Distributed Stratified Data Mesh (DSDM)

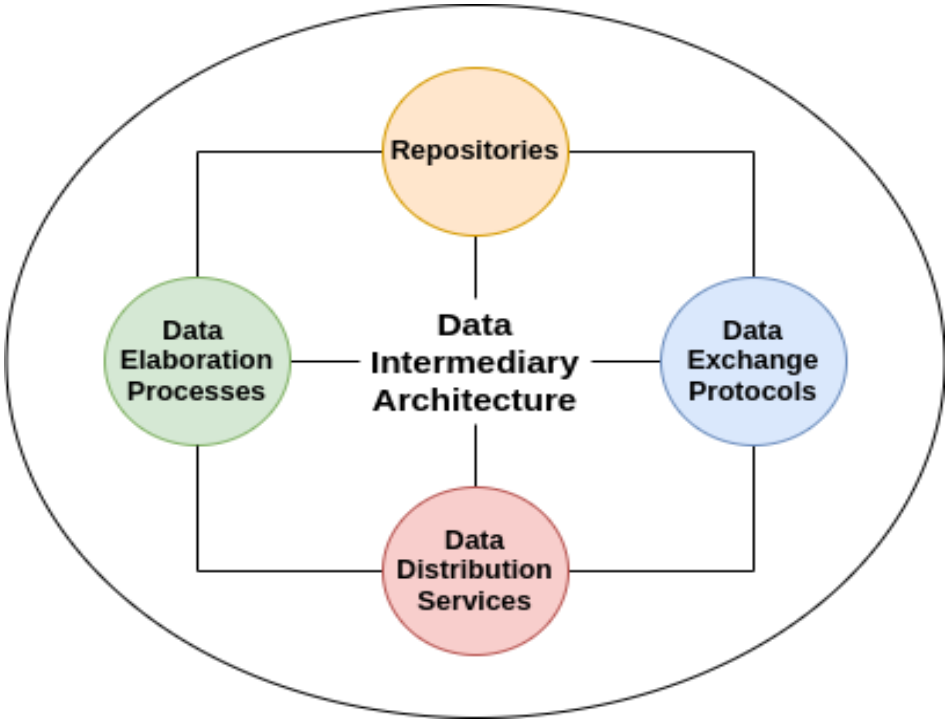
Data intermediary architecture

- To implements the iTelos methodology, thus supporting LTelos, KTelos and DTelos, the data intermediary needs a dedicated data architecture, called **The Distributed Stratified Data Mesh (DSDM)**.



Data intermediary architecture

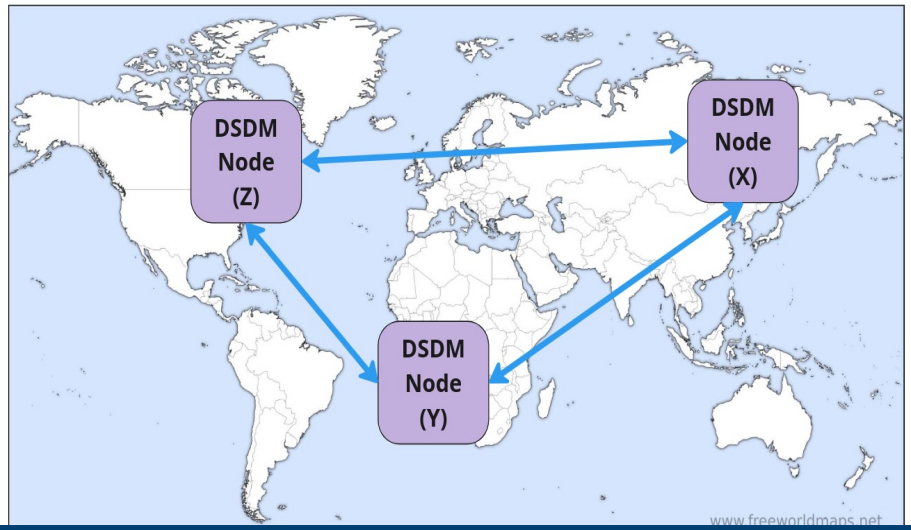
- The DSDM is defined by **different components communicating each other**.



