The iLog methodology for fostering valid and reliable Big Thick Data

Matteo Busso

Department of Information Engineering and Computer Science

University of Trento

A thesis submitted for the degree of

Doctor of Philosophy

January 2024

Acknowledgements

“— Vous êtes vraiment. . . Vraiment, vous êtes. . .
Sa main libre fait un geste d’impuissance heureuse. Il hoche affectueusement la tête.
L’intensité des projecteurs baisse. Peu à peu la lumière se fait chaude, presque intime.
— Je ne sais pas comment vous...”

Daniel Pennac *Merci*

Expressing gratitude is always a significant moment. As I always say, this thesis, like any creation of human endeavor, is simply the result of serendipitous encounters.

Therefore, it is with pleasure that I thank my advisor, Fausto Giunchiglia, for the guidance, all the opportunities of these years, and for the "frittella times". I thank Ivano Bison, without whom this unbelievable world of computer scientists and European projects would never have been revealed. And, speaking of European projects, I want to thank all the members of WeNet for the inspiration from their "deep" diversity. Special thanks to Amalia de Götzen for helping me discover the *Thick* side of IT.

Thanks also to the reviewers, Prof. Jahna Otterbacher and Prof. Kobi Gal, whose valuable and insightful comments helped me refine the style and contents of this thesis.

A special thank you to my colleagues and friends at KnowDive, who helped me during my PhD and revealed to me the human side of IT. And thanks to Bocca, who, in a short time, has become one of my lifelong friends (No, Grazie a te!).

Thanks to William and Mara, to whom I owe a lot of who I am, and thanks to the rest of the Turin daredevils.

Thanks to my family, thanks to Simone, Agnese, and Andrea.

And last but not least, thanks to Giulia Gaggero, for all the time she unraveled my tangled thoughts.

Abstract

Nowadays, the apparent promise of Big Data is that of being able to understand in real-time people's behavior in their daily lives. However, as big as these data are, many useful variables describing the person's context (e.g., where she is, with whom she is, what she is doing, and her feelings and emotions) are still unavailable. Therefore, people are, at best, thinly described. A former solution is to collect Big Thick Data via blending techniques, combining sensor data sources with high-quality ethnographic data, to generate a dense representation of the person's context. As attractive as the proposal is, the approach is difficult to integrate into research paradigms dealing with Big Data, given the high cost of data collection, integration, and the expertise needed to manage them.

Starting from a quantified approach to Big Thick Data, based on the notion of situational context, this thesis proposes a methodology, to design, collect, and prepare reliable and valid quantified Big Thick Data for the purposes of their reuse. Furthermore, the methodology is supported by a set of services to foster its replicability.

The methodology has been applied in 4 case studies involving many domain experts and 10,000+ participants from 10 countries. The diverse applications of the methodology and the reuse of the data for multiple applications demonstrate its inner validity and reliability.

Contents

Appendices

List of Figures

List of Tables

List Acronym and Definitions

Annotation . . . . It is a label associated with data coming from sensors to enrich their content and facilitate their design, validation, and interaction with the user. They are typical of Human Activity Recognition (HAR) approaches and are usually produced by experts who annotate sensor databases. One of the main contributions of this thesis is to extend the notion of annotation, integrating it with the sociological one of Time Diaries, ultimately favoring the in-the-wild labeling of sensors produced directly by the user (the leading expert of her own context).

Big Data . . . . . Generally, it refers to datasets that are too large or complex to handle with traditional approaches and software. In our case, the meaning of Big Data refers to all that data currently produced that surrounds the person, starting from the streams of social media information (from posts to photos to messages), passing through the whole of sensors to observe and measure human behavior (e.g., smartphone GPS) up to data from phone applications, streaming and e-commerce platforms and much more. With Bornakke, the main argument of this thesis is that these data do not allow us to fully recognize the person as immersed in his context, both due to their poor quality and because they do not consider some salient aspects to which, as people (and as social scientists) we are used to taking into consideration, that is, all the Thick characteristics that allow us to understand the meanings of our context.

BPMN . . . . Business Process Model and Notation (BPMN) is a graph that describes processes (similar to the activity diagram) designed to be intuitive and facilitate the management of business processes and their understanding even by non-experts. Therefore, it is useful for the purpose of this thesis, as it is designed to facilitate a broad and interdisciplinary audience in the execution of the methodology and in the use of the services. BPMN is made up of several elements, briefly described here for the reader (for an extended discussion see [1] or the relevant Wikipedia page). In particular, the BPMN is composed of:

- *Flow objects*, namely Events (◯), which denotes something that happens (for example, the beginning or end of an activity or process); Activities (□), which describes the action or work that needs to be done; and Gateways (◇), which indicate whether the process path forks or merges.

- *Connecting objects* (⟶), namely Sequence flow, or a solid line with an arrow that describes the order in which the actions are performed; Message flow, i.e. a dotted line with an arrow that describes the communications between different swim lanes; and Association, a dotted line that describes the relationships between artifacts and activities

- *Swim lanes* (=), namely Pool, a solid rectangle that contains all the activities of an organization and can be composed of one or more Lane(s) that organizes the activities according to a function or role.

- *Artifacts*, which add information to make the BPMN more readable, for example by indicating whether data, documents or external objects (e.g., a data storage unit) relevant to the process are included.

Context . . . . . . It is everything that surrounds the person in her daily life and can be represented from the person's point of view and other sources of information. In this thesis, we start from the assumption that the person is the best expert of their own context as they are immersed in it and can describe it. Starting from any time and space, the person can generally say where she was, what she was doing, who she was with, and how she felt. However, it is also possible to enrich the context information through other sources, such as sensors and databases relevant to the observed situation. The interaction between these two sources of information over time defines the person's situational context. It is understood here as a way to generate Big Thick Data quantitatively.

Cross-sectional S. Cross-sectional survey is a type of observational study that collects data from a population, or a representative subset, at a specific point in time. Usually, these studies are based on questionnaires and aim at the generalizability of the observations to an entire population.

ESM . . . . The experience sampling method (ESM), also referred to as a daily diary method or ecological momentary assessment

(EMA), is an intensive longitudinal research methodology. It is based on stimuli, like questions and psychometric scales, asking participants about their thoughts, feelings, behaviors, and environment on multiple occasions over time. Unlike cross-sectional or longitudinal investigations, the ESM approach allows us to observe the person's behaviors in progress (e.g., at what moments of the day a person experiences a particular emotion or how a therapy proceeds) and with greater granularity, as they are based on everyday life. Contrary to our approach, ESM does not consider the data coming from the sensors (except for GPS) and the possible interactions with them. For this reason, the ESM approach does not consider multiple aspects of daily life and how these interact with the person, nor does it have the possibility of lasting over time due to the high number of stimuli that should be used, increasing the respondent burden. Since ESM is only one of the possible data collection approaches, the methodology refers to the surveys with the broader term "intensive longitudinal survey".

Experiment . . . . An experiment is a study involving two groups to which the participants are randomly assigned, and there is an active manipulation of the input variable (also called independent variable or "feature" in machine learning fields). A similar form known in the field of user experience is A/B testing, which consists of an experiment that usually involves two variants or more (A and B) of the same variable (e.g., two different versions of the same webpage) to determine which of the variants is more effective. In this case, the random assignment of the participants to one of the groups is not considered; that's why they can also be defined as quasi-experiments. If the experiment is conducted in a controlled environment, it is called a laboratory experiment, while in real-life settings, it is called a field experiment. Since the experiment is only one possible approach, we prefer to refer to data collection research with the broader term "study".

GDPR . . . . Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons concerning the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).

HAR . . . . Human Activity Recognition

HITL . . . . Human In The Loop

iLog iLog is an application developed by the KnowDive group at the Department of Information Engineering and Computer Science of the University of Trento. It is designed to collect data for research purposes. With the user's consent, iLog collects data from the smartphone's internal sensors and sends context-sensitive questions. The final goal is to study the users' habits, allowing them to react accordingly with personalized services and generating research datasets for further studies.

Intensive L. S. An Intensive longitudinal survey involves repeated measurements (e.g., daily questions) taken on individuals, encompassing all types of data collection methodologies, such as Time Diary and experience sampling studies (ESM).

KG Knowledge Graph

Longitudinal S. Longitudinal survey consists of cross-sectional surveys administered several times over a period (months or years) to the same set of individuals, aiming to observe changes in behavior, opinions, or attitudes.

Sensors A sensor is any device that autonomously detects a physical phenomenon and returns a signal as output. In our case, we strictly refer to all the sensors surrounding a person's daily life and relationship with the environment. Therefore, the prominent examples of sensors are all those relating to the smartphone (and collected by iLog), as it is the ubiquitous device par excellence. These sensors allow you to observe physical phenomena such as movement (e.g., GPS, accelerometer), atmospheric and environmental conditions (e.g., temperature), but also the person's relationship with their device (e.g., number and type of applications used, screen touch events). Other devices can be included, considering smartwatches and biometric information (e.g., heartbeat) and extending to the IOT field, such as all home automation sensors. The combination with annotations from intensive longitudinal surveys is what, according to our thesis, enables the generation of quantitative Big Thick Data.

Study It is any research that involves the design, collection, management, and analysis of data, which, in our case, is data about the person in their daily life. A study can be exemplified in different methodological approaches, from the experiment to the intensive longitudinal survey. It can be based on various stimuli, from questionnaires to daily diaries, be conducted with different tools (online questionnaire platforms, iLog), and include a variety of data, from those of direct interaction with the person (e.g., answers to questions) to those coming from sensors. Given the interdisciplinarity of the concept of Big Thick Data and context and the reconfigurability of iLog, in this thesis, the term "study" is preferred where it does not refer to particular methodologies or their components.

Survey (S.) . . . . It is a method of gathering information from a group of individuals by asking them questions. For this reason, surveys are often associated with the survey instrument, i.e., the questionnaire. In our case, we refer to the survey as the classic studies conducted in the social sciences, as opposed to the studies proposed in this thesis, which include a variety of approaches and measures.

Time Diary . . . . Like ESM, Time Diary is an approach that is part of the daily diary methods. It was specifically designed to understand a person's (or family's) daily activities, considering aspects such as where they are, what they are doing, and who they are with during the day. It is a particularly relevant approach for understanding citizens' habits and the problems they face every day. For this reason, it is used by social research institutes such as EUROSTAT and ISTAT in Europe and Italy. In our case, the HETUS standard (ATUS in the American case) is applied to ask daily questions on the main aspects of the person's context.

All in all, it was all just bricks in the wall

— Pink Floyd *Another Brick in the Wall, Part 3*

1

Introduction: on the concept of Big Thick Data

We live in the age of ubiquity, where thousands of devices track aspects of our daily lives. Fixed cameras in shopping centres observe the passing of customers; vacuum cleaners build a map of our homes; smartphones observe us spending hours between different applications or chatting with our friends and colleagues; and smartwatches collect biometric details on our heartbeat. All these technologies generate an abundance of data flows which, due to their volume, speed and variety, have been defined as Big Data [2].

Yet, despite the abundance of information, many technologies fail to recognize a person's context, which is fundamental to their correct functioning. Without learning the personal point of view on the activities a person does, where they are, with whom they are, and their feelings at different times, devices cannot provide the information and personalized assistance people need. Supermarket cameras tell us a lot about the flow of people present and help prevent theft, but they do not recognize individual people's tastes and shopping preferences. Smartphones follow us everywhere in our everyday lives, but they learn little about all those complex interactions with the environment outside the device or when it is not present. Smartwatches give us metrics on our heartbeat, but they know little about our feelings and emotions, often providing us only with an aggregate of data that identifies our level of "stress" (for an extended discussion on the recognition of self in the person, informatics see [3]).

In other words, as rich as these data collections are, many useful variables are often unavailable. Therefore, people are "at best, thinly described" [4]. Indeed, despite the impressive size of some Big Data datasets, they often extend over a few variables, usually sensor data, making it impossible to recognize complex aspects of human behaviour. In other words, the granularity of the sensor data is essential but not enough to represent people's context. The lack of essential variables makes the data poorly reusable outside its context, or, as [5] would say, "data are often used 'out of context,' which decrease the 'meaning and value'". This is true in applications and research, where technologies are often developed under controlled settings, and participants are instructed to perform a predefined set of activities. Systems and technologies trained in this way often have poor implementation in the wild, where the human behavioural context varies in different environments.

Indeed, according to these assumptions, [6] propose a context-aware model to recognize people's physical activities. Based on a dataset with annotations of the activities carried out by participants (see ExtraSensory dataset [7]), such as doing exercises, watching TV, etc., and the locations in which they are located, the authors identify a series of patterns of activities via sensors, such as the accelerometer, validated in the personal context thus obtaining fifteen different context-aware activities. This approach has great potential, especially in the long term, where, once the person's complex activities have been learned, it will no longer be necessary to involve them by requesting feedback and notes. However, it is well known that social and personal contexts substantially impact our physical actions. The same activity can be conducted differently if you are in the presence of other people or depending on your mood. For example, on an evening at home with friends, a person could sit in the same place (perhaps with the TV on), but the main activity would be communicating with her guests. Similarly, our sports performances vary depending on our psycho-physical state of health: even if the accelerometer detects a different speed of our movements (so much so that we mistake them for a walk rather than a run), if we are in a bad mood or we have a cold, from our point of view we will still be doing physical exercise.

In the social sciences, the personal and emotional aspect of the context is explored more deeply. For example, a study by Pe and colleagues [8] investigated how people remember positive and negative stimuli using daily questionnaires on their well-being (assessed with an affective n-back task). The study shows how remembering more positive information helps to maintain and further enhance positive thoughts and emotions, influencing the person's general well-being. Clearly, this approach highlights how the aspects concerning the perception of self are fundamental for an accurate understanding of the personal context but not being framed in a more complex and sensorial notion of the context, also based on machine learning (as in the study presented previously), is challenging to scale to studies and applications that are considered long-lasting. Furthermore, neither this study nor the previous one considers the environmental aspects of the context: what happens to our activities and our mood when we are in a completely different location than usual? For example, when we are on holiday or when we change cities for work. In other words, how we relate to the environmental context and its stimuli.

Many studies deal with the environmental context by involving the person in observing himself, many of which can be attributed to the Citizen Science approach [9]. In this approach, people (citizens) actively monitor environmental events around them, from the migration of birds (see, e.g., [10] to the classification of plants and animals (see, e.g. [11]). Although these approaches have the positive aspect of exploring the environmental context in detail, they are not interested in deepening the relationship this has with the person, although, even in this case, there would be positive aspects. For example, a person who lives in the place under observation, who carries out many activities there and is interested in, will usually be able to provide reliable and quality information, unlike a tourist.

In summary, a detailed observation of the person's context (considering the activities that a person does, where they are, with whom they are, and their feelings) favours a more complex approach to human behaviour, facilitating the creation of complex and human-aware applications and facilitating wider reuse of the information collected. This thesis follows this intuition, articulating it in the concepts of Big Thick Data and Situational Context, as described in the following sections.

Before proceeding with the discussion, it is helpful to focus on a further aspect. Whether we consider the activities, the personal aspects, or the environment, all these approaches have in common the use of annotations, i.e. a direct interaction with the participant. The lack of such annotations and essential information is a cross-domain problem, often solved via *post-hoc* annotation or blending. For example, in the Human Activity Recognition (HAR) field, they call it the problem of the user diversity and transfer learning [12], i.e., transfer of knowledge from an existing domain into a new domain, which is mainly done via human annotation. Similar to HAR, in Human-In-The-Loop approaches[1], "annotating data is a complex but crucial task" [13]. At the same time, in context recognition, a dataset that integrates human feedback is needed to validate algorithms in the wild [14]. That's also why the health domain traditionally bases its observations on subjective reporting, sometimes through daily diary methods [7]. Adding annotations can be seen as part of blending techniques aimed at integrating datasets from different sources to obtain thick data descriptions, also called Big Thick Data [15], as described in the next section.

1.1 Big Thick Data

Figure 1.1[2] shows the Cartesian plane on which all the data and data sources referring to the person are plotted. The plane is split into two opposite couples:

[1] Human-in-the-loop are technical approaches that involve humans in every decision made by the system, even though this is often neither possible nor desirable due to the user burden. Similar to these approaches are Human-in-control. Some systems have built-in levels of human supervision. For example, in a military AI application/technology, the human operator may have the final say in activating system actions/implementation of system decisions.

[2] Adapted from [15, 16]

Thin and Thick and Extensive and Small. Big-Thin data sources are in the top left corner of the coordinate system, characterized by their extensive numbers but lack of contextual depth. Examples include datasets generated from sensors, such as GPS location data, which are abundant but do not represent the person's perspective. To stress this concept, considering another distinction between *space* and *place* can be helpful. The seminal work of [17] proposed the distinction and is often used by geographers and urban sociologists. Space is the set of physical objects that surround us, like a street or a building, and so on, while place refers to how people perceive the space, the feeling that they associate with that particular street or building, and the experiences they lived in. Without people, space doesn't have any substantial meaning. Now, the question is: how do we recognize the place?

