Part 1
The Reuse Problem

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

6 Part 5 - Exploiting the KG

Part 1.1
Information & information reuse

1 Information & information reuse

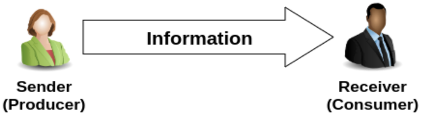
2 Data representation

3 Data reuse processes

4 Data architecture for reuse

Information

The information is stimuli (i.e., electromagnetic waves), created by a sender, **that has meaning in some context** for its receiver.



- The information is represented, using its raw form, by the **data**. The data, represented and managed in different ways, transports the information within a **communication**, from sender to receiver.
- Notice how a communication can be the creation of data (Producer) to be exploited by any kind of data service (Consumer).

Information Reuse

- The information in a communication is not always new, most of the time instead is **reused** from previous communication.
- As a consequence, the **reuse of information, and thus data**, is crucial in a communication between sender(producer) and receiver(consumer).

What does it means reuse of data ?

The data reuse is defined over three components:

1. Data representation
2. Data reuse processes
3. Data architecture for reuse

Part 1.2
Data representation

1 Information & information reuse

2 Data representation

3 Data reuse processes

4 Data architecture for reuse

Data Representation

- The data needs to represent different aspects that the information has to express in a communication.
- Reusable data is present, and available, in a (data) **world that is, apparently, disordered**.
  - ”disorder”: high level of **heterogeneity**, low quality, requiring **huge amount of pre-processing**.
- The data appears in multiple forms, highlighting what is called **data heterogeneity**.

Data Representation - Heterogeneity

- The general meaning of *Heterogeneity* is the *"quality or state of consisting of dissimilar or diverse elements"*[1].
- Heterogeneity is the key distinguishing feature of life: there will never be two identical moments, two identical places, two identical individuals, or two datasets!
- *Data Heterogeneity*, therefore, is the principle bottleneck in achieving reuse and integration data.

[1] https://www.merriam-webster.com/dictionary/heterogeneity

Data Representation - Heterogeneity

Data heterogeneity is defined over four encapsulated sub-layers of heterogeneity:

1. Source heterogeneity
2. Format heterogeneity
3. Structure heterogeneity
4. Meaning heterogeneity

Data Representation - Source Heterogeneity

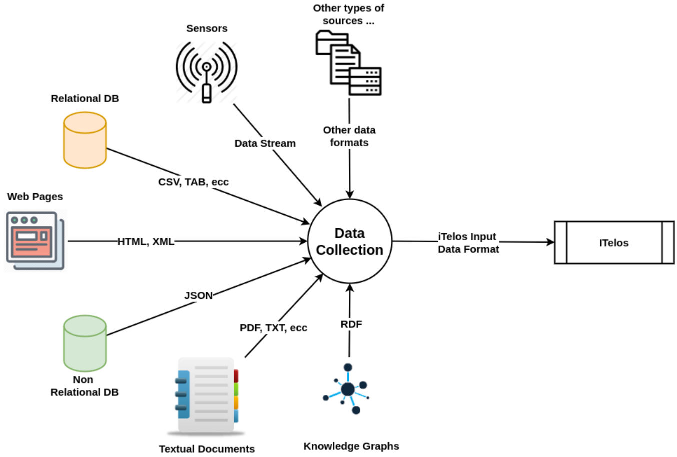
- Information can be transmitted through different modes, for instance, via:
  - Visual Mode.
  - Linguistic Mode.
  - Aural Mode.
  - Gestural Mode, ... etc.!
- Within each such mode, there can be several possible information sources, e.g.,:
  - Visual: Art, Photos, Videos, etc.
  - Linguistic: Written text in different languages.
  - Aural: Music, Speech, etc.
  - .... etc.!

Data Representation - Source Heterogeneity

- Source Heterogeneity refers to the *diverse possible sources of information* that can be employed to differently record information about the same *target reality*.
- For example, information about the same car can be differently recorded via:
  - Datasets recording different properties of the car.
  - Written textual description of the car in different languages.
  - Photos of the car from different angles.
  - Videos of the car from different angles.
  - A speech about the car.
  - ... etc.!