Distributed Stratified Data Mesh (DSDM)

More in details the intermediary data architecture is a Distributed Stratified Data Mesh (DSDM);

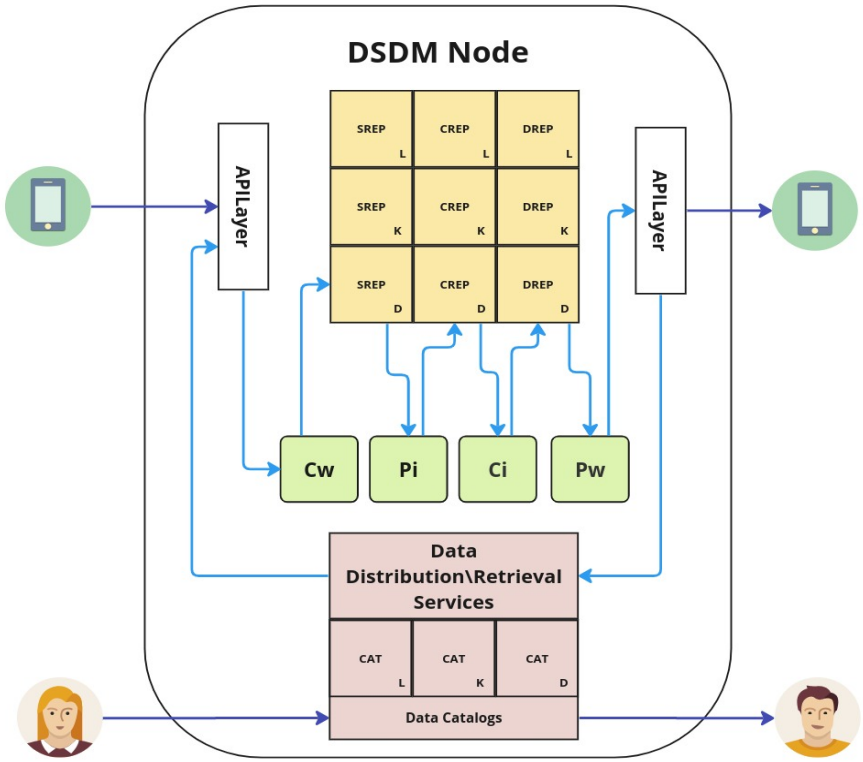
- it includes all the data mesh features;
- additionally, it is composed by different **nodes**, where each node;
  - is defined for a **specific domain** of interest, or purpose (i.e., geographically);
  - **autonomously manages resources** about its domain/purpose (local domain experts handling data);
  - has a **local implementation** of the intermediary data architecture;
  - For each node, **it handles stratified resources** (Language, Knowledge and Data)



Distributed Stratified Data Mesh - The Node

Each node of the DSDM
is composed by the following components
(Detailed in the following slides):

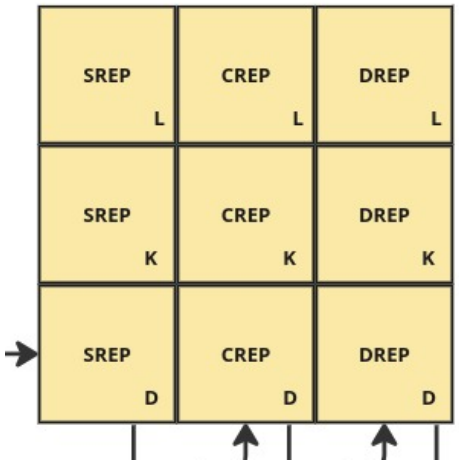
- Data repositories
- Data elaboration processes
- Data exchange protocols
- Data distribution services



Data Repositories

Each DSDM node includes three different repositories, which are distinguished on the basis of what data they contain.

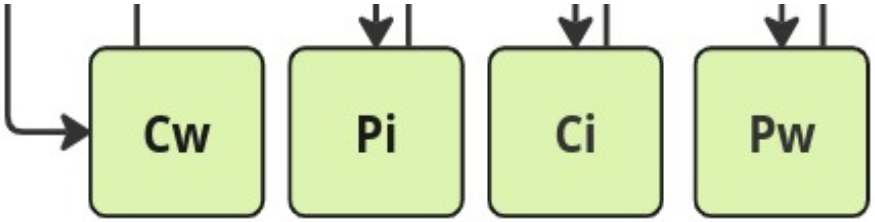
- **SREP**: the Source REPository stores the data collected form the "disordered" world, which need to be processed to make it compliant with the data intermediary data requirements about quality and reusability.
- **CREP**: the Core REPository stores the data that has been processed by iTelos, thus being compliant with the intermediary data requirements.



- **DREP**: the Distribution REPository stores the data which can be accessed by the data distribution services. In other words, the data which can be shared out of the DSDM node.