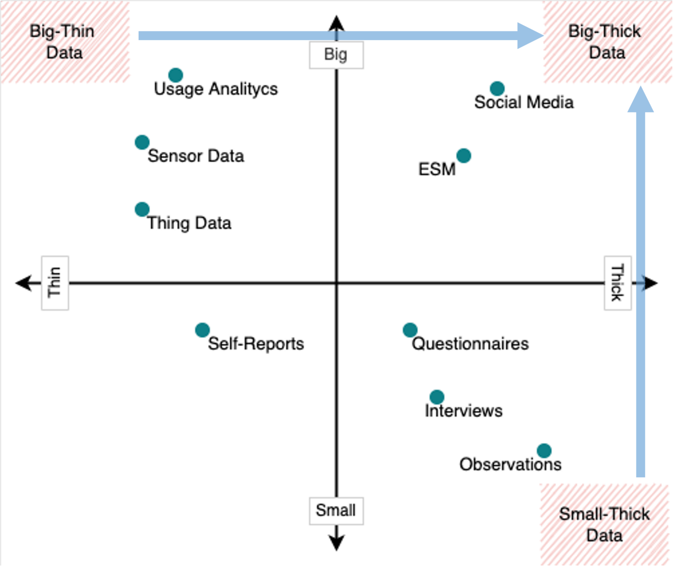


Figure 1.1: Big Thick Data

Through "thick[3] data" (i.e. the opposite end of the spectrum) which are collected through typical social science approaches and tools, namely participant observation and ethnography, interviews, questionnaires (listed from the thickest to the thinnest).

The "Big-Thick Blending" concept focuses explicitly on integrating ethnographically collected thick observational data. This blending approach combines the extensive reach of big data with the contextual richness of thick data, providing a more holistic and nuanced understanding of human behaviour and interactions.

The problem with these techniques is that adding annotations is time-consuming and, ultimately, as observed by [7], the richer annotation comes from the person's point of view. On the other hand, with the increase in the dataset's quality (i.e., specificity), blending can only be done by domain experts, namely, by a person who can understand in detail the context in which the data are collected. In other words, domain-specific datasets are essential but hardly bendable, i.e., reusable.

[3] The term "Thick" comes from the seminal work of Clifford Geertz [18] that focused on dense, detailed collection and analysis of the context in which human behaviour occurs

1.2 The context notion for quantifying Big Thick Data

To account for the quantification problem, the thesis aim to develop a novel methodology, considering a quantitative approach to Big Thick Data that integrates both the detection aspect of Thin data (i.e. sensors) and the aspect of Thick data (i.e. annotations). Therefore, we rely on a comprehensive formal model that captures and analyzes diverse facets of human experiences, emphasizing the representation of situational contexts, life sequences, and habits. According to [19, 20] the *Personal context C* can be defined as a 4-tuple (see also Figure 1.2

$$C(t) = \langle WHERE(t), WHAT(t), WHOM(t), WITHIN(t) \rangle$$

where: *WHERE*, the *spatial context*, defines the place where the person is, *WHAT*, the *event context*, defines the activities that the person is involved in, *WHOM*, the *social context*, defines the other people with which the person is, *WITHIN*, and the *internal context*, defines the internal (mental and/or physical) state of the person. In the equation above, the parameter t is an instant in time, meaning that the goal is to collect information continuously, in time. The sequence of moments t when information is collected, in terms of user-provided information and sensor data, is called the *TIME context*.

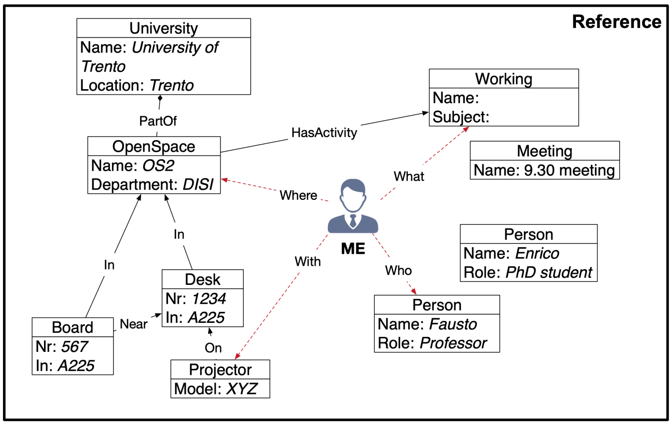


Figure 1.2: An informal representation of context

Each context other than time context can be modelled as a Knowledge Graph (KG) [19–22], namely as a set of Relation, Subject, and Object R(S,O) triples, where the Subject is, typically, the person.

Examples of such triples are *In (Person, Home)* and *With (Person, Friend)*, *Near (Location1, Object1)*, where *Person* is the person providing the information through her smartphone.

In other words, personal context is everything represented from the person's point of view, whether directly experienced and noted or collected through other sources.

In this sense, a second concept helps us, which is that of *Reference context*, which constitutes everything external to the person but can still be recognized within the context.

1.2.1 Situational Context and Life Sequences

The concept of a situational context serves as a foundational building block, encapsulating real-world scenarios through the lens of an individual, referred to as "me". The text outlines a life sequence, denoted as $S(me)$, constituting a set of situational contexts delineating specific time frames. Within each situational context, denoted as $C_i(me)$, crucial information is encapsulated, including the spatial and temporal aspects of the location, various events unfolding, and a spectrum of entities participating in the given scenario.

Indeed, all the triples mentioned above are annotated with timestamps. Triples with the same timestamp belong to the same context KG. This allows us to integrate the KGs of the single contexts into a more extensive timeline KG. Typically, a user-reported annotation allows the generation of one of these triples. However, these triples are also annotated by sensory information, which can also constitute approximate ground truth for the triple itself. For instance, the first triple could be annotated by GPS, and the third by GPS and Proximity information.

Therefore, it is possible to recognize recurring activities systematically, each annotated with its frequency. In this sense, the model can categorize a relevant aspect of human behaviour, namely habits [22, 23] observed into spatial, temporal, social, material, and action habits, offering a nuanced perspective on the various dimensions of repetitive behaviours over time. The structured categorization of habits enhances the model's adaptability to a broad spectrum of human activities and routines.

1.2.2 Similar notions

Various context notions have been proposed in the past; see, e.g., [24, 25]. Similarly to this earlier work, the schema of KGs is modelled as an ontology [19, 21, 22, 26]. However, the strength of the situational context perspective relies on the fact that the person is not asked to annotate some specific sensor value but, instead, to describe her subjective understanding of the current situation. This means that the main components of the context are always annotated and that these can be integrated or extended through the observations of the sensors and further data - both collected in the dataset and from other sources.

In summary, the formal model presented in the text provides a sophisticated and adaptable framework for representing and dissecting the multifaceted nature of human experiences and behaviours. By seamlessly integrating situational contexts, life sequences, and habits within the structured confines of a KG, this model emerges as a powerful tool for researchers and analysts seeking a comprehensive understanding of individual everyday life.

1.3 Populating Big Thick Data as a stream of situational contexts

If it is possible to frame Big Thick Data as a stream of Situational Context, the approach is still far from being operational. To be applied and validated, the idea needs a theoretical framework of reference, which makes it possible to define metrics and data collection methods that populate the different areas of the context. This presents several challenges leading to the definition of a new methodology for the design, collection, management, and sharing of valid and reliable Big Thick Data. The next sections focus on (i) the validity and reliability of the data, considering relevant aspects such as (ii) the theoretical framework, measurement methods and cultural context, taking into account (iii) ethical and privacy issues, but also (iv) how to manage the data collected, and their (v) validation through reuse.

1.3.1 Validity and reliability

The process of observing and measuring aspects of the world involves the identification of suitable instruments and methods which should be valid and realiable [27]. To further explore these concepts, le us consider the example of determining a person's location, which can be measured through GPS coordinates, direct questions posed to the person (e.g., "Where are you?"), or leveraging information from the person's smartphone connectivity.

Validity pertains to the accuracy of a measurement in capturing the intended aspect of the real world. In the context of location observation, if the goal is to pinpoint a physical space, GPS coordinates would be considered valid. However, if the objective is to understand the qualitative aspects of the place, such as its characteristics or activities, GPS alone might lack validity. Other methods, like direct questioning, may be more appropriate for capturing these nuanced aspects. In other words, the choice of measurement method should align with the specific observation goals.

Reliability focuses on the consistency and reproducibility of a measurement over time. Considering GPS coordinates, the reliability is associated with the accuracy of the measurement. High accuracy implies consistent latitude and longitude readings when a person is at the same point in space. In contrast, unreliable measurements may exhibit variability even when the person is in the exact location [4].

[4] The notion of systematic and random errors [28] is relevant to understanding reliability. Systematic errors consistently affect measurements by the same amount, while random errors introduce unpredictable slight variations.In other words, systematic errors are consistent across measurements and can be mitigated with standardized and careful procedures (e.g., use a weareable sensor to measure the location instead of the smartphone GPS). On the other hand, random errors introduce unpredictability, often stemming from uncontrollable factors during measurement.

In a nusthell, both validity and reliability are intricately linked to the purpose of the observation. The choice of measurement method should align with the specific goals and context of the observation. Different purposes may require different measurement approaches to ensure the meaningful capture of the intended aspects of reality.

1.3.2 Frameworks, measurements, and cultural context

The validity of an observation strongly depends on the reference framework, the objectives with which the data collection is conducted, and the cultural.

The notion of Big Thick data is inerently interdisciplinar, since the themes of a study that includes the self in a person's daily life can be among the most diverse, ranging from the individual experience during specific events to habits and routines in daily life, to the observation and cataloguing of external events. A computer scientist focusing on Personal Informatics and Machine Learning may be interested in levaring sensor data to produce accurate measurment of physycal behavior, while a psychologist would focus on the personal interpretation of the same behavior, and a sociologist would draw conclusion on how the behavior is shared within a group of people. Depending on the approach, the data collected could be valid and reliable according to one discipline but not so in another. The social scientist interested in knowing the locations visited by a person during her daily life needs to take observations in the wild and consider contextual information, such as the type of place visited or the reasons for visiting it. At the same time, GPS coordinates collected with reasonable accuracy (e.g., 50 meters) and a low frequency (e.g. once every 10 minutes) will be more than sufficient to reach the study objectives. On the other hand, the computer scientist focused on activity recognition needs high GPS accuracy and frequency, and a small and controlled study setting will be ideal to ensure greater validity of the results.

In addition, the cultural context, the people involved, and the measurement tools influence the results. A location may be called by different names or in different languages and be visited by multiple people with various aptitudes and abilities. A visit of the same duration to the same place, for example, a Buddhist monastery, can convey different activities and meanings if the visitor is a monk, a follower, or a visiting foreigner, producing different results for the observer. Furthermore, not all the observed people may be accustomed to using the same measuring instruments, just as not all measuring instruments are suitable for collecting valuable information for observations. For example, an elderly person unfamiliar with the use of a smartphone will have more difficulty configuring the GPS. At the same time, there are few applications capable of collecting the data necessary for investigation purposes.

1.3.3 Ethics and privacy issues

Big Thick Data is primarily data from and about the person. Therefore, their collection and use can have a personal and social impact, both positive and negative. Pervasive data collection can affect people's behaviour or subject them to unexpected stimuli, and the same data could lend itself to malicious use. For these reasons, Big Thick Data is strictly connected to ethical reflections and procedures aimed at protecting privacy and security, which must accompany all aspects of the process, from the purpose with which they are collected to their preparation and sharing.

1.3.4 Data preparation

The raw data collected about a person is never ready-to-use but must be cleaned and quality checked. This is not only for practical reasons of facilitating analyses and for the production of scientific literature but also ethical reasons, as incorrect or poor-quality data affect the validity of the conclusions reached, and poor management can lead to some data loss. For this reason, data management is particularly important, as are the security measures adopted during their processing.

1.3.5 Validation through reuse

Ultimately, the validity of the data is verified through approval from a scientific community or application in real-world scenarios. As we have seen, Big Thick Data is highly interdisciplinary and multicultural. Furthermore, given their abundance, they cannot be validated in the course of a single publication or a single application. For this reason, their distribution is necessary to facilitate their interdisciplinary and multicultural reuse, thus encouraging the proliferation of validations in the different scientific communities.

1.4 Towards a methodology for valid and reliable Big Thick data

Despite the abundance of data about people, the absence of fundamental variables makes their thick representation impossible. By considering the person's point of view on their context, it is possible to generate meaningful information fostering advanced research and new applications. However, as attractive as these statements are in theory, they require their scientific and interdisciplinary operationalization and validation and there is currently no end-to-end methodology capable of handling Big Thick Data. Therefore, a method is needed to put theory into practice.

The main contribution of this thesis is a methodology and a set of supporting tools for designing, collecting, managing and distributing ethics and privacy-aware, valid and reliable Big Thick Data quantified through the Situational Context notion.

Chapter 2 shows how different disciplines can contribute to the definition of the methodology, considering aspects of study design in compliance with ethics and privacy, data collection, preparation, and reuse. Furthermore, it considers available platforms and tools enable the replication of the methodology phases.

Chapter 3 describes the iLog Methodology for designing, collecting, and distributing the Big Thick Data in an ethical and privacy-compliant way. The methodology is inherently interdisciplinary, therefore flanked by a set of support services and materials to encourage its widest reuse by practitioners with different expertise.

Chapter 4 and **5** show how the methodology has been applied to collect Big Thick Data for different purposes and that it has been reused by researchers from different scientific communities. On the other hand, **Chapter 6** and **7** highlight the reuse of services and support materials connected to the methodology, as well as their validation provided by a panel of interdisciplinary experts.

Finally, **Chapter 8** draws the final reflections on the methodology, the reliability and validity of the data collected, highlighting its limitations and future steps.

About the WeNet Project

Before proceeding with the dissertation, it is important to underline how the methodology was developed within the broader framework of a Research Infrastructure, one of the most innovative products of the WeNet Horizon 2020 Project [29]. The methodology, the services connected to it, and its validation are, therefore, the result of an extensive collaboration with the consortium members, within which the writer participated in various work packages with different tasks. In particular, the writer collaborated in WP1, contributing to the pilots' design and operationalization [30]. Secondly, he collaborated in the implementation of the questionnaires, in the coordination of Diversity 1 (see Chapter 4) and 2 surveys, and in the preparation of time diaries and questionnaires data, contributing to the drafting of the deliverables [31, 32] of WP7 and drafting the materials in the Appendices A, C. He also participated in the design of the Research Infrastructure and the LivePeople Catalog, coordinating the deliverable [33] in collaboration with WP6 and following the legal aspects of copyright culminating in the materials in the Appendices B.6, D, and E. Finally, he collaborated with WP9 and WP11, contributing to the definition of the anonymization procedures and preparation of the data and materials reported in the Appendices B and in the deliverable [34, 35]. In a nutshell, the writer unique contribution are the conceptualization of Big Thick Data as Situational Context, the aligning of social science methodologies to the iLog procedures, the creation of protocols and services and their evaluation. On the other hand, the research idea underlying the Diversity 1 pilots and the first operationalizations of the context through time diaries are to be attributed to Bison and Giunchiglia (see, e.g., [36]). The development, implementation, and maintenance of iLog and its execution in data collections, the preparation of sensor data, and the development of the Catalog in JKAN must be attributed to the fundamental support of the KnowDive group of the University of Trento.

2

State of the art

Contents

When discussing methodologies, the study's replicability and the data's reproducibility are the first aspects to consider. A study is considered replicable when different researchers, by repeating the same procedures, are able to obtain the same results. Reproducibility is part of replicability and has to do with the possibility of examining the analysis of a given set of data. It is now known that various scientific disciplines are affected by the so-called "replication crisis" [37]. Although observed in the medical and psychological fields, the replicability crisis involves all disciplines, not least AI [38], where the lack of accuracy in validating the reliability of the data, the difficulties in explaining the algorithms and the possibility of playing with multiple parameters of machine learning have raised several concerns that AI "is driving a deluge of unreliable or useless research" [39]. On the other hand, a famous article published in 2016 [40] (see also [41]) shows how poor methods in data management have led to a situation in which 1,500+ researchers, more than 70% of researchers have failed to reproduce another scientist's experimental results, and more than 50% have failed to reproduce their own experiments. From our point of view, this problem has to do with the research field social dynamics (e.g., the publish or perish problem), which are outside the scope of this thesis, but also with the lack of supporting materials that could guide practitioners through the research process. This often requires interdisciplinary skills (in design, ethics, and the implementation of tools and instruments for data collection and management), which are difficult to achieve, mainly by individual researchers or small research groups. This latter aspect is fundamental for our thesis, which aims not only to provide a theoretical framework and guidelines but also the tools to put them into practice, facilitating the reproducibility of the studies and the replicability of the results.

Having said so, collecting Big Thick Data, namely involving people in research, adds a considerable complexity that can affect the replicability and reproducibility of the results. In fact, as highlighted in the introduction, while conducting a study it can occur both systematic and accidental errors [27, 28], which concern all phases of the investigation process: from the operationalization of constructs (i.e., moving from a theoretical concept to an empirical and measurable one), to the definition of investigation tools, such as questions and observations, to the reliability of people's verbal behaviour, up to privacy and data management [42]. Furthermore, given that Big Thick Data is eminently interdisciplinary, as is the methodology, there is a particular problem in making the data usable by different scientific communities and making the methodology executable by different types of experts.

Luckily, these problems have already been partially addressed within the social sciences. There is a plethora of handbooks that deal with methodology, in general, [27, 43, 44], and research design [45], which consider privacy [46], or provides guidelines on how to interact with people by asking structured questions and providing stimuli [47]. Other handbooks consider the possible errors that may occur during survey [28, 48], and even deal with data management and their dissemination for secondary analysis purposes [42].

In general, the different handbooks agree in identifying five phases for conducting a study about people, namely:

1. **Design**: this phase includes all aspects concerning the creation of a study, from the definition of the investigation topic to the structure and methods of data collection, to the definition of the people to be involved and the methods of involvement, up to the aspects concerning resources, timing and planning.

