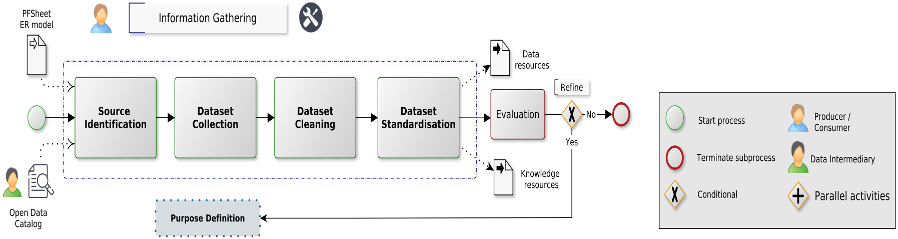
# Part 5.4
# Phase 2 - Information Gathering

1. KG Construction
2. iTelos
3. Phase 1 - Purpose Definition
4. **Phase 2 - Information Gathering**
5. Phase 3 - Language Definition
6. Phase 4 - Knowledge Definition
7. Phase 5 - Entity Definition

## Phase 2 - Information Gathering

- Phase 2 - Objective
- Phase 2 - Activities
- Personal context data collection

# Phase 2 - Information Gathering



- **Input**: Purpose Formalization sheet, ER model, Source list.
- **Objective**: collecting the resources, to be processed, to build the final EG, thus satisfying the formalized purpose.
- **Output**: Standardised datasets (Language, Knowledge and Data values).

# Information Gathering - Objective

- The second iTelos phase is focused on the collection of those resources required for building the final EG, accordingly with the Purpose detailed in the previous phase.

- **Note** - The ”resources” mentioned above include:
  - Data value datasets
  - Knowledge datasets (ontologies)
  - Language datasets

- Therefore, the whole phase with its activities can be applied to each information layer.

- The Information Gathering phase is not only responsible for the collection of such datasets, it aims also to improve their quality and reusability.
  - cleaning datasets noise
  - adopting well-known standards

# Information Gathering - Source Identification (1)

- The first activity of the current phase, aims at **identifying and accessing the sources of information** provided as input, and, eventually discovering other sources (if those in input are not sufficient).

- Clearly this activity is where the data reuse appears more evident. The user is literally looking for existing data sources providing the information she needs.

- As already discussed in the previous lecture, due to the data heterogeneity at information (data source) level, there are multiple types of data sources to be considered.
  - **"High quality" sources**: the sources considered are Catalogs where interoperable and reusable datasets (produced by an iTelos process) are distributed (i.e. LiveData catalogs, where the iTelos resources are published).

# Information Gathering - Source Identification (2)

- As already discussed in the previous lecture, due to the data heterogeneity at information (data source) level, there are multiple types of data sources to be considered.
  - **"Low quality" sources**: the sources considered can be different and distributing different type of data. The data distributed by this sources are less understandable (low quality metadata) and less interoperable (non standard, non diversity-aware).
    - Low quality data catalogs
    - Web pages
    - Purpose specific Databases
    - ...
  - Example: Open Data Trentino
    - Look into the catalog and into the resources.
    - Are they really useful ?

# Information Gathering - Dataset Collection (1)

- Depending by which source is considered, the collection of datasets requires less or more effort. **High-quality** data sources have:
  - clear policies of data distribution;
  - direct distribution/download;
  - or distribution on demand through request to the data owner (easy communication with data owners).
  - High quality metadata, describing the available resources. This increase the findability ad accessibility of the resources distributed.

# Information Gathering - Dataset Collection (2)

- Depending by which source is considered, the collection of datasets requires less or more effort. With **Low-quality** data sources:
  - the data is not always distributed, sometimes it is only **published** or **visualized** online.
  - It is not always clear how to concrete access the datasets. Sometimes the source has to be processed to get the data (i.e., Scraping webpages).
    - asking for datasets directly to owners (if the owner contacts are indicated);
    - accessing data through automatic or semi automatic portals/API;
    - scraping data from sources (this usually requires scraping libraries customization);
    - producing our own data (data collection apps and tools [iLog] [38] ).

[38] see last part of this lecture

# Information Gathering - Dataset Collection (3)

- Collecting data, in general, aims to achieve the following two results:
  - Increase the number of **entities** and/or entity types.
  - Increase the number of **entity attributes** and/or entity type properties.

- Are the resources collected covering your list of CQs ?
  - **yes** - let's proceed on.
  - **no** - go back to source identification.

# Information Gathering - Dataset Cleaning

- The cleaning activity aims to **remove ”noise”** from the set of resources collected.
- ”Noise” is intended to be:
  - entire **datasets** (could be downloaded as a bunch from a source) without any information required to satisfy the Purpose.
    - It happens often collecting data automatically or receiving huge amount of data.
  - **Etypes and/or entities**, within a single dataset, with no relevance for the Purpose.
  - **Properties and/or attribute**, within single datasets, with no relevance for the Purpose.
  - Any of the above, having **NULL** value.