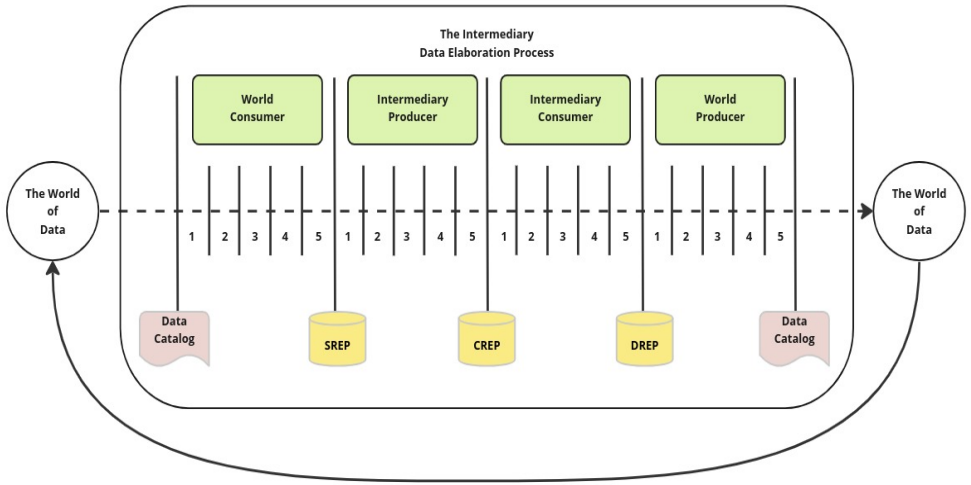
Data Elaboration Processes

- The data elaboration processes are responsible for the **collection and "transformation"** of the data, that before and after each process are stores in one of the above described repositories.



Data Elaboration Processes

- The data elaboration processes are **four different instances of the DTelos process** [29].
- Depending on the objective to be achieved, DTelos can be adopted, by exploiting the features offered by its different phases.
  - Some phases are more (or less) exploited then others depending by the elaboration process which need to be executed.



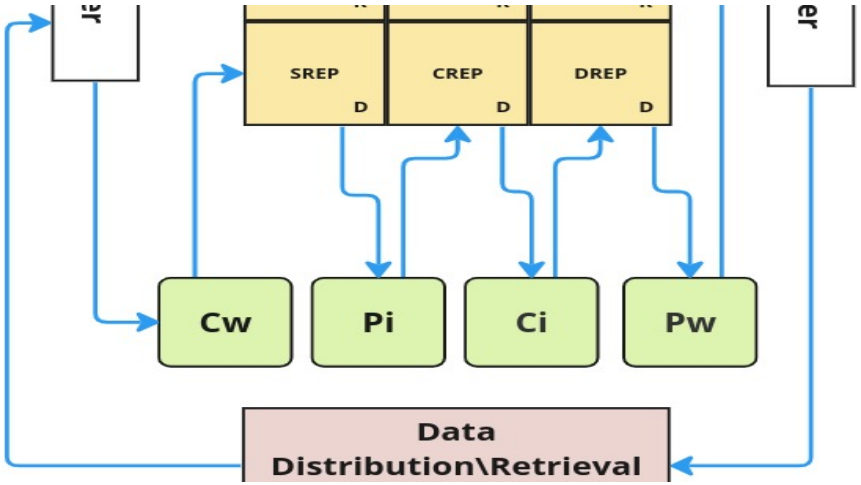
1. **Purpose Formalization**
2. **Information Gathering**
3. **Language Definition**
4. **Knowledge Definition**
5. **Data Definition**

miro

[29] different problems, one methodology

Data Exchange Protocols

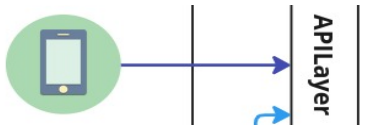
- The data exchange protocols are distinguished in two types:
  - **Internal protocols**: they define the exchange of data, **within the DSDM node**, between repositories, processes and data distribution services.



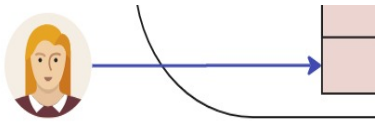
Data Exchange Protocols

- **External protocols**: they define the exchange of data, **across different DSDM nodes**, by considering

- automatic data exchange (device to device)

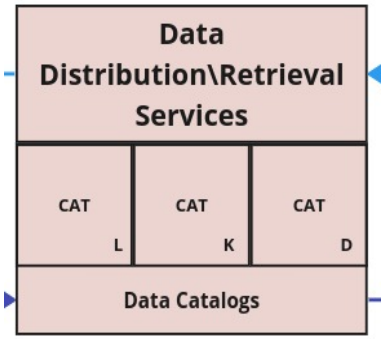


- human-driven data exchange (human to device)



Data Distribution Services

- The data distribution services aims at **sharing the resources** produced, and handled, by the DSDM node.
- Such services plays a crucial role in the **reusability** of the data accessible in the whole DSDM.
  - The exploitability of the data increases.
  - The effort in building new (EML-compliant) quality data, decreases.