2. **Ethics and Privacy**: in this phase, all aspects that could entail an excessive risk or burden for the survey participants are evaluated, both from a personal point of view and from a procedural point of view regarding data management

3. **Collection**: this is the operational phase of the study, in which data collection is carried out, covering both the active involvement of participants and the collection technologies.

4. **Preparation**: this phase is often not considered in survey design manuals (also because it is strongly dependent on the technologies used for data collection), despite it being essential for assessing the quality of the data and ensuring usability and analysis. It involves aspects such as data extraction and cleansing, storage and documentation.

5. **Distribution**: this phase consists of making the data accessible to third parties, through the use of catalogues, databases and publicizing the data and results in scientific journals.

If it is true that the literature is extensive, it is also true that the advent of new technologies for data collection and so called intensive longitudinal surveys (as described below) has posed a new set of problems. For example, is asking about personality on a synchronic scale (e.g., the Big Five [49]) the same as discovering its traits in the diachrony of everyday life [50]? Does people's verbal behavior have the same level of reliability when reported at different points in time [51]? But also, what role do data coming from sensors have (see e.g., [52])? Is it possible to use sensor data to discover characteristics of human behavior?

Some of these answers are found in new research and emerging handbooks, such as [53], but many issues still need to be addressed, which is the fundamental reason for the definition of a new methodology.

The next sections describe the state of the art of each of the five phases. Since the iLog methodology aims at being highly reusable in different disciplines through the provision of support materials, the sections will cover also the platforms and technologies currently available for supporting each methodological phase.

2.1 Designing intensive surveys with people

The research design identifies objectives and strategic practices to achieve the results. In general, in the research design [27] aspects such as (i) topic and objectives, (ii) study structure, (iii) measurement, (iv) sample design and incentives, and, in some cases, how to conduct research in (vii) Cross-country studies [54]. Each of the aspects will be described below.

2.1.1 Topic and objectives

The topic of a study that includes the self in a person's daily life can be among the most diverse, ranging from the individual experience during specific events to habits and routines in everyday life to the observation and cataloging of external events. According to our definition of Big Thick Data as operationalized within the situational context notion (see Chapter 1), three main topic categories can be defined:

Object context [55] provides an extensive scientometric review of this category, which includes all those studies that aim to recognize a set of activities carried out by the participant within single events or specific locations, such as sports training sessions or the movements of a patient undergoing rehabilitation within a hospital facility. It is defined here as the Object context as opposed to the Personal context as it does not involve interaction with the person to recognize their presence or activities, but is based on devices (i.e., objects, such as smartphones, smartwatches, cameras) which, through sensors, perceive the environment and report an "objective" measurement of it.

Personal context These studies aim to understand the person's point of view within their routines and habits in everyday life. This approach can be verticalised into specific studies aiming at understanding the single contextual components, namely:

1. *Interoception*: include all studies in which the participant is encouraged to reflect on the self and physical and emotional states. Interoception is the main approach of Experience Sampling Methods (ESM)[56] studies. ESM is an intensive longitudinal social and psychological research methodology, i.e., designed to reduce social and cognitive bias in data collection, where participants are asked to report on their thoughts and behaviors. [57] provides an extensive analysis of the literature.

2. *Social interactions*: focuses on the observation of social ties between people, from the creation of networks to the exchange of messages via social media [58] or chat application [59].

3. *Human-AI interaction*: they are approaches that involve a certain interaction between machine and human, where the stimuli are not generated on a table by a researcher but are the result of a circular exchange of feedback between human and machine from which an adaptation of the AI to the person derives (see, e.g., [60–62].

4. *Multi-purpose*: studies involving different topics within the same disciplinary field and combining various fields. An example is the multi-purpose surveys of Istat [63], which is "part of an integrated system of social surveys and collects fundamental information on individual and families daily life. The survey provides information on the citizens' habits and the problems they face in everyday life". This approach is often based on the time diaries [63–66], considering what a person does, who they are with, and where they are at different times of the day.

Reference context all approaches that involve the participant as a "sensor" or capable of detecting quality information about the surrounding environment, typically conducted within participatory sensing [67] or citizen science approaches [9]. A classic example are the challenges proposed by CornellLab [10] in which people are invited to photograph and classify different species of birds. But it can also refer to other aspects besides fauna, for example the weather, vegetation or the shape of a specific urban area. The peculiarity of the Reference context is that it does not belong to one person, but is shared by all the people who experience that specific point in space and time.

2.1.2 Study structure

In social science, there are many different study structures [27] (see also the definition provided by [68]). Study structure is the global organization of the investigation, from which derives the protocol, the sampling strategies, and the adopted measures. A first distinction in study structure is based on time and space and concerns *cross-sectional*, *longitudinal*, and *cross-cultural studies*. The main objective is to obtain observations representative of the population under study, usually considering the population's age (or other similar socio-demographic characteristics). In *cross-sectional* study, data are collected from a population at a specific time, while in a *longitudinal* study, the same data are collected repeatedly within the same sample over an extended period. Associated with these are *intensive longitudinal surveys*, i.e., longitudinal studies carried out in a short period but with a high frequency of questions. *Cross-cultural* studies can have both structures, but they differ because samples from more than one culture appear, generally identified with the nation. They aim to understand whether specific psychological, behavioral, or value characteristics are universal to humans or developed within different environments.

Another important distinction is in the study configuration, where social scientists [27] usually distinguish between *experiments*, *quasi-experiments*, and *surveys*. An *experiment* is a study in which there are two groups in which the participants are randomly assigned, and there is an active manipulation of the input variable (also called independent variable or "feature" in machine learning fields). A similar form known in the field of user experience is *A/B testing*, which consists of an experiment that usually involves two variants or more (A and B) of the same variable (e.g., two different versions of the same webpage), to determine which of the variants is more effective. In this case, the random assignment of the participants to one of the groups is not considered; that's why they can also be defined as *quasi-experiment*. If the experiment is conducted in a controlled environment, it is called a laboratory experiment, while in real-life settings, it is called a field experiment.

Unfortunately, most social phenomena do not provide such an artificial structuring, and it is often complex to understand the variables contributing to explaining a phenomenon. In other words, the occurrence of the observed variable is unknown. For this reason, social scientists often adopt *surveys*, which in our case are intensive longitudinal surveys, namely a study in which there is no active control on the input variable nor a division in experimental and control groups, but rather a complete observation of a population to study the relation between different input and output variables.

2.1.2.1 Study duration

The study duration is usually linked to three factors: the studied phenomenon, the amount of data needed, and the intensity of the measurements. The latter will be explored in depth in the next section. In this sense, there is no particular difference between the different investigation structures. Still, on average, *Objectual context* studies last a few days [7], as do some studies on *Interoception*, which, however, also include studies lasting six months [53]. For the other categories, the duration is variable, but it is advisable to consider at least one month of data collection.

2.1.3 Measurement

According to AAPOR, a scientific survey “uses reasonable tested methods to reduce and account for errors of measurement that may arise from question-wording, the order of questions and categories, the behavior of interviewers and respondents, data entry, and the mode of administration of the survey.” Measurements are how the participant is systematically observed in their daily life. Drawing a measurement is the input from which the data will arise and is, therefore, the cornerstone of the study. Measurement can be either passive or active. Passive measurements are all the information about a person that can be collected without direct interaction with the person herself, such as data coming from the smartphone sensor. On the other hand, active measurements are usually items used to evoke a reaction from the study participant, whether it is an answer to a question or feedback given to stimuli, which can be an audio, visual, or physical stimulus coming from the tools used in the study or from the environment.

2.1.3.1 Measurement reliability

In social sciences, there is extensive literature on how the interaction with the participants and the measurement quality may affect the results on the data [27, 28].

Regarding active measurements, it is helpful to remember that the relationship between participant and researcher is always asymmetric, where the researcher is interested in obtaining specific information from the participant. Within this relationship, a series of dynamics are generated, many of which are attributable to the *Hawthorne effect* [69], or the alteration of behavior by the subjects of a study due to their awareness of being observed. As [27] pointed out, during a measurement, *social desirability* effects can therefore occur when the questions asked concern behaviors and attitudes to which strong positive or negative values are associated, for example, if the number of hours spent by a student in preparing for an exam is investigated. In this case, the student could tend to overestimate or declare more hours of study than those done, trying to show a certain positivity in her behavior or, on the contrary, underestimate them due to excessive zeal, leading the answers to be less reliable. Similarly, there are conditions of *non-attitudes*, i.e., when the respondent knows little about the topic on which the question is asked, has not understood, or does not know how to answer but feels the pressure to respond, provides an answer at random. Another aspect that can lead to random or missing responses is the respondent burden [70], which is particularly important when considering intensive longitudinal surveys. The respondent burden is the effort required to answer a questionnaire or how the responder perceives the participation in terms of how long it will take the difficulty level, and the emotional toll. However, [71] demonstrated no significant difference in participant burden within the frequency of the questions since the burden is more affected by the length of the questions themselves.

2.1.3.2 Active data

As described above, active measurements are items used to evoke a reaction from the study participant, which can be divided into an open or a closed question. Open questions usually refer to a free contribution by the participant, which can be a text or an image. In contrast, in closed questions, the participant is presented with a range of alternative answers to the question [47]. Both have some advantages and disadvantages, but usually, closed questions are preferred [27] since (i) they provide a frame of reference, clarifying what is expected as an answer from the question; (ii) they help participants in retrieving information from their memories; (iii) they simplify the analysis task since they provide already codified answers which can be directly used in quantitative analysis. On the contrary, they may (i) miss some important alternative options, (ii) they can influence the answers, and (iii) they may not have the same meaning for all the participants. Other possible active data can be collected, for example, from social interactions via social media [58] or chat application [59].

Usually social scientist often follows standards. Indeed, as Sudman and Bradburn (in [27]) point out, "copying questions from other questionnaires is not plagiarism" but instead a "recommended practice, in that it enables knowledge to be accumulated and comparisons to be made". In addition, this practice reduces the possibility of incurring systematic errors [28]. For multi-purpose studies, generally, social scientist relies on the ATUS [65] or HETUS [66] standards. Other repositories provide state-of-the-art standard scales, like GESIS "Item and Questions" [72] and items that have been already used in ESM studies [73] within their timing, but there are also approaches in converting classic stimuli (e.g., a Likert scale) into ESM stimuli (see, e.g., [74]).

Timing According to the ESM framework [75], questions can be of three types:

1. Signal Contingent: questions are asked randomly throughout the day at random times.
2. Interval Contingent: questions are asked at fixed times. Time diaries are part of this category.
3. Event contingent: the participant reports on the events as they happen.

The frequency usually depends on the length of the study. Generally, in multi-purpose studies [63], questions can be even every ten minutes since the usual duration is one day. However, for more extended studies, the frequency usually decreases.

2.1.3.3 Passive data definition

As mentioned above, passive measurements are all the information about a person that can be collected without direct interaction with the person herself. There are numerous types of sensors and multiple ways of managing them. This issue in data collection will be addressed in Section 2.3, while additional insights are provided in Section 3.1.

2.1.4 Sample Design

Sample design is pivotal for the success of a study since the number of people and their level of involvement in data collection, i.e., the amount of information they are willing to provide and the assiduity with which they offer it, determines how much information is present to answer a research question and derive reliable conclusions. There are many ways to define a sample [27], usually divided into probability and nonprobability sampling strategies. However, unlike sample surveys, the sample size in intensive longitudinal surveys varies drastically depending on the disciplines and objectives, and there are no precise guidelines. On the one hand, there are studies with a few dozen participants, and, in some cases, even with just one; on the other hand, some studies reach 600 participants [53].

Although particularly recent, interoception studies have a more structured approach due to the type of data and data analysis. Indeed, this type of study has a multilevel structure in which repeated observations over consecutive days are nested within participants. Therefore, it is possible to quantify the number of samples through a variant of the well-known power analysis approach (see, e.g., [76]). In this area, a valuable tool for computing power analysis has been designed by [77].

And additional issue in sampling theory is the so-called ‘selection error’ [28], which involves not only sampling error, but also coverage and non-response errors. Coverage error stems from the fact that practitioners often have no access to a comprehensive list of the members of the population. Non-response errors may have two distinct causes: failure to contact selected subjects selected, and some subjects’ refusal to be interviewed.

2.1.4.1 Incentives

Unlike cross-sectional surveys, longitudinal surveys suffer from the problem of dropout due to the burden of compilation and the duration of the studies. For this reason, it is pivotal to provide incentives to the participants in the study. [78] has provided an extensive literature review on the type of incentives (monetary or non-monetary) showing how this can reduce respondent bias [28]. The article also presents helpful suggestions for research practitioners, like the fact that monetary incentives usually work better than gifts and prepaid incentives perform better than random prizes.

2.1.5 Cross-cultural studies

A final note concerns cross-cultural studies. Cross-cultural studies are particularly complex as there is a strong tension between the need for standardization and invariance of the stimulus and the sensitivity of the different cultures in which the research takes place [54]. Therefore, asking questions in diverse countries is not just a matter of translation, but it involves knowing and recognizing how norms, opinions, values, and beliefs are shaped within the culture. For this reason, many scholars focus on adapting their standard scales to other countries [79].

2.1.6 Design services

In addition to the various manuals, the repositories of standard or already used questions, and the tools for conducting a power analysis in the selection of the sample listed in the previous paragraphs, there are some valuable tools in the design of data collection. In particular, platforms with GESIS [72] offer consultancy services, while UKDataArchive [80] provides various information materials and summary guides to approach the different design phases. Furthermore, to guarantee the reproducibility of one's study and a correct approach to the investigation methodology in the field of cognitive science, it has become common practice to pre-register one's studies on the OSF [81] platform.

2.2 Ethics and privacy principles

Ethics and privacy encompass all aspects of a survey that involve people, from the design to data collection up to data preparation and distribution. Ethic can be defined as codes of conduct that define acceptable and unacceptable behaviors regarding the person that do not infringe on their autonomy or identity, while privacy as the set of legal principles to implement good behavior (i.e., legally sustainable), considering privacy and related rights such as the right to be let alone and control one's personal information.

In the digital era [1], ethics and privacy are highly relevant, in particular in the context of personal data. Although many strategies to mitigate the risks associated with treating personal data, especially in the AI fields, have already been proposed, such as explainability [83] or the report on Ethics guidelines for trustworthy AI [84], a growing body of literature is focusing on bias and bias management (see, e.g., the extensive work done by [85–87]) in scientific fields such as AI, health, and behavioral studies. Furthermore, how personal data is treated is not risk-free, especially when considering consumer privacy, non-transparent legal regulation, or even bias in the programming. Examples are the processing of data for advertising conducted by Google or Facebook [88], but also all the documented cases of "untrustworthy" AI (see, e.g., the cases related to face recognition [89]).

For this reason, various ethical and legal principles have been defined to follow during an investigation, as reported below.

2.2.1 Ethical principles

There is extensive literature regarding ethics in research, based on multiple approaches [90, 91] and different codes that have led to the definition of other principles. Furthermore, there are also various regulations at an international level [42, 90], as described in Section 2.2.3. The topic is particularly felt in the various scientific disciplines, so much so that many journals have their ethical code (see, e.g., ASA [92]). Although the debate on an ethical approach to science is generally relevant when dealing with methodologies, this thesis focuses on those principles that guide the relationship between researcher and participant. In this sense, four guiding principles [90, 93] can be identified as follows:

1. **Informed consent**: as far as possible, both participants and researchers must be aware of the methods and procedures of the investigation to guarantee

[1] This section is an adaptation of [82]

autonomy in decision-making. In this sense, informed consent is linked to aspects of voluntary participation, i.e., not forced.

2. **Confidentiality**: it is an essential aspect of research, both from the point of view of personal protection and the relationship of trust established between researcher and participant. The research may involve sensitive information of the participants, the disclosure of which could pose a danger to the person's safety or to their daily life (for example, in the case of research on the use and consumption of alcohol, the information could affect the participant's work career if made public). Furthermore, the awareness of a relationship of trust and secrecy between participant and researcher facilitates the sharing of personal information, ultimately promoting the success of the research.

3. **Integrity**: this aspect concerns the relationship with the participants and other colleagues and researchers. Integrity is often associated with three behaviors to avoid, namely the fabrication of results, the falsification of results, and the plagiarism of the works of others. Regarding the relationship with the participants, the first two are particularly important, as creating fictitious data and manipulating accurate data to obtain satisfactory results can ultimately harm the people involved. This is particularly true in cases where, starting from research, investments, policies, or processes are promoted, which, at different levels, include people.

4. **Avoiding harm and doing good**: although the concepts of "harm" and "good" are particularly uncertain, it is possible to identify the ethical aim of the research as not to harm the interests of the participants nor put their psycho-physical well-being at risk, especially in testing phases. Some ethicists argue that research must also "do good," that is, generate well-being at a local or global level for the population.

To these principles is added a fifth, which concerns the description and publication of research results to advance knowledge. In this context, it is interesting to note how, according to [42], the social research field moved from an approach aimed at the dissemination and distribution of data to encourage the proliferation of research to a more conservative approach aimed at putting protection in the foreground of people and their personal information. Section 2.5 will examine some of these aspects.

Ethical principles are usually operationalized into legal principles aimed at guiding researchers in correctly managing information concerning the person. These will be described in the next section.

