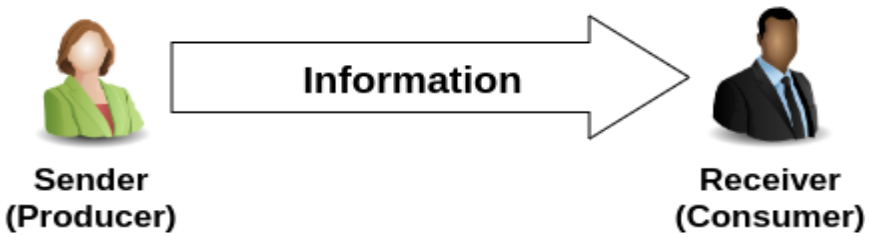
Part 1
The Reuse Problem

Part 1.1
The data representation problem

Information

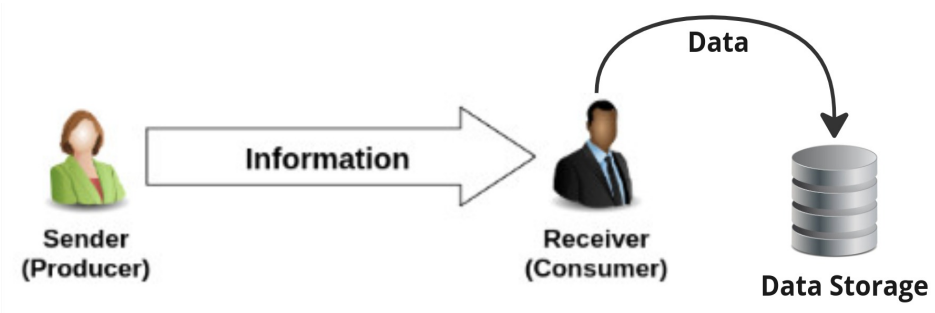
The information is stimuli (i.e., electromagnetic waves), created by a sender, **that has meaning in some context** for its receiver.



- Notice how a communication can be the creation of data (Producer) to be exploited by any kind of data service (Consumer).

Information & Data

The information is **perceived** by the receiver who transforms it into **data**.



- The data becomes exploitable once it is **stored in the receiver's memory**.
- Notice how the perception of information and the subsequent transformation into data, **depends on the receiver**, not on the sender who sent the information.
- The "point of view" or more formally, **The Purpose** of a user always affect the **perception** of the information and the **representation** of such information as data.
- The data user purpose gives the **semantic meaning** to the data.

Information & Data Reuse

- The information in a communication **is not always new**, most of the time instead is **reused** from previous communication.
  - In other words, the data obtained from a previous communication has been processed and used into a new one, **with a different purpose**.
- As a consequence, the reuse of information, thus the reuse of data, is a matter of **perceiving and representing, with a new purpose, a data created with someone's else purpose.**

The Representation Problem

- Reusable data is represented, and available, in a (data) **world that is, apparently, disordered**.
  - "disorder" := high level of **heterogeneity** (low quality data) requiring:
    - high effort in find the required information;
    - high effort in find the right data (representation);
    - huge amount of data processing.
- The data appears in multiple forms, highlighting what is called the **data heterogeneity**.
- The problem caused by the heterogeneity is, **to choose** the right representation of the data to be (re)used.

The Representation Problem - Example (1)

The same information can be represented in different forms.

- **Type of data source**: photo, video, text, table, hierarchical, ...

Assuming a fixed data source type, **different format** can be adopted.

- **photo**: jpg, png, svg, ...
- **video**: mp4, avi, flv, ...
- **text**: doc, docx, PDF, ...
- **table**: csv, tsv, xls, ...
- **hierarchical**: json, xml, rdf, ...

The Representation Problem - Example (2)

Assuming a fixed data format, **different languages** can be used to represent the information.

- Persona(Italian), Person(both English and German), Personne(French)

Assuming a fixed language (English), **different attributes** can be used to represent the information.

- Person =: <Name, dateOfBirth>
- Person =: <Surname, gender>

Assuming a fixed attribute set, **different values** can be used to represent the attribute values.