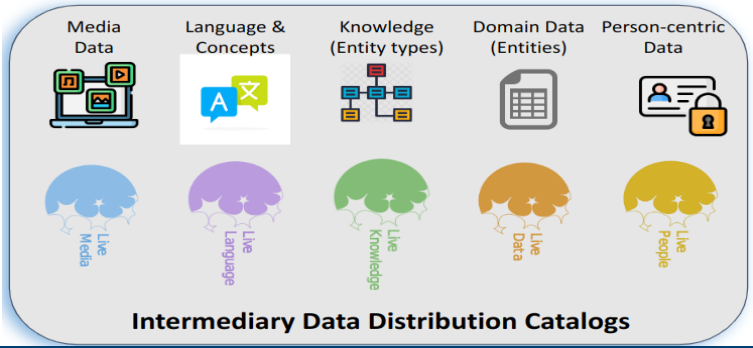
Note: more details about the intermediary data distribution are provided in the following slides.

Intermediary Data Distribution

- Within the data intermediary architecture, as already mentioned above, a crucial role is played by the data distribution.
- The importance of such component in the architecture, leads to the definition of a **dedicated architecture** for the data distribution.
  - **The Distributed Metadata Mesh**
- The metadata mesh is directly **mapped over the DSDM nodes**, but it handles **metadata** describing the data that has to be shared.
  - This allows the whole intermediary architecture to:
    - **provide information** about the data to be shared, through their metadata;
    - **protect** the data, and distribute it only once it needs to be exploited.

Distributed Metadata Mesh - Nodes

- The single node of the metadata mesh, enables the exploitation of the data distribution (and retrieval) services.
- To this end, a single node is composed by 4 resources catalogs, the we can divide in two categories:
  - **data catalogs**: these catalogs shares metadata about domain data and person-centric data. Their names are **LiveData** and **LivePeople**, respectively. (Media data will be considered in future)
  - **interoperability catalogs**: these catalogs shares metadata about language and knowledge resources, which can be used to represent the data. The resources considered by these catalog are provided to enhance the **interoperability** of the data. The catalogs have been called, **LiveLanguage** and **LiveKnowledge** respectively.



Distributed Metadata Mesh - Catalogs Links

- The KG(s) produced by DTelos, within each DSDM nodes, are **stratified**, thus composed by language, knowledge and data resources.
  - The link among the different resources, composing a KG, is maintained also at metadata level.

Example: a data resource, shared in LiveData, has a specific metadata linking to the knowledge resources, in LiveKnowledge, defining the initial data schema. Using the same approach the knowldge resources is linked to one (or more) language resources in LiveLanguage.

Linked data catalog example

Distributed Metadata Mesh - Navigation entry point

- The distributed data mesh can be accessed by the users who wants to navigate it, through the catalogs webportals.
- The entry point for such catalog navigation, is the **Main LiveData** catalog.
  - This "top-level" catalog, unlike the others catalogs, collects metadata about the the local LiveData catalogs, providing the direct access to them.

Distributed Metadata Mesh - Services

- The metadata mesh provides a set of services which can exploited through the catalogs. Here below the list of the available services.
  - **Catalog deployment**: offered by the Main LiveData catalog, it allows new organization and/or users, to easily deploy a new domain specific data LiveData catalog (service that can be adopted also for the other types of catalogs).
  - **Data Upload**: offered by each catalog, it allows the user to upload new data, in the relative DDM node (notice that such data is still not published).
  - **Data Publication**: offered by each catalog, it allows the user to publish a set of metadata for a new resource (already available in the DDM node repositories) to be shared.

Distributed Metadata Mesh - Services

- **Data Search**: offered by each catalog, it allows the user to find resources by searching for its metadata values.
- **Data download**: offered by each catalog, in a different way depending by the data access policy defined by the DDM node, this service allows the user to download the data required.
- **Data composition on demand** (under development): this services will allow the users to select and compose the resources (language, knowledge and data) they need to build a new KG.