2.2.2 Legal principles

From a legal point of view, since it concerns information concerning people, the regulatory framework of reference concerns privacy and the protection of personal data. Added to this are the aspects of Copyright and intellectual property protection about research products, i.e., articles, datasets, and applications. Regarding privacy, different regulations developed by other countries, such as the California Consumer Privacy Act (2018), were created following the new European legislation.

Since the technologies used are located in Europe, the thesis takes as reference the European regulatory framework, as defined in the General Data Protection Regulation (GDPR) [94], and operationalized in the Article 29 Data Protection Working Party [95] (see also [96] for a critical appraisal) and the recent ISO/IEC standard 27559:2022 [97] on "Information security, cybersecurity and privacy protection – Privacy enhancing data de-identification framework". In general, this regulation applies in all cases where the use of data could pose a risk to people. Personal and sensitive data are particularly problematic. According to the definitions of Articles 4(13), (14) and (15), Article 9 and Recitals (51) to (56) of the GDPR, personal data is all information that allows a person to be directly or indirectly identified. By standardizing this definition, ISO indicates two main categories of identifiers: direct and indirect. Direct identifiers are all information that allows us to trace back to a natural person without intermediaries or calculations, such as name, home address, email, passport, or identity card code. Information such as gender, date of birth, and nationality are indirect identifiers. Sensitive data is a subcategory of personal data with the peculiarity of exposing the person to greater risk if disclosed. This category includes personal data revealing ethnic origin, political opinions, religious or philosophical beliefs, health-related data, and data concerning sexual orientation.

According to Article 5 of the GDPR, there are seven principles to follow when personal data is processed, namely:

1. **Lawfulness, fairness, and transparency**: according to this principle, personal data must be processed according to the relevant legal regulations so that their processing is not misleading for the person. This occurs through honest and clear communication of the purposes of data processing.

2. **Purpose limitation**: according to this principle, the researcher must clearly define the purposes of his research through appropriate documentation. This principle is attenuated in the research field, as unexpected uses of the data may occur, which must, in any case, be documented in compliance with the regulation.

3. **Data minimization**: just as the purpose of the research must be clarified, the data collected must also be treated in an adequate, relevant, and limited way. In other words, the data collected and managed must not exceed that which can be rationally deduced from the purpose of the research and must be traceable to it.

4. **Accuracy**: according to this principle, all reasonable measures must be taken to protect personal data at every stage of the collection and management process, including data cleaning and consolidation.

5. Storage limitation: personal data must be saved for the shortest possible period, i.e., the period necessary to achieve the set purposes. In the case of research, it is helpful to consider archiving personal data for historical record purposes or future investigations, especially in the context of public and administrative interest studies.

6. **Integrity and Confidentiality**: this principle is a direct consequence and operationalization of the ethical principle described in the previous section and essentially concerns security, i.e., all personal data must be managed in such a way as to ensure its quality and validity as well as protected from possible risks, such as data breach and the consequential loss and/or illicit disclosure of data.

7. **Accountability**: According to this principle, the researcher must take responsibility for what happens to personal data. The GDPR identifies two prominent figures: the data controller, who is legally responsible for the data as its owner, and the data processor, who is technically responsible for data management.

To facilitate the application of the principles, the GDPR also considers data anonymization and pseudonymization procedures. The former is aimed at the complete removal and deletion of any personal data, while the latter is aimed at separating and replacing personal data. As highlighted by [96], talking about anonymization in the era of Big Data is particularly complex as tracing personal data through blending and integration operations is often possible. For this reason, it is often helpful to consider risk minimization techniques, such as the definition of a Data Protection Impact Assessment and the use of tests to evaluate the security of your datasets.

2.2.2.1 A note on copyright and data distribution

Once personal data has been collected, it is usually the intellectual property of the data controller, who can choose to share it based on its purpose. In this case, aspects relating to Copyright come into play. Also, in this case, the legislation is comprehensive [42]. For this thesis, it will be enough for the reader to know that in the Open Data and Open Science fields, Copyright is often managed through the Creative Commons [98] approach, even if it is usually advisable to use licenses more restrictive, to guarantee correct use of the data. In particular, public institutes such as EUROSTAT [99], which deals with "microdata" [100], usually consider the reliability of the entity that intends to use the data, the purpose of use, and the techniques and methods of analysis and management of the data that will be shared.

2.2.3 Privacy across cultures

According to [101][2], "Norms and behaviors regarding private and public realms greatly differ across cultures [. . .], And even within cultures, people differ substantially in how much they care about privacy and what information they treat as private". Indeed, the understanding of privacy varies across cultures and has different positive and negative connotations. For instance, in "Western" cultures, the individual is at the center of attention and considered an atom in a society with atomistic autonomy. This Western philosophical conceptualization of the individual (and individual privacy/autonomy) is informed by the traditions of Greek philosophy, Christian thought, enlightenment, and human rights (especially the firm belief in the dignity of a person [102]). On the other hand, in Asian and African cultures, family and community are more relevant to society than the individual [103]. Individuals are understood as relational to their families and communities, i.e., they are constituted by the "web of social relations" in which they find themselves. Therefore, in many Asian cultures, privacy has traditionally been considered shameful and suspicious, as a claim to privacy seems like a person wants to hide something wrong from the community or society.

Understanding these different philosophical and cultural traditions is important because they inform the process and content of policy-making and data protection regulations in the respective countries. Indeed, concepts of privacy across cultures are changing with intensified global exchange (economic, educational, academic, or political exchange and engagement). Yet, despite some (emerging) similarities between privacy concepts in "Western" and "Eastern" cultures, there are also unique and potentially insurmountable differences in the conceptualization of self and privacy. For instance, social norms might prevent people from claiming their rights and demanding transparency, e.g., because confronting a data controller with such requests is considered rude and aggressive [103], like in the case of many Asian cultures.

Finally, privacy preferences and social norms vary not only across cultures but also within societies and according to situation and context. The information age is characterized by the blurring of lines between public and private spheres, the emergence of semi-public spaces, and the constitution of new and potentially hybrid identities in online and offline worlds. There is much uncertainty about the degree of privacy one can expect; individuals then do not hold clear privacy preferences [101]. Even if they claim that privacy is essential to them, individuals might contradict such beliefs by sharing personal information. In some situations, individuals might be concerned about privacy but feel motivated to share personal information for personal gains (also, when building intimate relationships, sharing private details is seen as necessary). Hence, even within a given culture, people vary in their privacy preferences and performances according to the situation and context [101].

[2] This section is an adaptation of [35]

In conclusion, dealing with privacy in cross-country research is particularly complex. Not only because different regulations need to be understood and adapted but also because of people's social norms and personal attitudes in dealing with privacy, which generate various behaviors and responses in recognizing and enforcing their rights.

2.2.4 Ethics and Privacy support

Ethics Usually, for conducting research, the local institutional review board (IRB) is addressed, whose purpose is to ensure that appropriate steps are taken to protect the rights and welfare of humans participating as subjects in a research study and the procedures required by the IRB to comply with their regulations.

Legal Usually, for conducting research, the local institutional legal office (ILO) is addressed, whose purpose is to ensure that the regulation and the procedures required to be compliant with it are attended to (considering, e.g., the informed consent to be presented to the participants).

In case of the absence of an ILO, an external ILO, such as "ICT-Consulting" can be consulted.

2.3 Data collection tools and approaches

Chapter 1 and Section 2.1 show how data collection can be based on numerous data types concerning a person's daily context. From these, it is clear how the information must be collected over time, encouraging interaction with the person and guaranteeing a high number and granularity of the sensors collected. The relationship between sensors and interaction with the person is fundamental to enriching the context with the annotations necessary to generate Big Thick Data. To better define the relationship between these two data types, the following sections explore (i) available datasets and their features, (ii) the different technologies available for data collection and management and their related activities, namely (iii) pretest, (iv) participant recruitment, and (v) data collection monitoring.

2.3.1 Datasets, sensor data and annotations

Within the domains previously described (see Section 2.1), several datasets are collected based on smartphone and smartwatch data. Table 2.1 shows some of the exemplary datasets available for download and their main features. The tables also show our dataset for comparison. With a glance at the table, it is possible to notice how most of the datasets are collected on a small sample, with few (or low detection rate) sensors data, and with short time coverage (except for StudentLife[104] and Real-life HAR[105]). The lack of the first two features decreases the diversity in observed activities and behaviours. At the same time, the short duration does not allow the observer to recognize the pattern of routine behaviour, which typically takes more than two weeks to train [23] and, consequently, to be observed and identified.

Dataset	Sample	Duration (day)	Annotations	No. of sensors
Real-life HAR [105]	19	28	.	4
MobiAct [106]	57	Trials	.	3
StudentLife [104]	48	70	Sleep-related	10
ExtraSensory [7]	60	7	Activity	8 + 2
ContextLabeler [107]	3	14	Activity	18
ETRI [108]	22	28	Activity, Location, Relation, Mood	10 + 4
SmartUnitn2 [109]	158	28	Activity, Location, Relation, Mood	28

Table 2.1: Public available datasets for activity and context recognition

Considering Table 2.1, the first category of the dataset contains only data from sensors, e.g., [105] [110], which is a significant bottleneck in recognizing human activities, as also pointed out in [111].

Considering annotations, many data collections often consider a minimal, if absent, interaction with the user for collecting run-time labels. Some datasets are collected in a controlled environment from experimental protocols of simple and low-level activities. Examples are the MobiAct [106] datasets and similar datasets such as MobiFall [112], collected to detect falling episodes, or the PAMAP2 [113] and UCI-HAR [114]. However, they are inadequate for analyzing multilateral characteristics of human behaviour that are inherently complex.

Other datasets add a layer of complexity, considering behaviours via surveys and activities through repeated annotations. Datasets collected within the Experience Sampling Method (ESM)[56] framework can be considered part of this category. The ESM approach is worth mentioning for its growing importance in many disciplinary fields and the detailed behavioural observations, even if most datasets are unavailable due to the GDPR and the sensitivity of their content. Other datasets of this type can be found starting from the kind of data collection tool, e.g., the AWARE system [115]. However, sensor data are a corollary of the ESM research as they provide the context for the interpretation of annotations, but not the other way around. And their size and precision are often not sufficient for computational tasks.

On the other hand, there are datasets with *about* sensor annotations, which considers *post-hoc* annotations, i.e., with annotations done after the data collection, and *ad-hoc* annotations, i.e., with annotations done at run-time. The most relevant dataset with post-hoc annotations is [116]. This work is based on time diaries, with annotations generated using the ATUS (American Time Use Surveys) [65] reference ontology. The main weakness is annotations are a posteriori without direct interaction with the person generating them. The main consequence is that these annotations miss the intrinsic truthfulness of a run-time annotation and, consequently, the subjective view of the user, with a substantial increase of the cross-annotation (context-dependent) inconsistency and annotation mistakes (which is known to be very high, see, e.g., [117]).

Examples of datasets with *ad-hoc* annotations are StudentLife [104] (which contains only labels related to sleep), [7], [107]. However, these datasets focus only on user activities and have labels that do not follow reference standards (i.e., social and cognitive bias in data collection are not directly addressed).

For completeness, Table 2.1 also reports the SmartUnitn2 dataset [109], which is the first example of a dataset collected considering both self-reported annotation and sensor data integrated. This approach allows both the use of sensors to learn and validate context annotations (see, e.g., [60, 62]) and annotations enriched with information from sensors (e.g., extending the location annotation via Points of Interest obtained from GPS positions). The approach adopted in the generation of SmartUnitn2 dataset guided the definition of the WeNet pilot called Diversity 1 and described in Chapter 4.

2.3.2 Data collection tools

Numerous tools have been developed for data collection that allow the collection of both data from questionnaires (synchronically or diachronically) and sensors. As regards questionnaire data, the tools often used are LimeSurvey [118] and Qualitrics [119], which also guarantee excellent protection from a privacy point of view (see Section). An extensive list of tools and their features is in [120]. Regarding diachronic and sensor data, [121] shows a rather extensive and complete list of the applications available for the web and smartphones. Some of these worth mentioning are AWARE [115] and m-Path [122] developed by KU Leuven. Both are designed to conduct ESM studies and are constantly updated. The first has already been used in 2,722 studies across institutions worldwide, while the second has 38,000 active users.

A noticeable absence from the previous lists is `iLog app` [123]. [36, 124–126] is a list of publications that describe the use of iLog and iLog collected data in various studies. Currently, iLog runs on Android and iOS devices, with some differences in sensor collection capabilities. The possibility of collecting both user answers and sensor data makes iLog unique (see [104, 127, 128] for a list of other tools currently available). This double facility is important because it improves the state-of-the-art in time diary [64, 129] methods, mainly if structured [130].

2.3.2.1 Tools adaptation

An essential aspect that encourages non-experts to use applications is the user interface. While all tools developed for questionnaires are equipped with a web interface that allows you to configure the questions and send them to participants, most applications for conducting longitudinal surveys do not have this feature. In this sense, the work carried out by CS Logger (adapted from [131]) is particularly innovative, an application that aims to make intensive longitudinal surveys accessible in Citizen Science communities.

2.3.3 Pre-test

Before starting the data collection, running a test to simulate the main aspects and test the technologies is a common practice. The pretest has characteristics similar to the data collection itself. It can be done with a sub-sample of the sample collected in the recruitment phase or conducted and validated by a pool of experts who send feedback on the procedures. An example of such a service is the GESIS Cognitive pretest [72].

2.3.4 Recruitment

There are several approaches to participant recruitment, but the citizen science and crowd-sourcing approaches are more relevant for the purposes of the thesis. According to [132], Citizen Science has some aspects in common with crowd-sourcing, especially involving non-expert people in fulfilling research tasks. Examples are participatory sensing [67] and Mobile Crowd Sensing [133]. They are particularly relevant as they rely on the pervasiveness of smart devices to collect data on large panels, even though based on people often coming from Western countries, which is a central issue for considering the diversity of people. However, crowd-sourcing considers the participant only as a contributor to the data collection [134], but rather than an active community member, as it does not consider involvement in the research process nor education and information projects. On the contrary, projects like [135], [11], or [10] can be considered actual Citizen Science communities, even if their focus is on natural sciences and not on the diversity of people and their behaviour. In contrast, [136] and [137] have a broader focus, even though they are not based on an end-to-end RI.

The participants' recruitment phase can take place in two steps, with two different types of services:

1. Participants attributes (filtering): identification of the main characteristics that the sample must have (for example, age, gender)

2. Contacting: contact of people, which can take place through different channels (e.g., app, mail, call)

Various participant recruitment services like Amazon Mechanical Turk [138] and Prolific [139] provide access to a crowd of potential participants upon payment. [140, 141] present an extensive review of the quality of these platforms, showing how AmazonTurk is among the least reliable of all, as it has a strong selection bias based on a pool of participants experienced in answering surveys. In addition to these platforms, many companies, such as Nielsen or Toluna, are experts in market research (many of these are listed on this website [142]).

2.3.5 Field supervision

Monitoring data collection is essential to ensuring data quality, particularly in longitudinal investigations. Many participants may need to contact the researcher for help or clarification, and active presence in the field can help reduce dropout rates.

Questionnaires are often carried out via telephone contact, particularly in CATI-type interviews.

In the case of intensive longitudinal surveys, however, monitoring requires a constant commitment to evaluating the daily data sent by participants. Usually, this occurs through queries in the database or, in some cases, through the use of an ad-hoc dashboard that allows you to view all the necessary information from a bird's eye view (e.g., amount of data from the sensors sent, number of responses, rates of abandonment). Although this second solution is the most friendly, particularly to facilitate monitoring by non-expert users, most of the apps described in [121] do not provide this function.

2.4 Managing research data

The replicability of results is a fundamental aspect of the scientific process. This is essentially based on the quality of the data [143], which has now become particularly important, especially in the context of Big Data [144]. Although pivotal, few manuals in the social sciences address the data preparation and quality enhancement topics from an operational point of view, relying on external standard practices and technical programming guides (see, e.g., [145]). Furthermore, these manuals often focus on the management of data typical of the social sciences [42], i.e., data coming from qualitative sources (such as transcripts of interview texts) or questionnaires, or even from images and videos collected in the context of structured or laboratory investigations. However, little is said about other data sources, such as those from social media, and even less about the management of sensor data and its preparation for computational analysis and machine learning.

However, the research in the sector is quite advanced. Indeed, [144] provides a good description of the data preparation process, which goes from data analysis to the definition and operation of the data transformation workflow up to its validation, and a list of approaches and methodologies for data processing (e.g., SCARE [146] or KATARA [147]). [148] goes even further, showing how data preparation is not just about cleaning but also about all the procedures to increase the quality of the data, such as the imputation of missing values and the extraction of features. From the service side, [149] has developed a tool based on R and Java for knowledge and discovery databases, industry, and data science, which deals with aspects such as preprocessing and data cleaning, data reduction, and projection but also with verify data quality and cleaning data steps.

An essential aspect rarely taken into consideration is that of data documentation. [42] offers a reliable guide for the entire process. While [150] provides an accurate description of all the metadata that should be associated with the dataset, starting from the information describing the reason why the dataset was created to the actual composition of the dataset (including the description of the variables that compose it), up to the processes used to clean the dataset and instructions and suggestions for future reuse.

Generally, the primary data preparation steps range from (i) data extraction and (ii) transformation to (iii) data curation and cleaning and (iv) data quality improvement, namely adding annotations and imputing missing data. As described in Section 3.4, our approach is based on this primary step. However, since the methodology deals with personal data, privacy operations must be considered to consolidate and anonymize the data. Furthermore, as described in the introduction 1 and in Section 3.1, the peculiarity of our approach is that the domain expert does not create annotations, but the person herself, so that contextual labels are a built-in of our output datasets.