- dateOfBirth =: "July 22, 2024"
- dateOfBirth =: "22/07/2024"
- dateOfBirth =: 1721657941

Part 1.2
The data heterogeneity

The Data Heterogeneity

- The four data heterogeneity layers:
  1. Information
  2. Data Language
  3. Data Schema
  4. Data Value

The Data Heterogeneity - Information layer

- This kind of data heterogeneity is related **the way the data is provided by a data sources**.
  - Multiple types of source.
  - Multiple types of formats.
- **The heterogeneity, at the information layer, constraints the capacity of processing the data, and thus its reuse.**

The Data Heterogeneity - Data language layer

- This kind of data heterogeneity is related **the multiple languages which can be used to express an information** into the data.
- Different languages could use different words/terms to express the same concept (maybe more then a single words is used).
- **Not always a concept expressed in a language can be expressed in another one.**
- The heterogeneity, at language layer, has a strong impact on:
  - the **understanding of the information** carried by the data;
  - the **interoperability** level **of the data** (cross-lingual applications);
  - the **reuse** level **of the data** (translation effort).

The Data Heterogeneity - Data schema layer

- This kind of data heterogeneity highlights **multiple ways in which a real-world entity can be described by a set of attributes** (or properties), called schema.
  - **Entity** (top level def.): A data entity is a representation of a set of information about a given concept. Each entity is defined **by its own attributes**, and attributes values.
- Different cultures, geographical contexts, even different single persons can define **the same entity with a different attribute set**, depending by their "point of view".
- **The heterogeneity, at data schema layer, is the root of the problem of identifying whether or not two entity representations describe the same entity.**

The Data Heterogeneity - Data value layer

- This kind of data heterogeneity highlights **multiple ways in which a real-world entity can be described by the values of its attributes.**
  - **Entity** (top level def.): A data entity is a representation of a set of information about a given concept. Each entity is defined by its own attributes, **and attributes values**.
- As for the above heterogeneity layer, the data producer **"point of view"** defines which are the data types used to represent an entity attribute's value.
- The heterogeneity, at data value layer, **increases the complexity of the entity identification process**, which aim to understand if two entities (even if with the same attribute set) are actually two representation of the same real-world object.

The Data Heterogeneity - Data reuse costs

- Each data heterogeneity layer **introduces its own cost** to be paid while dealing with data reuse.
- Such a cost is defined in terms of **data identification and data processing** required when the data has to be reused.
- For this reason, all the four different layers have to be considered, while reusing data, trying to minimize the overall data reuse cost.

The Data Heterogeneity - Bug or Feature ?

- It seems that the data heterogeneity is the main problem to be addressed while we need to reuse data. **A kind of bug on data representation.**
- Nevertheless, the heterogeneity is an **ineradicable** characteristic of the data.

Are we sure that the heterogeneity is a bug ?

Part 1.3
The data diversity

Data Diversity

- Let's forget for a while the cost of data reuse, and let's concentrate on the data representation only.
- The data heterogeneity expresses, at different levels, the **local features of information** implicitly represented in the data.
  - **"information feature"** := specific portions of information properly represented through the heterogeneity.
  - **"local"** := such information features reflects the "point of view" of the data producer, as well as of the geographical and social context in which the data has been created.

Such a local information features is called **Data Diversity**.

Data Diversity Layers

- Unlike the heterogeneity (seen as bug causing expensive data processing), the data diversity is information which makes the data **richer**, more **useful**, and more **reusable**.
- The main difference (and the major benefit) is that the diversity is **usable information** represented into concrete (diversity-aware) data.
- Like the heterogeneity, the diversity appears at **multiple layers**.
  - Information diversity
  - Data language diversity
  - Data schema diversity
  - Data value diversity

Information Diversity

- As already discussed Information heterogeneity refers to the different data source, and data formats, from which data can be collected.
- The information diversity aims at "homogenizing" the heterogeneity by **selecting reliable and standardized sources, and data formats**, thus **limiting noisy interpretation of the target reality** respect to a specific purpose ("point of view").
- Concretely, Information diversity is the **set of data sources** from which the data can be collected (and reused).
  - Media
  - Websites.
  - Data catalogs.
  - Databases and Knowledge bases.
  - Others.
- Moreover, for each data source, **standard data formats** are fixed to improve the data interoperability and data reuse.

Information Diversity - Example

- Working with geographical data means to deal with several different data sources (heterogeneity information layer) from which data can be collected.
  - Google maps; GPS mobile sensors; purpose-specific (i.e., transportation, environment, ...) web portals; others.
- One of the most important source of geographical data is OpenStreetMap (OSM). For this reason it is recognized as a **standard data sources for the geographical data domain, providing data in standard formats.**
- OSM with its data formats, plus other standardized data sources, **defines the Information diversity in the geographical data domain**.
- Nevertheless, **highly specific data**, produced with specific tools, **could not be** provided by **standard data source** and/or represent through **standard formats**.
  - In such a case **the information diversity preserves the local data diversity** by maintaining the original source and format (which could become the reference standard in future).