Source heterogeneity

- The heterogeneity at source layer



Data Representation - Format Heterogeneity

- For each source of information, there can be possibly many types of data *formats* which can be used to encode information about a target reality.
- For example, following are some formats which employ different syntax to encode information:
  - Images: JPEG, PNG, TIFF, BMP, SVG, etc.
  - Videos: WebM, MKV, FLV, etc.
  - Text: Doc, PDF, RTF, TXT, etc.
  - Datasets: JOSN, CSV, RDF, etc.
  - .... etc.!

Data Representation - Format Heterogeneity

- Even assuming the same source, Format Heterogeneity refers to the *diverse possible data formats* that can be employed to differently encode information about the same *target reality*.
- For example, information about the same car can be differently encoded via:
    - Datasets recording different properties of the car in CSV or JSON or RDF.
    - Written textual description of the car in different languages in PDF or DOC or TXT.
    - Photos of the car from different angles in JPEG or PNG.
    - Videos of the car from different angles in FLV or MKV.
    - ... etc.!

Data Representation - Structure Heterogeneity

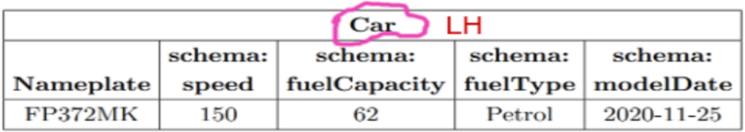
1. Even assuming the same source and format, heterogeneity appears over the structure of the information within the data, that we can call **Structure Heterogeneity**.

2. Structure Heterogeneity is conventionally understood as the existence of variance in the representation and description of the same target reality, e.g., of the car, when modeled through different properties by different sources.

3. Structure heterogeneity (but in general all layers of heterogeneity) appears at **three different levels** within the data:
   - Language
   - Knowledge
   - Data

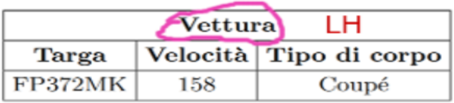
Data Representation - Structure Heterogeneity (SH) - Language

SH in Language (LH) refers to the different levels of abstraction in the concepts employed to describe the same target reality in a language. For example:



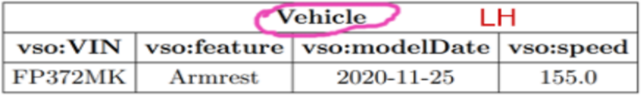
		Car		
Nameplate	schema: speed	schema: fuelCapacity	schema: fuelType	schema: modelDate
FP372MK	150	62	Petrol	2020-11-25

LH



	Vettura	
Targa	Velocità	Tipo di corpo
FP372MK	158	Coupé

LH

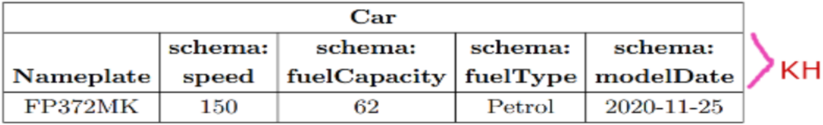


	Vehicle		
vso:VIN	vso:feature	vso:modelDate	vso:speed
FP372MK	Armrest	2020-11-25	155.0

LH

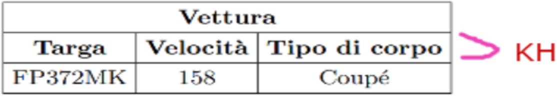
Data Representation - Structure Heterogeneity (SH) - Knowledge

SH in Knowledge (KH) refers to the different (set of) properties employed to describe the conceptualization of the same target reality. For example:



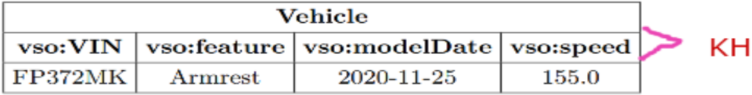
Car				
Nameplate	schema: speed	schema: fuelCapacity	schema: fuelType	schema: modelDate
FP372MK	150	62	Petrol	2020-11-25

KH



Vettura		
Targa	Velocità	Tipo di corpo
FP372MK	158	Coupé

KH

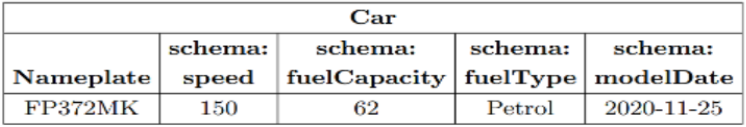


Vehicle			
vso:VIN	vso:feature	vso:modelDate	vso:speed
FP372MK	Armrest	2020-11-25	155.0