An important note concerns potential errors that may occur in data preparation. [28] in his Total Survey Error approach describes some errors that can occur in the data preparation process, especially for survey data. According to Weisberg's approach, a Big Thick Data methodology should consider the following potential sources of error:

1. **Data management error** That occur in the process of encoding and cleaning the variables, which can be divided into:
   (a) *Coding error* Verbal material is put into numeric categories incorrectly
   (b) *Domain error* Illegitimate data codes appear for a variable
2. **Imputation error** That occur in the data processing and aggregation phase (for example, for the anonymization purposes described in Section 3.4), which can be divided into:
   (a) *Estimation error* Incorrect weighting or aggregation of data

   (b) *Adjustment error* Wrong cleaning or statistical techniques are employed

3. **Participants error** that happen either on the part of the participant (see Section 2.1) or in interaction with the participant, which can be divided into:
   (a) *Completeness error* Surveys are not filled out fully

   (b) *Skip-pattern error* Instructions for skipping questions were not followed

   (c) *Consistency error* Answers to different questions are not logically compatible

4. **Reporting error** Survey results are presented incorrectly to an audience, or there is an inconsistency in the dataset documentation (e.g., variables have different names in the dataset, report, codebook, and additional materials)

The next sections consider a set of services currently available for the researcher who wishes to prepare data (in addition to the various data management programs such as Python, R, SPSS, and Stata), even though, according to [151], most data cleansing tools may have limitations in usability due to:

1. Project costs: costs typically in the hundreds of thousands of dollars
2. Time: mastering large-scale data-cleansing software is time-consuming
3. Security: cross-validation requires sharing information, giving application access across systems, including sensitive legacy systems

2.4.1 Overall data preparation

In addition to the tools proposed by [149], which are developed in the research field, there are some automated cleaning tools and software offered to the researcher, like Google Open, refine [152] and IBM InfoSphere QualityStage [153]. Automated cleaning also includes algorithms and programming codes (see section 2.4.3).

Furthermore, there are consulting companies and companies specializing in data cleaning.

2.4.2 Documentation

Data documentation is essential to ensure correct reuse and facilitate replicability. There is a set of documentation concerning the description of the data collection process and outcome (also called data descriptor), the descriptive metadata (identification, interpretation, authentication, finding, i.e., data about the original context of the documents), and the description of the collected data (codebook). This section does not consider another type of metadata, namely Administrative or management metadata (authorization, logistic data, ownership, formal origin, accountability of management activities, i.e., data for the archive system) and

Technical metadata (software, hardware, storage format, i.e., data of the system with which the information is created and managed).

Generally, packages that offer codebooks also do metadata, such as, for instance, Python Pandas Profiling [154], and the R package called dataMaid [155].

2.4.3 Improving data quality

Data quality can be improved through additional services concerning data cleaning, formatting (especially if textual data are considered), and imputation. There are several algorithms for imputing data, although they are often superficial. The following are some examples of data quality improvement tools, packages, and libraries:

- [156] describes a list of R Packages for handling missing data
- The Deequ Library from AWS built on top of Apache Spark [157]
- Great Expectations [158] which is a specialized data quality tool
- Griffin Apache [159] which is an Open Source data quality solution for Big Data

2.5 Data sharing approaches and platforms

As stated at the beginning of this chapter, the replicability of research is the fundamental principle on which the scientific method is based, which passes through the reproducibility of the findings and data collected. This is also described by UNESCO [160] as recommendations for Open Science.

In this sense, data distribution is one of the pillars of science and, in our case, a fundamental aspect of validating the adopted method. As demonstrated in Chapter 1, the approach to Big Thick Data is eminently interdisciplinary because it adopts theories, guidelines, technologies and standards from different disciplines in the data generation process. To guarantee that the output of the process is valid, i.e. that it measures the theoretical concepts to which they refer, the data collected must be able to be reused in different disciplinary fields for the production of reliable results and, ultimately, reliable scientific literature (e.g., published in indexed journals, validated through a peer review process, etc.).

There are several principles for making data accessible, two of which are particularly important. The first is the FAIR principle [161], and the second is the guidelines promoted by Tim Berners Lee to facilitate data distribution on the Web.

According to the FAIR principle [162], data must be:

1. **Findable**: The first step in (re)using data is to find them. Metadata should be accessible to both humans and computers. Machine-readable metadata

are essential for automatically discovering datasets and services, a critical component in the FAIR process.

2. **Accessible**: Once the user finds the required data, she/he/they need to know how they can be accessed, possibly including authentication and authorization.

3. **Interoperable**: The data usually needs to be integrated with other data. In addition, the data needs to interoperate with applications or workflows for analysis, storage, and processing.

4. **Reusable**: FAIR's ultimate goal is to optimize data reuse. To achieve this, metadata and data should be well-described so they can be replicated and combined in different settings.

[163] organizes the accessibility of the data on a 5-star score, as reported below:

⋆ make data available on the Web under an open license

⋆⋆ make data available as structured data (e.g., Excel instead of an image scan of a table)

⋆⋆⋆ make data available in a non-proprietary open format (e.g., CSV instead of Excel)

⋆⋆⋆⋆ use URIs to denote data so that they can be cited

⋆⋆⋆⋆⋆ link data to other data, using, e.g., RDF so that concepts that are part of an ontology can be defined using OWL) to provide context

To facilitate the dissemination of data, many (i) platforms have been developed, which aim to provide the services of data (ii) finding and (iii) storage (and Download), as well as journals specialized in (iv) dissemination of datasets, as described below.

2.5.1 Open Data platforms for data distribution

The disciplines that deal with personal data often develop research infrastructures (RI) or platforms to support researchers in data management and sharing. For instance, within social sciences, [80] and [72] provide support for ethics assessment and data management, and, alongside [81], they enable the documentation and distribution of high-quality survey data. Data distribution is particularly advanced in the healthcare sector, with leading RIs such as [164] or [165], which also have played a key role in managing the COVID-19 pandemic.

However, despite the apparent support provided by such infrastructures, they are not end-to-end. Furthermore, they remain tied to individual research communities, not favouring effective interdisciplinary exchange.

2.5.2 Finding

In addition to search engines, there are several specific repositories in which it is possible to identify datasets according to the topic and features. Some examples are Google Dataset Search [166], re3data [167], and Github [168].

2.5.3 Storage

Several companies provide data storage (and finding) services, like *Zenodo* or *fighsare*. An exhaustive list can be found on the Nature website [169].

2.5.4 Dissemination

Dissemination refers to the drafting and publication of scientific articles concerning the dataset. While this is the researcher's job, advice and support can be provided to write the articles. In addition, the leading journals that publish datasets can be indicated, such as conferences with Resource Track, like Nature Scientific Data [170], and Data in Brief [171].

3

The iLog methodology

Contents

As described in Chapter 1, a Thick approach to Big Data favors a more accurate representation of the context of the person immersed in daily life. This broadens the research potential for all those disciplines that aim to investigate aspects relating to the Object, Personal, and Reference Contexts (see 2.1.1 for an extensive description). In the introduction, we also showed how it is possible to quantify the notion of Big Thick Data through Situational Context, which identifies the fundamental entities of a person's context, enabling their modeling over time through knowledge graphs. However, this operationalization lacks the essential aspects to generate the data that make up the context, i.e., to design the interactions with the person and the sensors, collect the data, and manage them, encouraging interdisciplinary reuse.

Chapter 2 shows how, although different approaches and methodologies address one or more aspects of the generation of Big Thick Data, there is no end-to-end methodology nor, much less, a set of services and operational procedures that enable its reuse and replicability for non-experts. The latter is an essential aspect, as both the topic and the process of generating Big Thick Data are eminently interdisciplinary, involving aspects of social sciences and HCI for the design of interactions with the person, passing through aspects of IT for the configuration and management of data collection, up to ethical and privacy aspects. Making the methodology suitable even for non-experts means that any researcher interested in Big Thick Data can replicate it.

In a nutshell, to collect Big Thick Data through the notions of situational context, a methodology must be developed that starts from the systematic approach of the social sciences and adapts it to the possibilities arising from new technologies. Furthermore, Big Thick Data is an eminently interdisciplinary notion. For this reason, the methodology must apply to different practitioners.

These are the reasons behind the development of the iLog methodology, which, in addition to the operationalization of the concepts described in the state-of-the-art, will present a section dedicated to the services implemented to encourage their reuse (see Appendixes from A to E). Furthermore, each section will present a Business Process Model (BPMN), which characterizes the main activities to be conducted in each phase (for a definition of BPMN and its components see the List of Acronyms and Definitions). Each phase is designed as a module so researchers can access the different services without necessarily following all the phases. As anticipated in Chapter 2, the phases of the methodology are:

1. *Design*: considers all the aspects to prepare a Big Thick Data collection.
2. *Ethics and Privacy*: covers the aspects to handle all necessary bureaucracy to comply with proper research analysis, ethical considerations, and privacy legislation.
3. *Data Collection* (iLog Platform): concerns the process for collecting the Big Thick Data.
4. *Data Preparation*: covers the aspects of cleaning, minimizing, and preparing the collected data to facilitate the data analysis.
5. *Data Distribution*: considers the main features of data sharing, i.e., metadata creation, searching, upload, and download of datasets.

Figure 3.1 shows the overall process, while the following sections will provide the details for each phase.

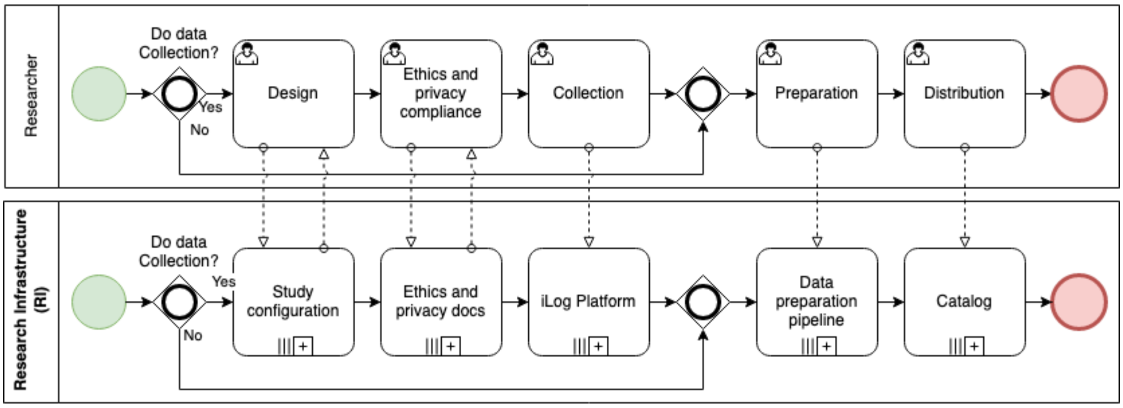


Figure 3.1: The overall iLog methodology BPMN

3.1 Design

The design phase corresponds to the general strategy defined to ensure that the research problem is answered, considering the different components of the process. Section 2.1 showed how, usually, the study design phase ranges from (i) the definition of the research topic to (ii) the structure of the study, to (iii) the measures to be used, (iv) the sampling strategies, (v) the adaptation of the survey, up to (vi) planning. Aiming to make the methodology easily reusable, design aspects are supported by the materials in Appendix A.1, while the last section describes (vii) the BPMN for designing a study.

3.1.1 Topic and objectives

A topic is a subject or issue that the practitioner is interested in. In the case of the iLog methodology, it is designed to respond to a wide variety of research interests that concern the person in her context, ranging from themes that concern the *Object context*, the *Personal context*, the *Reference context*, or transversal approaches (i.e., multipurpose) which focus on the multiple aspects that make up the person's daily routine. According to Chapter 1 and Section 2.1.1, *Object context* encompasses all the studies that aim at recognizing people's activities through sensors and devices, while *Personal context* studies involve direct interaction with people, aiming at recognizing the person's point of view within their everyday life. Finally, *Reference context* studies involve the participant as a "sensor" to gather quality information about the surrounding environment.

The choice of the topic is an iterative process that is refined during the different phases of the design. Therefore it does not only depend on the interest, but also on the objectives, the resources available, and the state of the art. It is not always easy to anticipate all the aspects that depend on the choice of topic. For this reason, in addition to the literature specific to each study field, it is possible to consult the LivePeople Catalog [172], which presents examples of different studies about context, as described in Table 3.1.

Project	**Acronym**	**Topic**	**Focus**
SmartUnitn2 [109]	SU2	Personal context	Students habits and routines
Qrowd [173]	QR19	Reference context	Parking and bike racks
Diversity1 [174]	DV1	Personal context	Multi-purpose
Chat application 1 & 2 [175, 176]	CH1 & CH2	Social context	Interactions in chat application

Table 3.1: Big Thick Data collections and topics

Finally, the topic has the main purpose of being a guide in the design choices of the study, as highlighted in the following sections.

3.1.2 Study structure

The study structure is the global organization of the investigation, from which it derives the protocol, the sampling strategies, and the adopted measures. Section 2.1.2 has shown the different study structures. Since the iLog methodology has the main aim of observing people in their daily lives through streaming of granular information, the most relevant study structures are experimental and quasi-experiment, and intensive longitudinal surveys. Unlike synchronic surveys, i.e. which take place in a single moment (e.g., cross-sectional surveys), these structures consider the evolution of actions and activities or perceptions of self and surroundings, both in a controlled environment and in the wild.

Considering *Object context*, experiment or quasi-experiment is the most common choice since the objective is to isolate and recognize specific patterns of actions, controlling the influence of external phenomena. A typical case is HAR studies involving visual stimuli [177], where both photographs and videos are recorded in a specific setting (e.g., a living room [178]), not only due to the availability and costs of the detection technology, but also to exclude a set of variables typical of daily life, but which could complicate the cleaning and preparation of the dataset or interfere with the models you intend to create. However, with the advancement of technologies and the ease of obtaining sensors (e.g., accelerometers present on smartphones and smartwatches), many datasets are collected in the wild (see, e.g., the datasets presented in [179]).

Personal context have a propensity for intensive longitudinal surveys, as they favor interaction with participants in the wild. Examples are [8, 180] in [53] where participants received random questions on aspects of psycho-physical health during waking hours. There are different variations in the duration of the investigations and the stimuli used. *Interoception* studies will prefer psychometric stimuli sent at a regular interval and a low frequency, unlike *Multi-purpose* ones which aim to represent in detail the individual events of daily life (see, e.g., [63]). On the contrary, similar to Object context studies, *Human-AI interaction* studies prefer a controlled environment, often with an A/B testing approach. There are differences depending on the topics of investigation and objectives. Again considering [180], an experimental and a control group was planned, with a field experiment approach. In contrast, recently, studies of *Human-AI interaction* in the wild have been conducted (see, e.g. [181]).

Generally, in *Reference context* participation is "free", in the sense that it is not defined at a regular frequency of stimuli sent to the person. On the contrary, participants are invited to describe and catalog aspects of the reference context at the moments they prefer (see, e.g., [182]). Even in this case, however, there are studies with a similar approach to intensive longitudinal surveys where participants receive daily notifications based on their location (see, e.g., [183]).

The structure of the study depends on its overall duration, as presented in the following paragraph.

3.1.2.1 Study duration

The duration of the study is strongly correlated with the intensity of the measurements, i.e. with the frequency of the stimuli used and, consequently, the respective respondent burden. In addition, the practitioner should also consider the type of detection tools used and the amount of data needed for the analyses. For example, a *Object context* study on sports activities [184] often has a concise duration (e.g. a few hours) as high-intensity active participation is required from the participants involved, as well as the use of different data collection instruments. On the contrary, studies on *Personal context* often last longer, from a few days to several months (see, e.g., [185], which lasted 239 days) as the intensity of the phenomenon studied and the respective frequency of stimuli and involvement of the participants is lower. Likewise, studies on the *Reference context* can have longer durations, especially if participation is free and there is no use of the stimuli sent regularly.

3.1.3 Measurement and tools

Section 2.1.3 describes two types of measurements: active and passive. Active measurements involve direct interaction with the participant and are commonly used in *Personal context* and *Reference context* studies. Their opposite is passive measurements, typical of *Object context* studies, which involve automatically collected data, such as data from the smartphone sensor.

Practitioners can use different tools to collect these types of measures. Tools are the set of objects that convey the measurement and gather the data. A question can be asked orally, via pen and paper, or via software for conducting online surveys, such as LimeSurvey [118] or Qualitrics [119], which also allows the collection of some passive data (called para-data), such as the timing, the duration of the survey, or the IP address from which it was done. However, a question can also be asked through a data collection application like iLog, which also can collect sensor data and provide multiple other stimuli, whether they are collected from the internet (e.g., a link to a video on YouTube) or a photo taken and annotated via the app. Additional tools such as video cameras or data from secondary sources can be used to refine observations and expand analysis possibilities.

3.1.3.1 Measurement and tools reliability

Measurements are always subject to reliability problems whether they are active or passive. Active or self-reported measurements have often been questioned in the scientific literature. Section 2.1.3.1 describes some of the issues related to active measurement, like *Hawthorne effect* [69], or the alteration of behavior by the subjects of a study due to their awareness of being observed. This can happen due to *social desirability* [27], that is, trying to direct the perception of behavior towards aspects considered more positive; or *non-attitudes*, i.e., when the respondent knows little about the topic on which the question is asked. Moreover, respondent burden [70] effect may lead to a response set or missing answers.