Data Language Diversity

- Heterogeneity at language layer underlines how **multiple languages can be used to express the same concept**, used then to represent data. Therefore, causing ambiguity to be handled during a data reuse process.
- The data language diversity is referred to the **set of terms (or words) which define a terminology (vocabulary) for a specific domain, context or even personal "point of view".**
- Such terms are isolated into **language dataset**, where the language diversity is concretely defined.
  - In a language dataset each term is associated to: an **identifier**, a specific **word** (which can be the term itself or a synonym), and a textual **description** of its meaning.
- It clearly appears how the data diversity, at language level, **becomes exploitable through the language datasets**.

Data Language Diversity - Example (1)

- Assume two projects requiring data for the development of digital services in two different domains; project A is focused on transportation domain, while project B focused on tourism.
- **Case 1: a single term can be used in both the two data domains to indicate two different concepts**.
  - In project A the term "user" means a bus or train passenger.
  - In project B the term "user" means a visitor/tourist.
- The language diversity in this case is represented by **the specific meanings of the same terms into the different domains**.
- To make such a diversity processable, the two projects generate two language datasets where the term "user" is described through different meaning, and associated to other terms like "passenger" and "tourist" for project A and B, respectively.

Data Language Diversity - Example (2)

- Assume two projects requiring data for the development of digital services in two different countries, with two different languages; project A use English defined data, while project B use Italian defined data.

- **Case 2: a single concept is represent by two different words one in English and the other in Italian**.
  - In project A exist the term "Bear".
  - In project B exist the term "Orso".

- The language diversity in this case is represented by **the specific terms adopted locally to represent the same concept**.

- To make such a diversity processable, the two projects generate two language datasets where the term "Bear" is described through its meaning, so that it can be associated to the Italian term "Orso" (in the other dataset) which is in turn described with the same meaning.

Data Schema Diversity

- Heterogeneity at data schema layer underlines how **multiple set of attributes can be used to describe a real-world entity**. Therefore, causing ambiguity to be handled during a data reuse process.
- The data schema diversity is referred to the **set of attributes used to define an entity for a specific domain, context or even personal "point of view".**
- The data schema diversity is concretely defined by (and become exploitable through) a **schema dataset**, indicating the set of attributes used to represent the real-world entities.
  - In other words, a schema dataset defines how the **Entity Types** (or EType) are modelled.
- The explicit representation of the data schema allows for **the concrete exploitation of the data schema diversity**, as a feature of the data to be reused.

Data Schema Diversity - Example

- Two projects, requiring data to develop digital services, operate in two different domain. The domain of project A is car manufacturing while project B is focused on car sale.
- Project A and B recognize the entity "car" in two different ways.
  - Project A - Car := <chassis number, model, manufacturer>
  - Project B - Car := <color, price, model>
- **The data diversity is concretely represented by the two data schema**, created for the data to be used by the two projects respectively, where the above data schema (for the car entity) are explicitly defined, thus becoming exploitable.

Data Value Diversity

- Heterogeneity at data value layer underlines how **multiple set of attribute values can be used to describe a real-world entity**. Therefore, causing ambiguity to be handled during a data reuse process.
- The data value diversity is referred to the **set of attribute values used to define an entity for a specific domain, context or even personal "point of view".**
- The data value diversity aims at "homogenizing" the heterogeneity **by selecting reliable and standardized data types**, to limit noisy interpretations of the target reality respect to a specific domain, context or personal "point of view".

Data Value Diversity - 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 language (same concepts), and same data schema. 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?

Data Value Diversity - Example (cont'd)

- Consider the Car entity represented in two different datasets A, and B.

Car in dataset A:
- Vehicle-ID: <**Integer**>
- Manufacturer: "Renault"
- Engine-type: "Electric engine"
- Fuel-type: "Electricity"

Car in dataset B:
- Vehicle-ID: <**Integer**>
- Manufacturer: "Renault"
- Engine-type: "Electric engine"
- Fuel-type: "Electricity"

- The data value heterogeneity defines the standard data type to be used to define identifiers in the automotive domain, to avoid the above ambiguity.

- Nevertheless, **highly specific data could not adopt standard data types**.
  - In such a case **the data value diversity preserves the local data diversity** by maintaining the original data type (which could become the reference standard in future).

Data Diversity

- The definition of data diversity at each level makes possible to exploit that portion of information that before was only implicitly represented into the data.
  - **Information diversity** -> standard set of data source and formats.
  - **Language diversity** -> language datasets (concepts definition).
  - **Data schema diversity** -> schema dataset (EType definition).
  - **Data value diversity** -> standard set of data types.