# Information Gathering - Dataset Standardisation (1)

- Now the set of datasets has been finalized.
- The Standardisation activity aims to:
  - **Align the differ formats** present in the heterogeneous dataset (datasets formats and data types), by adopting reusable and interoperable standards.
    - **Example**: For a tabular datasets, the Microsoft Excel format could be ok, but it is a proprietary formats. CSV, being an open format, is more reusable (i.e., for the user that do not use Microsoft apps).
  - **Anonymise the data** collected; required only if sensible information (like personal data) are included in the datasets collected.

**Note**: the format alignment over common standards (CSV, XML, TSV, JSON, RDF and OWL) is strongly required, mainly for two reasons:

- Reusability, Interoperability.
- Compliance with iTelos process.

# Information Gathering - Dataset Standardisation (2)

The best practices for the generation of highly reusable and interoperable data are provided by:

- FAIR Principles
- 5-Stars Open Data

# Personal context data collection

- Personal context data requires a particular procedures and instruments.
  - **Sensitive Person-centric data**
- An example of person-centric data collection is the **iLog application**, and the methodology for it has been developed [39].

[39] for lack of time this methodology will not be discussed in this course.

# What is iLog

- iLog is a mobile application developed by the Knowdive group @ UniTn
- it collects data to study **human behavior** on a large group of participants in the wild
- Allow the collection of **sensor data** from participant mobile devices.
  - Supports a wide range of sensors, both physical and software.
- Allow the collection of **contextual data** using questionnaires.
  - They can be used to infer the behaviour of participants.
  - They are inspired and can be used to implement the experience sampling method (ESM) approach.

https://datascientia.eu/ilog/

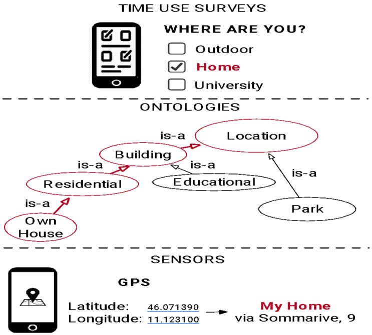
Available on the App Store

GET IT ON Google Play

Annotations

Knowledge representation

Sensor (+ ML)



## Active data *vs.* passive data

- The person can be systematically observed in their daily life using passive or active measurements.

### Active data

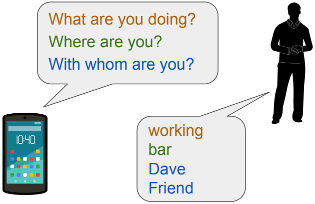
- is the result of person's reaction to a stimulus
- *e.g.*, answer to a open or closed questions
- affected by Hawthorne effect, respondent burden, non-attitudes conditions,...

### Passive data

- collected without direct interaction with the person
- *e.g.*, smartphone sensors

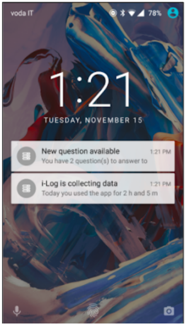
# Active data: time diaries

- measures the frequency and duration of human activities, behaviours, and experiences, offering a detailed view of social behaviour.
- **self-answered** questions asked at fixed intervals
- gets people's **perspectives** on their daily lives
- question frequency varies from a few times in a day to every 10 minutes (it depends on the study duration)



- Matteo Busso et al. "The iLog methodology for fostering valid and reliable Big Thick Data". In: (2024).
- Tamlin S Conner and Barbara J Lehman. "Getting started: Launching a study in daily life.". In: (2012).

# Time diaries in iLog



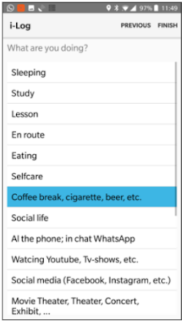
voda IT 78%

1:21

TUESDAY, NOVEMBER 15

New question available 1:21 PM
You have 2 question(s) to answer to

i-Log is collecting data 1:21 PM
Today you used the app for 2 h and 5 m

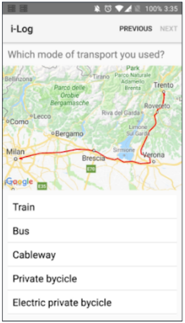


97% 11:49

i-Log PREVIOUS FINISH

What are you doing?

- Sleeping
- Study
- Lesson
- En route
- Eating
- Selfcare
- Coffee break, cigarette, beer, etc.
- Social life
- Al the phone; in chat WhatsApp
- Watcing Youtube, Tv-shows, etc.
- Social media (Facebook, Instagram, etc.)
- Movie Theater, Theater, Concert, Exhibit, ...

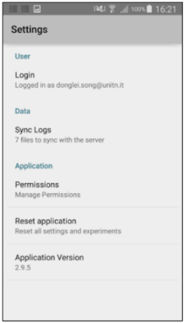


100% 3:35

i-Log PREVIOUS NEXT

Which mode of transport you used?