For these reasons, with the advent of new data collection technologies (e.g., smartphone applications), studies rely more on "objective" measures of behavior and attitudes, such as sensor data. However, the fact that they are more objective does not exempt them from problems similar to self-reports. For example, in studies conducted via smartphone, participants may actively choose to silence notifications, turn off the cell phone, or turn off specific sensors (such as location or Bluetooth) for logistical and privacy reasons.

All the errors relating to the measuring instrument are linked to the reliability problem, which in this case correspond to technological problems. In the case of a smartphone app, these may involve problems with:

1. *Server*: i.e., the device that provides the service of creating, sending, and receiving active and passive data, which may crash during the data collection, losing some information
2. *Notification sending system*: i.e., the service that sends notifications from the Server to the smartphone, which may not work properly in all cases, losing some notifications
3. *Phone operating system*: which may not work properly with the app used, resulting in loss of notifications and information
4. *Phone hardware*: which may not have all the sensors needed for data collection

All these aspects can influence the results' reliability and should be considered in the study's design. The following paragraphs provide some configuration details of the measurement instruments.

3.1.3.2 Active data definition

Section 2.1.3.2 shows how "copying questions from other questionnaires is not plagiarism" but instead a "recommended practice, in that it enables knowledge to be accumulated and comparisons to be made" [27], which reduces the possibility of incurring systematic errors [28]. Therefore, as a general guideline for the design of questions and stimuli, it is always good practice to rely on previous literature. In this sense, the LivePeople Catalog [172] offers the possibility of exploring different configurations of questions.

As regards active measurement of the *Personal context*, the Diversity 1 study (see also Chapter 4) shows how it is possible to ask questions about people's daily lives through the use of the HETUS [66] standard asked every half hour. The standard comprises three questions (where are you, who are you with, and what are you doing) and a set of possible answers designed to map the main aspects of daily life. It also considers aspects of *Interoception* through questions on Mood, always asked every half hour, as well as aspects relating to eating behavior, asked every two hours intending to observe both the food and drinks consumed during main meals than those consumed between meals. On the other hand, studies [175, 176] focus on aspects of social interaction, through the creation of chat applications that allow participants to interact on the issues of daily life in which they are most interested. Other sources of questions can be found in GESIS "Item and Questions" [72] and ESM repository [73], in particular regarding the in-depth analysis of aspects relating to interoception.

As regards the *Reference context*, the study [183] (available in the LivePeople Catalog) shows different ways of interacting with the participants with the aim of photographing, tagging, and mapping the facilities present in the area. Other sources of measurement can be found in major citizen science platforms, such as Zooniverse [135], iNaturalist [11], or SciStarter [136].

The active measurements available in the literature do not always reflect the practitioner's interest. In these cases, it is possible to adapt the available measurements or create new ones. In this case, several manuals guide the process of designing questions (see, e.g., [47]), as well as a set of guidelines and practical advice. For instance, [27] highlights how the formulation of the question can influence the response and provides a list of 21 aspects to consider considering the language, syntax, and content of the questions, such as:

1. *Simplicity of language*, namely avoiding the use of technicalities and unusual words
2. *Question length*, since too long questions will induce the participants to skip them or spend too much time reading, increasing the respondent burden
3. *Questions with non-univocal answers*, which is particularly important in closed questions since it will induce the participants to answer randomly
4. *Presumed behavior*, or posing a question assuming that a set of behaviors are normal or common for the culture under study

Timing As shown in Section 2.1.3, [75] has identified different questions based on time. Following this approach, it is essential to consider the time frame in which the question is asked, which must be found in the type of survey you intend to do. A second important aspect is the frequency with which questions are asked. In this case, it is helpful to consider both the respondent burden - that is, trying to ask questions at a limited frequency - and the duration of the study. Depending on the survey length and the number of questions, it may be helpful to add break options to lighten the participant response burden. Please note that in iLog, the ability to silence notifications is always available.

Similarly, in studies that plan to delve into aspects of social interaction - for example, via chat - it is helpful to consider features such as the possibility of muting or not receiving notifications so as not to overload the participant.

3.1.3.3 Passive data definition

As mentioned in Sections 2.1.3.3 and 2.3, passive measurements are all the information about a person that can be collected without direct interaction. There are different types of sensors and tools for passive data collection, from video cameras to various home automation and IoT sensors. Below is information that specifically concerns sensors collectible via iLog [1]. In particular, there are several criteria to consider when defining sensors, namely:

1. *Available technologies*: how many and which devices are needed for data collection. For example, the iLog application can collect all the smartphone sensors. Still, in addition to this, it may be necessary to use other types of technologies, such as weather stations or devices, home automation, and IoT, depending on the studio's purpose.

2. *Devices capabilities*: this is particularly true for smartphones, as they are subject to different operating system configurations and wear and tear. Considering the operating system, iLog can collect all the sensors of Android smartphones. Still, it does not work with Huawei devices (as they do not use the Google services used by iLog), and problems have been found with some other devices, such as Xiaomi, while in iOS, not all sensors are available (see next section). Regarding wear, some functions or the age of the device can influence the correct functioning of the device, in particular on the battery life. iLog is tested to collect all sensors in conjunction with daily queries and consume 7% of the battery, but this may increase in older devices or with different configurations.

3. *Interaction with devices*: some sensors and devices are simple to use as they are now part of many people's daily lives. For example, almost everyone knows how to install an application or turn location tracking on and off on their smartphone. However, depending on the type of device and the population under study, it may be necessary to select a reduced number of sensors or prepare instructions and user manuals if more complex activities are envisaged.

4. *Storage capacity*: an essential aspect in selecting sensors is the availability of server space to save and manage them. Data collection involving high-frequency sensors such as the accelerometer can quickly reach a terabyte of data in a month.

[1] For a complete list of the sensors available in iLog, see Appendix C.1

5. *Number and type of sensors*: sensor selection strongly depends on the study purpose and the technologies, devices, and storage available, but also on the expertise and tools to manage them, in particular in the case of Motion and Environment sensors. There is no precise rule to decide which sensors collect, but GPS location is usually collected due to its versatility.

6. *Sensor frequency*: it depends on the observation's granularity that needs to be collected. In general, for Object context tasks, it is suggested that sensors be collected at the highest frequency possible to observe each person's movement and compensate for collection errors. For other purposes, the frequency can often be reduced, up to collecting the sensor information only for particular events (e.g., concurrently to an answer to a notification).

3.1.3.4 Other data sources

Before starting data collection, it is helpful to consider the presence of data sources that are useful for the research, namely not only additional primary data, i.e., collected directly by the researcher, but also secondary, i.e., downloaded from pre-existing databases. In addition to the datasets already present in the LivePeople Catalog, the researcher may be interested in considering Open Data resources such as OpenStreetMap to integrate the information collected with GPS, for example, for labeling points of interest visited by the participants or the European Open Data portal [2].

3.1.3.5 Devices, applications and user experience

Depending on the purposes of the study and the data to be collected, the researcher may have to select multiple applications and devices for data collection. There are devices and applications specialized in collecting specific data. Although iLog provides a wide variety of collectible data through smartphones, from questionnaires to intensive longitudinal surveys to sensors, it may be necessary to integrate the application with online tools for questionnaires, which allow them to be viewed on a computer screen or which have implemented and optimized aspects of question formulation and visualization (e.g., multiple conditional questions), or with specialized applications in messaging and chat services, or with other devices and external sensors, such as smartwatches and IoT technologies. In this case, it is good practice to try to minimize the technologies in use and, if necessary, limit the switching between different technologies. Therefore, it can be helpful to define a user journey, clarifying for each phase which devices will be used, how the participant will obtain the information to access them, and how to use them.

[2] https://data.europa.eu/en

3.1.4 Sample design

Sample selection was described in Section 2.1.4. From the section, it is clear how useful it is to identify:

1. Target population: a sub-sample of the total population that has a set of characteristics or factors of interest for the study
2. Sampling scheme: a proportional or nonproportional approach to sampling
3. Sample size: This varies depending on the disciplines and objectives, but recently, some communities and scientific journals expect a minimum of 150 participants.

In addition, it is always advisable to recruit at least 30% more participants due to the numerous problems that can occur during the registration and management of data collection and due to the dropout of participants, whose rate is between 30 and 70% [186].

3.1.4.1 Recruitment strategy

As seen in Section 2.1.4, many sampling errors can occur during the recruitment phase. The recruitment strategy should be planned to avoid this error, considering different aspects, such as how the communication with the participants will happen and the kind of support materials that should be created.

Communication can occur online, for example, via email, social channels, or platforms to which users are registered, or offline, via flyers, word of mouth, or events presenting the data collection campaign. Particularly in the case of probability samples, the selection of people and communication must be standardized. Regarding materials, intensive longitudinal surveys often present tasks that are not intuitive. Therefore, in addition to standard communication materials such as email texts, slides, and contact scripts, it is helpful to create a short, easy-to-use instruction manual that the participant can download online and use whenever they have doubts (e.g., see Appendix C.2).

3.1.4.2 Incentives strategy

As described in Section 2.1.4.1, [78] has provided an extensive literature review on the type of incentives (monetary or non-monetary) showing how this can reduce respondent bias presenting helpful suggestions for research practitioners, integrated into this section.

In the case of monetary incentives, the participant will receive compensation at the end of the data collection, be it in money, vouchers, or other types of objects. In this case, if not necessary, it is valid to specify the terms within which the participant will be rewarded in addition to the timing. For example, in the case of studies that involve questions repeated daily via applications, it is helpful to remind the participant that it is not sufficient to install the application to receive compensation but that it is necessary to answer a certain number of questions. Drawing prizes for the most active participants is an excellent tool to incentivize participation and data quality. In general, the amount of incentives usually depends on the tasks the participants need to do. In the case of Diversity 1 study conducted with iLog, the effort to respond to a notification is cognitively minimal (there are five very simple closed-ended questions) and takes about 10 seconds, so for around 4 hours of commitment over 2 weeks, the study paid €20 (i.e. €5 per hour) plus bonuses and prizes. For the response to a questionnaire of medium duration (about 10-15 minutes), which requires greater effort than notifications, a platform such as Prolific pays between 2 and 4 euros.

Non-monetary incentives can involve various aspects aimed at arousing interest in the research project or encouraging and supporting the participant during the multiple phases of the study. Regarding interest, it is helpful to reflect on the user experience of apps, even if it is often challenging to make answering questions engaging daily. For this reason, it is beneficial to define good communication of the research project, the objectives it intends to achieve, and how people will benefit from the results in the future. In this regard, thinking about several intermediate or final presentations showing the research results is also helpful.

Regarding reminders and support, the fundamental aspect is to have a constantly active helpdesk function that responds promptly to questions asked by users. The helpdesk can also be designed proactively, i.e., starting from the daily results, it can choose to contact the less active participants to provide support. In this case, it is helpful to define the messages that will be sent in agreement with an ethics committee to avoid privacy problems and excessive intrusion. Furthermore, it may be beneficial to define a set of automatic messages that give the participant a perception of how the data collection and their contribution are progressing or stimulate them to provide contributions more regularly. This can be accompanied by forms of gamification, for example, through a badge system that can be obtained as contributions grow.

3.1.5 Cross-cultural studies

Cross-cultural studies are particularly complex because of the adaptation procedures, which involve not only the translation of stimuli but also the understanding of cultural norms and values. In case of a need for translation, it is helpful to adopt the standard procedure in social science, namely to find three experts who know both the language in which the study is designed and the local language. Two experts will independently translate all the stimuli, while the third expert will revise both translations and finalize the stimuli in the local language. Therefore, in case of disagreement between the translations of the two experts, the third party will define the best solution.

Cross-cultural studies may also involve the adaptation of the tool. For example, iLog is based on Google services (which doesn't work in China) for sending notifications. In this case, the solution was to use different services and adopt "offline" time diaries directly uploaded on the app.

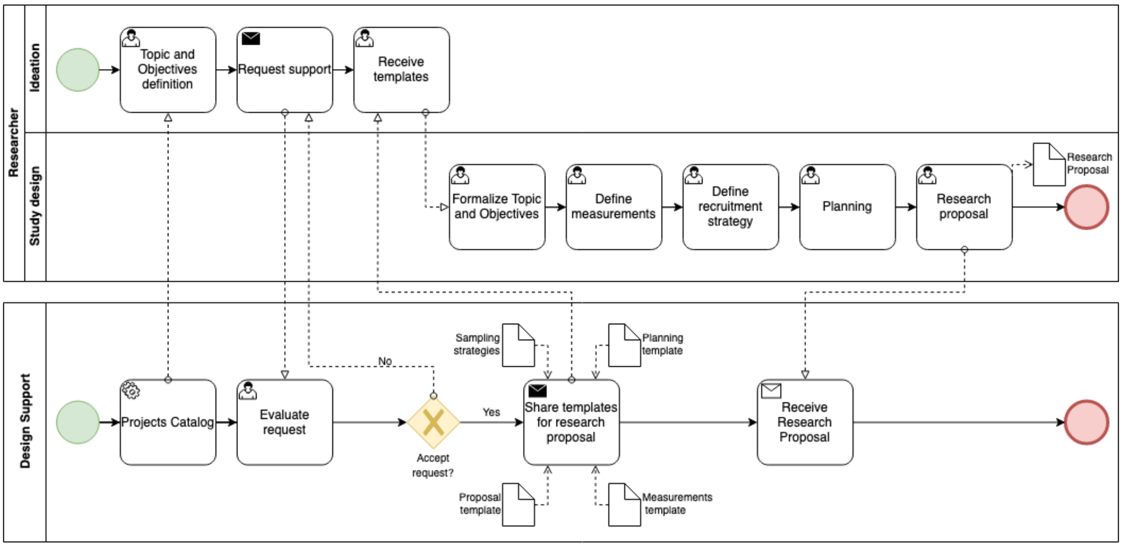
3.1.6 Planning

Once the research design has been concluded, it is helpful to plan the subsequent phases, taking into consideration the description of the processes in Sections 3.2, 3.3, 3.4, and 3.5. It may be helpful to define a Gantt for the entire process (and a data management plan in more complex cases) by identifying different task roles. Usually, there are three leading roles:

1. Research leader: The role of the research leader is to oversee the project's life cycle, focusing on the research design and management.
2. Study leader: The role of the study leader is to oversee the study, especially concerning the sample and personnel design, which also includes designing incentives and recruitment strategies while ensuring the overall compliance of the experiment with privacy and ethical regulations. Furthermore, it also manages the actual running of the experiment.
3. Technology leader: The role of the technology leader is to oversee the technological development of the platform by collaborating with the Study leader to meet the requirements of the study, especially concerning the iLog app configuration and the helpdesk.

3.1.7 Design BPMN

Figure 3.2: Design phase BPMN



The design of a study is the least automatable aspect of data collection as its implementation strongly depends on the investigation objectives defined in the researcher's purpose and shared practices that vary depending on the scientific community. Nonetheless, Figure 3.2 shows a BPMN that highlights the main activities related to support services, while tables 3.2,3.3,3.4,3.5,3.6 describe the various components of the model.

Regarding timing, the sub-process involves several unknowns, such as the choice of measuring instruments, which, if created from scratch, can require several months of work. For this reason, the times of the different tasks have not been indicated. Still, it is assumed that the process can last from an elapsed time of two weeks in the case of almost total replication of studies already done up to about six months in the case of defining a completely new study.

Name	Type	Contained in	Description
Researcher	Pool	-	The user of the system that creates and runs a study
Design Support	Pool	-	The main service for supporting the researcher in the design of a study
Ideation	Lane	Researcher	The sub-entities of the researcher who are responsible for defining a topic of interest for the study
Study design	Lane	Researcher	The sub-entities of the researcher who are responsible for configuring the study towards the actual running

Table 3.2: Design pools and lanes

Name	Type	Definition	Description
Researcher.Begin	Start	Start a study	When a researcher decides to initiate a data collection study
Researcher.End	End	Research proposal submission	When the researcher submits the research proposal
DesignSupport.Begin	Start	Catalog and Support	Through the catalog showing the study resources and after the researcher requests support.
DesignSupport.End	End	Receive research proposal	When the proposal is submitted

Table 3.3: Design events

Name	Type	Description	Timing
Topic and Objective definition	User	Looking at the catalog and the state of the art, the researcher defines a topic of interest for the study	.
Request support	Send	Process for sending a request to Design Support defining the study objectives	.
Receive templates	Receive	Process for receiving the materials for setting up the study design	.
Formalize the study structure	User	Process for choosing which type of structure is more suitable for the study	.
Define Measurements	User	Process for choosing the set of stimuli and sensors to collect, and the timing of the data collection	.
Define recruitment strategy	User	Process for defining the number of participants, how to contact them, and the type of incentives	.
Planning	User	Process for defining the roles and the data collection and management Gantt	.
Research proposal	User	Process for writing and submitting the proposal to Design Support	.

Table 3.4: Researcher - Design tasks

The Researcher Lane (see Figure 3.2 and Table 3.4) shows the main tasks to be conducted by the researcher, following the approach described in this section. The design phase will begin with the ideation, supported by state-of-the-art studies and examples present in the LivePeople Catalog. Once the topic and the research objectives have been identified and the need to continue with an investigation following this methodology and using the relevant tools, the researcher will be able to contact the Design Support who will provide the necessary material for the design of the study. By following the templates, the researcher will be able to define the research proposal that should state a clear definition of the (i) type of study, (ii) the data that will be collected, (iii) the timing, (iv) the sample involved, and the recruitment strategy, and (v) a plan of the data collection and management

Name	Type	Description	Timing
Projects Catalog	Service	Showcases previous data collections design and data	.
Evaluate request	User	Process for evaluating the researcher proposal	.
Share templates for research proposals	Send	Process for sending the necessary support material for designing a study	.
Receive research proposal	Receive	Process for the complete research proposal	.