KH

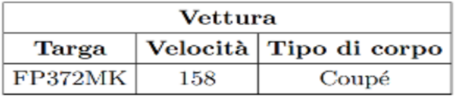
Data Representation - Structure Heterogeneity (SH) - Data

SH in Data (DH) refers to the different (set of) data values (belonging to different data types) employed to describe the conceptualization of the same target reality. For example:



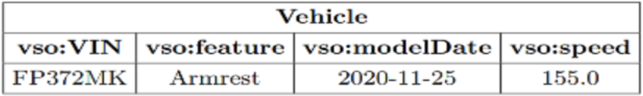
Car				
Nameplate	schema: speed	schema: fuelCapacity	schema: fuelType	schema: modelDate
FP372MK	150	62	Petrol	2020-11-25

DH



Vettura		
Targa	Velocità	Tipo di corpo
FP372MK	158	Coupé

DH



Vehicle			
vso:VIN	vso:feature	vso:modelDate	vso:speed
FP372MK	Armrest	2020-11-25	155.0

DH

Data Representation - Meaning heterogeneity

- Even fixing a source of information from which data is collected and represented through a specific data formats, as well as adopting clear data structures, a final layer of heterogeneity has to be considered.
- **Meaning Heterogeneity**, is defined over the values of the information properties which can be used to identify a real world entity, thus distinguishing one entity from one another.

Data Representation - Meaning heterogeneity

Example: consider the Car entity represented in two different datasets A, and B.

Car in dataset A:

- Vehicle-ID: 1234
- Manufacturer: "Renault"
- Engine-type: "Electric engine"
- Fuel-type: "Electricity"

Car in dataset B:

- Vehicle-ID: ABCD
- Manufacturer: "Renault"
- Engine-type: "Electric engine"
- Fuel-type: "Electricity"

From the same source, we have two datasets in the same format, using the same structure of information. Nevertheless ..

- how can we know if the two car are the same entity or different ones ?
- is the identifier in dataset A equivalent to the identifier in dataset B ?
- the "Manufacturer" term in datasets A has the same meaning of "Manufacturer" in dataset B ?

Part 1.3
Data reuse processes

1 Information & information reuse

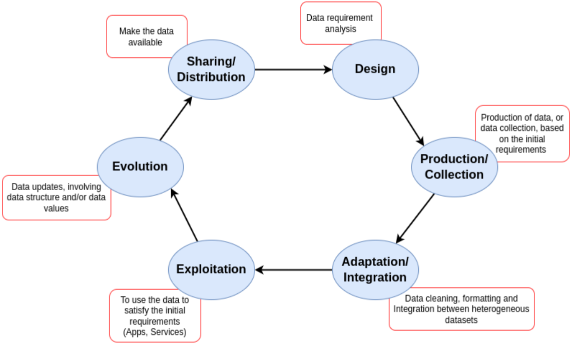
2 Data representation

3 Data reuse processes

4 Data architecture for reuse

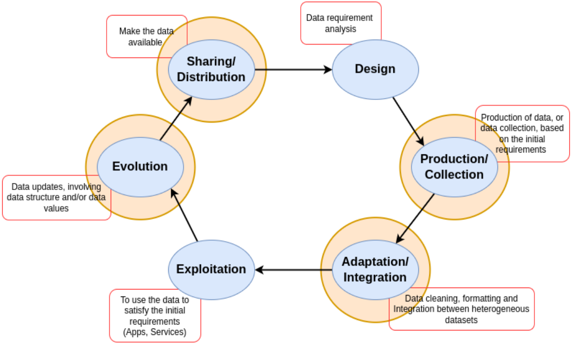
Data Reuse Processes

- The reuse of data, is not only a matter of representing data, it involves also **the processes** required **to get, exploit**, and **make reusable** such data.
- To understand the role of such processes in data reuse, we can see them into **the data life cycle**.