- Train
- Bus
- Cableway
- Private bycicle
- Electric private bycicle



100% 16:21

Settings

User

Login
Logged in as donglei.song@unitn.it

Data

Sync Logs
7 files to sync with the server

Application

Permissions
Manage Permissions

Reset application
Reset all settings and experiments

Application Version
2.9.5

The user can also be asked to take pictures and videos and share his/her location.

# Passive data: sensor data

- data collected in the background with no user intervention
- the data are generated as time-series, consisting of tuples composed of a timestamp and one or more values

## Hardware

- physical sensors of the device: detect and respond to physical environment

- they generate **low-level** information, usually at a very high frequency

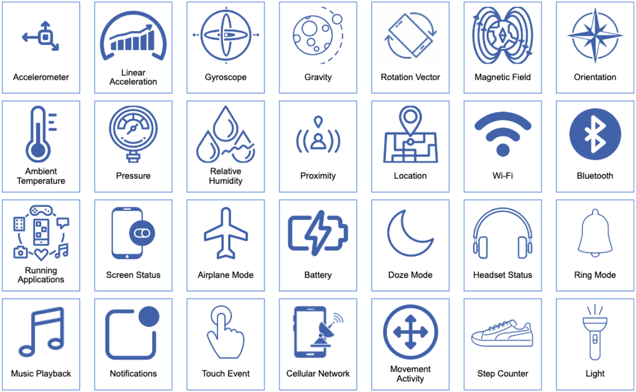
## Software

- software component collecting events from the operating system and software

- generally they generate a **higher level** information

# Passive data: iLog sensors

- Connectivity
- Environment
- Motion
- Position
- App-usage
- Device-usage



## Connectivity

Sensors that report connections with other devices.

1. *Bluetooth Devices* returns all Bluetooth devices detected by the phone;
2. *Bluetooth Low Energy* lists the discovered low-energy Bluetooth devices in the local area;
3. *Cellular Network info* returns information related to the cellular network (cellid, dbm, type) to which the phone is connected to;
4. *WiFi Network connected to* describes the WiFi network to which the phone is connected to;
5. *WiFi Networks available* returns all WiFi networks detected by the smartphone.

## Environment

Sensors collecting data about the environment are hardware-based, and thus their availability depends on the device model.

1. *Light* measures the ambient illumination around the phone, measured in illuminance (lux);
2. *Pressure* measures the ambient air pressure to which the phone is subjected.

# Motion

Motion sensors detect device movements along the three coordinate axes.

1. *Accelerometer* measures the acceleration to which the phone is subjected and it captures it as a 3D vector;
2. *Gyroscope* measures the rotational forces to which the phone is subjected and it captures it as a 3D vector;
3. *Movement activity label* identifies the activity performed by the user.
4. *Step Counter* uses the Android API to measure the number of steps made by the user (while carrying the phone) since the phone was turned on;
5. *Step Detection* similar to the previous, uses the Android API to generate a step value each time the user takes a step.

## Position

Position sensors determine the device's physical position.

1. *Location* returns the geographical coordinates of the phone. To increase the accuracy, this sensor combines GPS and WIFI/cellular connections;
2. *Magnetic Field* measures the magnetic field to which the phone is subjected along the three coordinate axes;
3. *Proximity* measures the distance between the user's head and the phone. Depending on the phone it may be measured in centimetres (i.e., the absolute distance) or as labels (e.g., 'near', 'far').

## App usage

These sensors identify how people use and interact with social media and applications.

1. *Headset status* returns whether the headphones of the phone are connected;
2. *Music Playback* returns whether music is being played on the phone (yes or no) using the default music player from the operating system;
3. *Notifications received* measures when the phone receives a notification and when it is dismissed by the user;
4. *Running Applications* reports the name of the application (or application package) currently running in the foreground of the phone.

## Device usage

These sensors identify how people use and interact with their devices.

1. *Airplane Mode* returns whether the phone's Airplane mode is on or off, Airplane mode turns off all the connectivity features of the phone;
2. *Battery Charge* returns whether the phone is currently charging its battery;
3. *Battery Level* returns the phone's battery level;
4. *Doze Mode* returns whether the phone's doze mode is on or off. Doze mode is a low battery consumption state in which the phone enters after some time of not being used;
5. *Ring Mode* returns the current ring status of the phone (normal/silent/vibrate);
6. *Touch event* generates a touch value each time the user touches the screen;
7. *Screen status* returns whether the phone's screen is on or off;
8. *User Presence* detects when the user is present near the phone, for example when the user unlocks the screen.

# WeNet case study

- Observe the **diversity** in the **everyday life** of students considering the main aspects of the context and their articulation over time in terms of **routines** and **social practices**
- Researchers from **4 disciplinary fields** (Sociology, Psychology, Computer Science, Design)
- Students from **8 universities** (AAU, AMRITA, JLU, LSE, NUM, UC, UNITN)
- Results:
  - 1 month of data
  - 21073 participants to first questionnaire, 757 with sensor data
  - multi-purpose dataset