Table 3.5: Design Support - Design tasks

Name	Type	Direction	Description
Exclusive Gateway 1	Exclusive	Diverging	Design Support Gateway for checking the research support request. Options are Accept and Reject

Table 3.6: Design Gateways

activities.

The Design Support Lane (see Figure 3.2 and Table 3.5) shows the main tasks conducted by the support service. In particular, once the request for documentation has been received, the support service will evaluate it and then send the necessary templates to the researcher. Different support materials have been developed, considering the whole study design phase (see Appendix A) and the definition of questions and sensor data to be collected (see Appendix A.2). Once the supporting material has been submitted, a consultation will be provided. The activity will be concluded with the receipt of the proposal to move on to the subsequent phases.

3.2 Ethics and legal aspects

As discussed in Section 2.2, an assessment of the research design is fundamental to guarantee the correct application of the various ethical principles (i.e., informed consent, confidentiality, integrity, avoiding harm and doing good) and to ensure that the legislative bases concerning the data collection and management are respected.

The following sections describe the ethical and privacy guidelines, considering the GDPR [94] as a legal basis. In particular, they cover aspects regarding (i) The need for an institutional revisory board (IRB, also called ethics committee or ethics board) assessment; (ii) Personnel, responsibility, and conflict of interest; (iii) Project purpose, objective, and expected results; (iv) Participants burden; (v) Participants selection criteria; (vii) Informed consent; (viii) Data management, safety measure, and risk evaluation; (ix) and, the Privacy in cross-country research. Finally, a BPMN will be presented with various supporting materials, to guide the researcher through the ethical and privacy assessment process and the preparation of the relevant documentation.

3.2.1 The need for an IRB assessment

Generally, any study that directly involves people requires approval from an IRB. Following the approach defined by [187], there are some criteria to identify the need for an IRB assessment. In particular, it is necessary to request a consultation if:

1. The project involves the collection of new information about individuals
2. the project compels individuals to provide information about themselves
3. the information about individuals is disclosed to organizations or people
4. the project involves new technology that might be perceived as being privacy intrusive (e.g., the use of biometrics or facial recognition)
5. the project results may involve decisions or actions against individuals in ways that can have a significant impact on them
6. the information about individuals may raise privacy concerns or expectations (e.g., health records, criminal records, or other information that people would consider to be private)
7. the project requires to contact individuals in ways that they may find intrusive

The following sections describe the main components of the IRB assessment documentation.

3.2.2 Personnel, responsibility, and conflict of interest

The first aspect to consider is the identification of the personnel who will participate in the research. According to the GDPR definition, there are two roles: the Data Controller and the Data Processor (see Section 3.1.6).

The data controller determines the purposes for and how personal data is processed. In the case of research, the data controller is also the data owner. For this reason, this legal responsibility is generally given to the research institute, represented by the Data Protection Officer (DPO). In contrast, managerial responsibility is given to the Project Manager, i.e., the person who defined the purpose of the data collection. In our methodology, the data controller generally coincides with the figure of the Research leader (see Section 3.1.6).

The data processor processes personal data only on behalf of the controller. Given the size of the data and the resources in the research field, often the data controller and the data processor are embodied by the same person or members of the same research team. However, there is the possibility that the data processing is entrusted to a third party. In our methodology, the data processor generally coincides with the figure of the Technology leader (see Section 3.1.6).

The processor's duties towards the controller must be specified in a contract or another legal act, usually called a Data Processor Agreement. A template of the contract can be found in the Appendix B.5. For example, the contract must indicate what happens to the personal data once the contract is terminated. A typical activity of processors is offering IT solutions, including cloud storage. The data processor may only sub-contract a part of its task to another processor or appoint a joint processor when it has received prior written authorization from the data controller.

An essential aspect in selecting personnel, whether controller or processor, is the absence of conflict of interests, here understood as personal interests (i.e., non-scientific) deriving from the reuse of data or the products of their analysis.

3.2.3 Project purpose, objective and expected results

The purpose, objectives and expected results should derive from the information defined in the design phase. From an ethical point of view, it may be helpful to consider two aspects, namely scientific relevance and public interest. Scientific significance is the impact the study would have in its community of reference, i.e., how much the study results could broaden knowledge and advance understanding of the subject and how much the scientific community accepts them. Generally, scientific research is of public interest, namely how impactful the study can be for the well-being of a population, both directly and indirectly.

3.2.4 Participants burden

As described in Section 2.1 and specified in operational terms in Section 3.1.3, the respondent burden is an essential aspect of research based on intensive longitudinal survey methods. From an ethical and privacy point of view, the survey tools used mustn't put participants at risk from a personal and cognitive point of view. Therefore, the survey should not subject people to excessive stress, e.g., by asking a disproportionate number of questions with too high frequency or by presenting stimuli that are too different or can create too much discomfort. This can happen even in investigations not directly dealing with sensitive information. For instance, through a survey on activities carried out in daily life, a participant might realize that they spend too much time on leisure activities. In this case, the minimization principle described in Section 2.2 (i.e., data should be adequate, relevant, and limited to the purpose of the study) is a handy guide in defining the smallest number of interactions possible and determining the impact on the participant.

3.2.5 Participants selection criteria

An important aspect is to define the criteria by which participants will be selected and involved. The selection criteria are usually dictated by scientific reasons (e.g., the random selection of a subset of the population) and the availability of the sample, i.e., the availability of resources such as telephone and email contacts of the participants and the participant's willingness to participate in the study. However, there are also aspects related to the technologies used, which are particularly important in intensive longitudinal surveys. Suppose it is true that 80% of the world's population now owns at least one smartphone [188]. In that case, it is also true that not all models in use are compatible with data collection applications, both for reasons of the operating system and the age of the device. Furthermore, there are substantial differences in the ability to use these devices. From an ethical point of view, it is therefore essential to reflect on the inclusiveness criteria and what scientific and personal consequences exclusion from the study may entail.

3.2.6 Informed consent

Informed consent is the fundamental aspect of an ethical and legal approach to data collection. Informed means that the participant must have all the information necessary to choose to participate in a study voluntarily. Therefore, the participant mustn't feel forced to provide their data. As seen in the design phase (see Sections 2.1 and 3.1), due to the asymmetric relationship between researcher and participant, the researcher will have to pay particular attention to the context and the methods of requesting consent. On the other hand, it is necessary to define all the required information and a transparent communication method so that the participant can effectively understand the activities proposed in the study and the procedures for processing the data. This is not always possible, so the participant must always have the possibility to withdraw from the research and change their consent. Therefore, there are three aspects to consider when seeking informed consent, namely (i) the information to provide, (ii) how to ask for consent, and (iii) when to ask.

Information to provide A privacy statement should include:

1. The legal basis that regulates the consent
2. The contact details of the Data Controller to ask for further details or communicate the withdrawal
3. The purpose of the study, stated in a clear and non-technical way
4. The type of data collected and how they will be processed, managed, stored, and shared
5. Benefit and risk for participating in the study
6. The rights of the participants, e.g., accessing the personal data, asking for rectification or deletion

Appendix B.3 shows a template for the definition of the Privacy Statement. Other examples can be found on OSF [189] or can be provided by different research institutes.

Since transparency and conciseness in communications are fundamental for privacy reasons and the experiment's success, it is also helpful to define a simplified version of the privacy statement for dissemination purposes. A template for such a document can be found in the Appendix B.2. Furthermore, various materials such as flyers and instruction booklets can be prepared to make communication more transparent and intuitive regarding participants' activities during the research.

Given the newness of the technology, by default, iLog requests consent to use each of the sensors by explaining each one. Furthermore, explicit acceptance of the study privacy statement is required before starting the registration procedure, as described in Section 3.3.

How to ask Consent can be requested in oral or written form. In both cases, keeping a copy of the participant's expression of consent is essential. The oral form may result in a video or audio recording of the consent presentation and its approval. In written form, the consent must be signed if presented on paper, or it can be done via an acceptance check for online forms, which must be recorded and maintained in the researcher's database.

When to ask It is generally advisable to request consent before data collection begins. If the data collection has several phases and uses different survey tools, it is advisable to show consent before each phase or immediately after opening the survey tool. There are some cases in which consent is requested retrospectively (e.g. when participants should not be aware of being observed to avoid the Hawthorne effect).

In these cases, it is essential to select a method in advance to trace participants and keep track of information without consent, which cannot be disseminated or reused for research purposes.

Often, in research, there are uses of data that were not foreseen at the beginning of the investigation. Since maintaining personal data usually has a high cost and high risk, it is helpful to provide participants, in addition to informed consent, with a way to stay updated with any changes (e.g., a web page) or further communications (e.g., publications of research results) - which must, in any case, be approved by a privacy office or an ethics committee. In this way, participants will always be able to exercise their rights on the data collected if they disagree with the new methods.

3.2.7 Data management, safety measure and risk evaluation

One aspect to consider is the security measures adopted during all phases of data management. The following Sections 3.3 and 3.4 provide details about the methods and practices to address with a privacy-by-design approach, while below the main aspects of data management are briefly introduced.

Data collection As described in the design, various tools that rely on servers to save the data can be used during data collection. Therefore, ensuring that the tools used have the necessary security measures and guarantee data control and protection is essential. For this reason, with regards to online questionnaires, for example, it is advisable to use tools such as LimeSurvey[118][3] or Qualitrics[119][4]. However, to date, Google Forms has also adopted sufficiently robust measures.

As for servers, most cloud services, such as those offered by AWS, provide excellent security measures. If a server needs to be configured, it is helpful to consider security measures such as SSH keys, Firewall, VPN and Private Networking, Public Key Infrastructure, SSL/TLS encryption, and Intrusion Detection Systems.

Data processing The data management and cleaning process is central to guaranteeing the reliability of the data in compliance with the ethical and privacy principles described in Section 2.2, namely that of Accuracy, Integrity, and Confidentiality. This process is based on three essential aspects, namely cleaning, pseudonymization, and anonymization.

Paraphrasing the definitions described previously, data cleaning must be done both for reasons of research and reuse of data and for reasons of correct assignment of records to the respective participants to avoid the risk of obtaining incorrect results

[3] LimeSurvey follows the German directives dictated by the Federal Data Protection Act (Bundesdatenschutzgesetz, BDSG) and from the point of view of data protection is based on European regulations such as the European Data Protection Directive 95/46/EC and the GDPR

[4] Qualitrics security measures are ISO 27001 certified, which guarantees best practices for information security

(particularly problematic when dealing with information that may have an impact on the well-being of the person).

Another fundamental step is the pseudonymization procedure, i.e., removing the participant ID and replacement with an alpha-numeric code.

Finally, various procedures should take place to remove all participants' personal information (and replace it with anonymous information) and to analyze and share data in a GDPR-compliant way.

Section 3.4 will describe these aspects in more detail.

Storage and distribution The Accuracy, Integrity, and Confidentiality principles (see Section 2.2) regards also data storage. In this case, in addition to the server security measures listed above, it is essential to focus on the organization of the storage and file system. This is done by taking care, especially in the case of pseudonymous data, that the ID decryption keys are not easily reachable by external bodies or are in a place separate from the rest of the database.

Regarding distribution, two other fundamental aspects must be considered: copyright and the data request and download procedures (see 2.2.2.1). Concerning copyright, obtaining permission to redistribute the data through the participants' authorization and to a proper agreement with the data owner (especially in data collections involving multiple data controllers) is essential. Furthermore, there should be a correct procedure for the attribution of copyright. This can happen through drafting technical reports and metadata, as presented in Sections 3.4 and 3.5.

Regarding the download, even if the data has been correctly processed, there is always a high risk of de-identification. It is therefore helpful to define a risk assessment procedure, adapting, for example, to the standards proposed by ISO/IEC 27559:2022 [97] and a distribution procedure for some data. This will have to consider aspects such as authentication and recognition of the people requesting the data and the purpose for which they are requested or evaluating whether they are adequate, relevant, and limited (Article 5(1)(c) of the GDPR). Furthermore, a contract of agreement for the use and management of the data must be drawn up, considering conditions such as (i) the use of the dataset for specific purposes (e.g., for research purposes) for which consent has been obtained from the participants; (ii) prohibition on redistribution or publication of the datasets; (iii) request to delete the dataset once the purpose of use has been achieved.

From a privacy point of view, the distribution procedure can be particularly complex and involve different types of documentation. For this reason, the Appendix B.6 contains all the documents helpful in distributing a dataset.

Risk evaluation According to Article 35 of the GDPR[94]

"Where a type of processing, in particular using new technologies and taking into account the nature, scope, context, and purposes of the processing, is likely to result in a high risk to the rights and freedoms of natural persons, the controller shall, before the processing, assess the impact of the envisaged processing operations on the protection of personal data."

Therefore, given the newness of iLog technology, it could be relevant to conduct a Data Protection Impact Assessment (DPIA), namely a process designed to identify risks arising out of the processing of personal data, aiming at minimizing them as early as possible. The output of the process is the measures applied and a document that can be based on the template in the Appendix B.4.

3.2.8 Privacy in cross-country research

As described in Section 2.2.3, privacy is understood differently depending on different cultures, and its implementation can also vary. From a methodological point of view, it is therefore recommended to follow an inclusive approach in defining the documentation and privacy procedures.

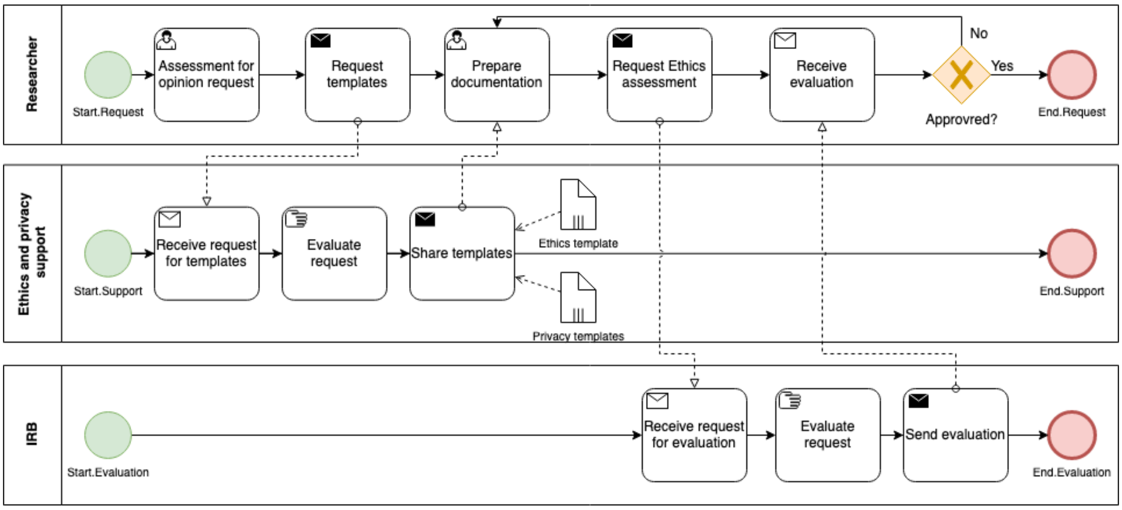
Considering ethics in the case of cross-country data collection, the data controller of each country should indicate a local ethics committee that supports and approves the actions carried out within the project. If, at some universities or research centers, the figure of the ethics committee is absent, they should nominate one. In this case, it can be helpful to proceed with the approvals from the institution that already has an ethical committee to facilitate the process.

Considering privacy, given that iLog is managed in Europe, the current methodology adopts a Eurocentric approach, thanks to the fact that the GDPR is one of the most extensive and comprehensive regulations regarding processing personal data. It is, therefore, advisable to start from the GDPR and then integrate it with additional clauses and procedures as required by the local regulation. In other words, the rules in force in each country, where they do not conflict with the GDPR, should be considered supplementary to the GDPR. In these ways, the enforcement in these countries will be based on the procedures and documents produced in the context of European legislation, with measures that safeguard local law (even though this approach may have some limits, as highlighted in Section 2.2.3).

3.2.9 Ethics and privacy BPMN

The ethical and privacy documentation should result from the choices made in the design phase and implemented in the data collection, management, and distribution phases. Ethical and privacy compliance verification usually occurs through an Institutional Review Board (IRB) or similar. Although each IRB has its templates and guidelines, it is possible to define some common principles and tasks for all

Figure 3.3: Ethics and Privacy phase BPMN



IRBs. For this reason, a BPMN process described in the Figure 3.3 and detailed in the tables 3.7,3.8,3.9,3.10,3.12 has been defined. This process is associated with various templates and support materials helpful in this phase, which can be found in Appendix B.

As regards the timing, as in the case of design, they are very variable and depend on specific aspects such as the purpose of the research and the type of data processed. For instance, if the study requires the processing of sensitive data to involve children or non-EU countries (i.e., with different privacy regulations) and with the transfer of personal data, the preparation and review of the documentation could take several months. The tables refer to timescales for studies that do not involve such complex variants.