Make the data available
Data requirement analysis
Sharing/ Distribution
Design
Production of data, or data collection, based on the initial requirements
Evolution
Production/ Collection
Data updates, involving data structure and/or data values
Exploitation
Adaptation/ Integration
To use the data to satisfy the initial requirements (Apps, Services)
Data cleaning, formatting and Integration between heterogeneous datasets

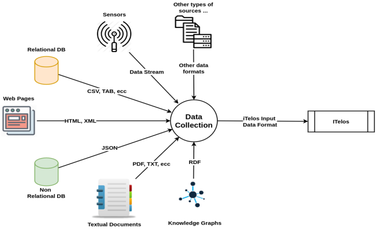
Data Reuse Processes

- Data life cycle activities most involved in data reuse.


Make the data available
Sharing/ Distribution
Data requirement analysis
Design
Production of data, or data collection, based on the initial requirements
Production/ Collection
Evolution
Data updates, involving data structure and/or data values
Exploitation
To use the data to satisfy the initial requirements (Apps, Services)
Adaptation/ Integration
Data cleaning, formatting and Integration between heterogeneous datasets

Data Reuse Processes - Data Collection

- The data collection activity requires processes for the extraction (scraping) of data from data sources.
- Such processes need to be defined and implemented considering the **source heterogeneity**.
    - Potentially each data source requires a dedicated implementation of data collection processes, thus increasing the effort to be paid for the whole data life cycle.



Data Reuse Processes - Data Production

- The production of data does not affect the **reuse of existing data**.
- Nevertheless, it plays a very crucial role in **the reuse of new data**, being th activity responsible of the creation of data which can be potentially reused.
- Data production processes that do not consider reusability and interoperability of data, increase the overall cost of the data life cycle.

Data Reuse Processes - Data Adaptation/Integration

- **Data adaptation**: activity which aims at cleaning and formatting (format heterogeneity) the data to be exploited for a specific purpose.
- **Data integration**: activity which aims at integrate together different datasets to obtain a merged information resource (structure heterogeneity), able to satisfy a specific purpose.

The KGE course is strongly focused on Data Integration processes.

Data Reuse Processes - Data Adaptation/Integration

- The adaptation and integration processes are fundamental in the reuse of data, mainly for two reasons:

  1. (input side) The efficiency of such processes has a strong impacts over the data life cycle.
     - **cleaning**: how much the a reusable datasets can be cleaned out from noise, respect to a specific purpose to be satisfied ?
     - **formatting**: which, and how many, standards the process is able to apply to the dataset to be formatted ?
     - **integration**: how much the integration process is able to deal with Data Heterogeneity ?
  2. (output side) The way the data are cleaned, formatted and integrated, strongly affects their future reusability.

Data Reuse Processes - Data Evolution

- **Is a data able to scale up ?**
    - How much effort is required **to extend the data produced/collected and adapted/integrated**, in order to satisfy new feature, respect the data initial purpose ?
        - evolution at schema level (schema update);
        - evolution at data level (data values update). For example, data expiration
- Low quality data evolution processes can increase the cost to be paid in the data life cycle.

Data Reuse Processes - Data Sharing

- As already anticipated by discussing about previous data reuse processes, **the reuse is not only a matter of getting existing data in input**.
- The reuse of data strongly involves the process for **data sharing** (or data distribution).
- Low quality data sharing processes introduce difficulties for the retrieval of the data, thus **limiting its potential future reuse**.

Part 1.4
Data architecture for reuse

1 Information & information reuse

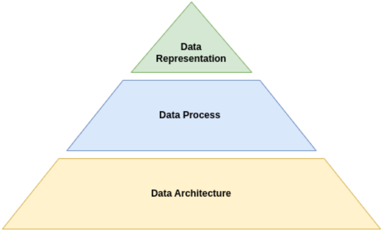
2 Data representation

3 Data reuse processes

4 Data architecture for reuse

Data Architecture for Reuse

- The data needs to be represented properly to handle its heterogeneity, and
- processes are required to implement the reuse of data.
- Moreover, to fully address the data reuse problem, we need to consider the environment in which such a data can be properly represented, and the reuse processes correctly supported.
- Such environment is defined by the **architecture** (or infrastructure) **enabling the reuse of data**.



Data Architecture for Reuse

The fundamental requirements required for such architectures are:

- **Data collection support**: components and services for the upload of reusable data (Fundamental to support the data collection processes).
- **Data store support**: storage components and services.
- **Data elaboration support**: component and services supporting the adaptation, integration and evolution processes.
- **Data distribution support**: component and services supporting the data sharing processes.

The lack of data architectures where the above requirements are not considered, and/or not well composed together, limits the possibilities of data reuse.