Name	Type	Contained in	Description
Researcher	Pool	-	The user of the system that requests the ethical and privacy assessment
Ethics and privacy Support	Pool	-	The main service for supporting the researcher in the request
IRB	Pool	-	The Institutional Revisory Board in charge of evaluating the study

Table 3.7: Ethics and privacy pools and lanes

The Researcher Lane (see Figure 3.3 and Table 3.9) shows the main tasks to be conducted by the researcher, following the approach described in this section. In particular, the researcher will have to recognize the need for an evaluation by

Name	Type	Definition	Description
Start.Request	Start	Start a request	When a researcher needs an evaluation of her data collection design
End.Request	End	Ethics and privacy approved	When the researcher receives the approval of her data collection design
Start.Support	Start	Support	When the researcher request for support
End.Support	End	Share templates	When the templates for ethics and privacy are shared
Start.Evaluation	Start	IRB evaluation	When the researcher request for evaluation
End.Evaluation	End	IRB send evaluation	When the results of the evaluation are shared with the researcher

Table 3.8: Ethics and privacy events

Name	Type	Description	Timing
Assessment for opinion request	User	The researcher fills the questionnaire for assessing the need of an ethics and privacy evaluation	1 day
Request templates	Send	Process for sending a request for templates to the Ethics and privacy Support	1 day
Prepare documentation	User	Process for setting up the ethics and privacy documentation	5 days
Request Ethics assessment	Send	Process for requesting the assessment of the documentation by the IRB	1 day
Receive evaluation	Receive	Process for receiving the results and assessing the needs for further work	3 days

Table 3.9: Researcher - Ethics and privacy tasks

the IRB. Subsequently, he can contact the Ethics and Privacy support service to request the necessary documentation. Support will send essential documentation (as described below) if not specified. Once received, he will have to fill it out and

Name	Type	Description	Timing
Receive request for templates	Receive	Process for receiving a request for support templates	1 day
Evaluate request	Manual	Process for evaluating the request and prepare the templates	2 days
Share templates	Send	Process for sending the necessary support templates	1 day

Table 3.10: Ethics and privacy Support - Ethics and privacy tasks

Name	Type	Description	Timing
Receive request for evaluation	Receive	Process for receiving a request for ethics and privacy assessment and notify the receipt	7 days
Evaluate request	Manual	Process for evaluating the request and prepare the templates	30 days
Send evaluation	Send	Process for sending the final evaluation	7 days

Table 3.11: IRB - IRB tasks

Name	Type	Direction	Description
Exclusive Gateway 1	Exclusive	Diverging	Researcher Gateway for checking if the documentation is Accepted and Rejected by the IRB

Table 3.12: Ethics and privacy Gateways

take care of identifying and submitting the documentation to the IRB. If this is not approved, he can refine it or request further documentation from the support service.

The Ethics and Privacy Support Lane (see Figure 3.3 and Table 3.10) shows the main tasks carried out by the Ethics and Privacy support service. In particular, once the request for documentation has been received, the support service will evaluate it and then send the necessary templates to the researcher. If not specified, the support service will send the essential documentation or the template for the IRB assessment and the one for the privacy statement and informed consent (see Appendix B.1, B.3 and B.2 respectively). In the case of particularly complex research, it may also be necessary to share a template for the DPIA (see Appendix B.4) as well as a Data Processor agreement template (see Appendix B.5) if the researcher requires external support for data collection and management.

The IRB Lane (see Figure 3.3 and Table 3.11) shows the procedures generally followed by an IRB for receiving and evaluating documentation. Being an external body, it will not be detailed here.

3.3 Data collection

Data collection is the operational phase of the research which involves the implementation of the tools and the involvement of the participants. This phase necessarily takes place after the design (Section 3.1) and the approval of the study (Section 3.2) and consists of (i) measurement specification; (ii) iLog (and other software and platform) configuration; (iii) Test; (iv) invitation of the participants and (v) onboarding processes; the (vi) monitoring of study (to ensure data quality); and, finally, the (vii) closing procedures. The last section describes the BPMN of the data collection service. Each of these aspects is presented in the sections below.

3.3.1 Measurement specification

To facilitate the configuration of iLog, once the measurements have been designed (see Section 3.1), they must be communicated to the Technical leader (see Section 3.1.6 for a description of the role) in an orderly and precise manner. The information that must be loaded into iLog is as follows:

- The code of the Ethical Committee approval (not mandatory for the test).
- The GDPR-compliant privacy statement (not mandatory for the test).
- The study timing (Begin and End date and time).
- A short description of the study (purpose, collected data, payment requirements).
- The language of the study.
- The questions and their relative timing.
- The sensors selected.
- The timing of the break options (e.g., for sleeping)

Since the iLog configuration is based on `json`, sharing the files in table format with all fields separated is preferable. This will make uploading easier. The table templates are found in the Appendix A.2.

3.3.2 iLog configuration

In addition to defining the questions, the researcher must indicate a personal or cloud server to deploy iLog. The server specifications will be determined depending on the duration of the study and the amount of data collected.

An essential aspect to consider is the participants' internet connection stability. In a nutshell, iLog notifications can be sent from the server to a service, such as Google Firebase, which takes care of sorting and sending notifications to the participants' smartphones. The responses from the smartphone to the server will be synchronized in the same way. In this way, the system is remarkably efficient. Still, notification reception and synchronization depend, in this case, on the stability of the connection and the server itself, which could be subject to various problems such as overload or shutdown. It is recommended to use this type of configuration in contexts where the connection is reliable, and participants often have access to WiFi. Otherwise, the notifications can be sent offline, i.e., uploaded to the smartphone at the start of the study and managed by the application on-site. However, this involves more significant use of the smartphone's battery.

Finally, it is helpful to remember that although iLog works on both iOS and Android devices, some problems are reported during past data collections. As for iOS, not all sensors are available, while for Android, the app does not work on Huawei devices, as they do not use the Google services on which iLog is based, and problems have been encountered with some versions of Xiaomi smartphones.

3.3.3 Usage of additional data sources

It may be necessary to use information from sources other than iLog for research purposes. In these cases, it is essential to ensure that when using the different tools, the participant is assigned a single identifier to integrate the datasets. The identifier can be an email address or a unique code given at the time of registration. In both cases, it is essential to ensure that the participant always uses the same identifier.

Furthermore, as already described in Section 3.1, using different data collection sources can worsen the user experience by creating misunderstandings and difficulties for participants. Clear and redundant communication (e.g., defining a single web page where all tools are described and linked) will facilitate the process.

3.3.4 Test

Once iLog is configured, the app will be tested. From the participant's perspective, the test aims to validate:

1. issues in the registration and installation process
2. the correctness of the text of the questions
3. the correct receipt of notifications
4. the user experience, namely the usability of the study and the burden of filling out the notification

From a technical point of view, they will have to be considered

1. Server capacity and overload
2. Consistency in sending notifications

Once the test has been completed and any problems resolved, it will be possible to generate the "study code," i.e., the code the participants will use to access the configured study and begin data collection.

3.3.5 Participant invitation

Based on what is defined in the recruitment design phase, the invitation to participants can take place through direct or indirect contact with participants. Direct contact can occur through the organization of data collection presentation events or by sending a message (email, text message, WhatsApp, or Telegram) to the participant's address. Indirect contact can occur through flyers, posts on the leading social and university channels, and word of mouth (especially in snowball sampling strategies). As highlighted in the previous sections, regardless of the type of contact, the participant must receive the following information:

1. Link to download iLog
2. Study code
3. Duration of the study
4. Required tasks
5. Incentives and selection criteria
6. Privacy information and data management
7. Instructions for managing iLog and tasks

Despite the simplicity of the activities required of the participant, it can be seen that this is a considerable amount of information. For this reason, it is advisable to consider a certain redundancy in communication, sending the same materials through different channels and diluting over time. For example, in the case of direct contact via a presentation event, it is helpful to create slides in which all the material is presented. Once the participants have registered with iLog (and have viewed and accepted the privacy statement), they can be sent a reminder email containing the instructions for participating in the surveys (points 3., 4., 5., 7. of the list above) and, at a later time, an email containing the privacy material (point 6. above). An example of the instructions can be found in Appendix C.2, while the privacy material can be found in Appendix B. Privacy documents should always be shared to remain accessible to the participant.

3.3.6 Onboarding and data collection start

Once the invitation has been received, participants can register for the study and begin data collection. During this phase, some problems of both a logistical and technical nature may arise.

From a logistical point of view, participants may have difficulty installing the application or have further curiosities and requests for clarification before starting the study. Furthermore, some participants may learn about the survey a few days after the invitation if they do not regularly check their email or the communication channels they use, or if they know about it through word of mouth.

From a technical point of view, some server overload problems may occur due to the high number of accesses. In this case, the iLog app will not finalize the recording. While frustrating, this problem is usually resolved by restarting the application or uninstalling and re-installing it. This may lead to the arrival of various participant requests and an extension of the registration procedure.

For this reason, it is helpful to plan an onboarding phase of the study, which can last several days depending on the design chosen. If the invitation took place directly and no snowball sampling was planned, it was found that three days of onboarding were sufficient to reach the desired number of participants. Another fundamental aspect is the presence of a help desk that must be particularly active during the first days, responding promptly to participants' requests to limit the number of dropouts in the initial phase.

3.3.7 Monitoring

It is well known that in longitudinal surveys, especially in intensive ones, dropout rates are exceptionally high, averaging around 30% of the participants who signed up for the survey. To limit them, in addition to a correct study design that considers the respondent burden and a set of incentives, it is necessary to conduct active monitoring for the entire data collection campaign. This is divided into two aspects: the help desk function, the daily verification of the back end, and the data quality obtained.

As mentioned in the previous section, the help desk must be active and promptly respond to participants. Without sounding alarmist, every hour that passes without the participant being able to solve their problem with the app, there is less data detected, risking affecting the quality of the final result. The Appendix C.3 contains a series of Frequently Asked Questions (FAQs).

From the point of view of back-end verification and data quality, it is a good practice to have an alarm or daily monitoring system in case there are crash problems with the server to resolve them promptly. It is also helpful to provide a daily report of the data collected, i.e., the number of questions answered per participant and the amount of sensors collected. This way, it will be possible to check the study's progress and proactively contact participants experiencing problems with the app, even without realizing it. For example, they may have inadvertently turned off GPS or not received some notifications.

3.3.8 Closing procedures

Once the study has been completed, the following activities must be conducted:

1. Notify participants
2. Wait two weeks so that all participants are able to synchronize their data
3. Dump the data from the iLog backend
4. Save data in a safe space (if data preparation is needed) and/or share the data with the data owner

If the study provides for the remuneration of participants, two additional activities will be performed:

1. Assess qualified participants for payments
2. Issue the payments (by contacting the participants and the administrative staff)

Regarding the assessment of the contributions, not only the number of total responses must be considered, but their percentage of completeness must be compared to the notifications sent by the server and those actually received on the participants' smartphones. This is to avoid excluding some deserving participants, as it is possible that some notifications have been lost due to errors in the backend.

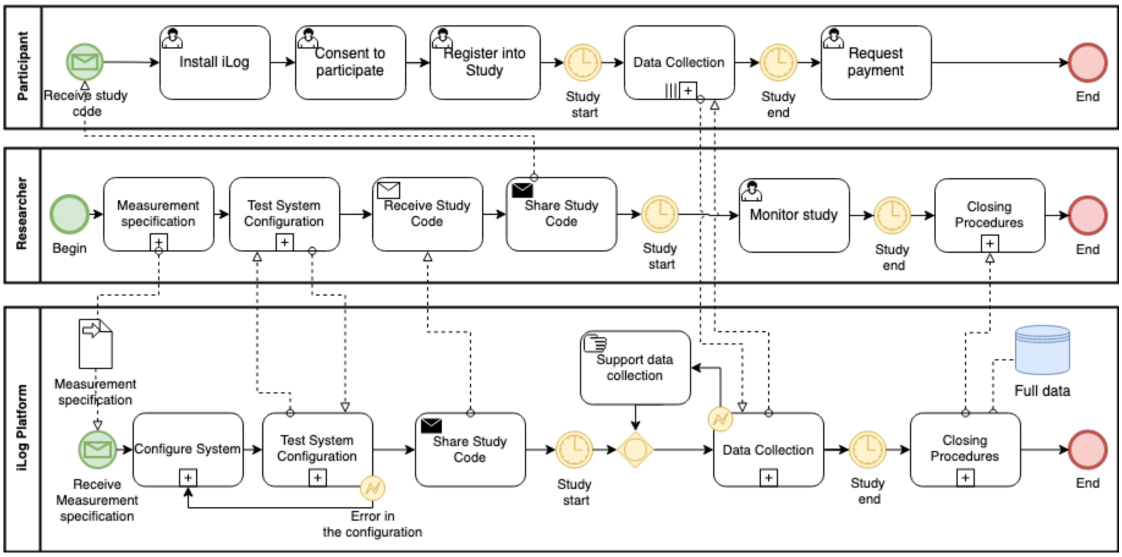
It is also advisable to consider the number of daily contributions for each participant. In this way, it will be possible to choose to pay even participants who, despite not having reached the required percentage of contributions, still participated assiduously in the study (i.e., responded to the notifications for most of the days).

3.3.9 Data collection BPMN

The sub-processes of the overall service of iLog Platform (Data Collection) show the steps and documentation used for data collection that should be approved before this sub-process. Figure 3.4 shows the configuration, testing, and data collection processes for the actual collection of data for the Research Infrastructure. BPMN process described in the Figure 3.4 and detailed in the tables 3.13,3.14,3.16,3.17,3.18 has been defined. This process is associated with various templates and support materials helpful in this phase, which can be found in Appendix C.

The Participant Lane (see Figure 3.4 and Table 3.15) shows the activities that the participant must carry out. Once the survey invitation and information materials have been received, the participant can choose to install the iLog app (and other applications or links depending on the survey procedure). Once the app has been installed, the participant will have to give consent to participate and then proceed with the registration procedure, which consists of providing your email, consenting to the use of the different sensors, and (if present) completing the participation questionnaire. Afterwards, the participant will be able to participate in the study,

Figure 3.4: Data collection phase BPMN



Name	Type	Contained in	Description
Participant	Pool	-	The person invited to participate in a data collection
Researcher	Pool	-	The user of the system that requests the data collection execution
iLog Platform	Pool	-	The main service for supporting the researcher in the data collection process

Table 3.13: Data collection pools and lanes

answering questions and stimuli and managing the collection of sensors. Once data collection has been completed, the participant will be informed of the procedures to be conducted and, if required by the study, will receive the forms to make the payment request.

The Researcher Lane (see Figure 3.4 and Table 3.16) shows the activities conducted by the researcher. Following the design of the experiment and its approval by the ethics committee (as described in Sections 3.1 and 3.2), the researcher will be able to move on to the operational phase of data collection. In this phase, the first task to be conducted is the measurement specification, i.e. the transcription of the stimuli into a format that facilitates uploading to iLog and, if applicable, to other measurement instruments (see Section 3.3.1 and Appendix A. Furthermore, they must indicate the characteristics of the study, as indicated in section A.2. Once this information has been shared with the Technical leader (see Section 3.1.6 for a description of the roles) the researcher will receive a code to test the iLog app, considering the aspects indicated in Section 3.3.4 and send a report to the Technical leader highlighting the problems encountered. Once the test is concluded, the researcher will receive the study code to share with the participants (see also

Name	Type	Definition	Description
Receive study code	Start	Participate in study	When a participant receives the invitation for participating in a study
End.Participant	End	End participation	When the data collection ends for the participant
Begin.Researcher	Start	Start data collection	When the researcher sends the questions and sensor to be configured in iLog
End.Researcher	End	Closing data collection	When the researcher finalizes the procedures for closing the study namely receives the collected data and issues the payment request for the participants
Receive Measurement Specification	Start	Start configuration	When the iLog responsible receives the resources to be uploaded on iLog
End.Collection	End	End data collection	When the results are shared with the researcher

Table 3.14: Data collection events

Section 3.3.5 and 3.3.6), together with all the material necessary information such as the description of the study (Section 3.3.1), privacy documentation (Appendix B.2 and B.3), instructions (Appendix C.2) and other material. Once the study has started, the researcher can proceed to the monitoring phase, described in Section 3.3.7 and supported by the material in Appendix C.3. In this phase, the Study Leader will interact with technical support to receive information on the progress of the study and for further help. Once the study is concluded, the researcher will be able to carry out the closing procedures, as described in Section 3.3.8.

The iLog Platform Lane (see Figure 3.4 and Table 3.17) shows the activities conducted by the technical leader who manages iLog. Once the researcher has received the study specifications, the technical leader starts the iLog configuration, loading the questions and sensors with their respective frequencies. When the configuration is completed, the technical leader sends the codes to the researcher to

Name	Type	Description	Timing
Install iLog	User	Process for downloading and installing the iLog app	3 days
Consent to participate	User	Process for reading the informed consent and accept to participate in the study	1 day
Register into study	User	Process for registering in the study	1 day
Do data collection	User	Process for participating in the data collection and fulfilling the assigned tasks (e.g., answering questions, providing sensor data, etc.)	.
Request payment	User	Process for requesting the payments	1 day

Table 3.15: Participant - Participant tasks

Name	Type	Description	Timing
Measurements specification	User	The researcher prepare the documentation to be uploaded on iLog	3 days
Test system configuration	User	Process for testing the configuration of tasks (e.g., questions) on iLog	3 days
Receive study code	Receive	Process receiving the study code from iLog	1 day
Share study code	Send	Process for sharing the study code and instructions to the participants	1 day
Monitor study	User	Process for monitoring the collected data and helping the participants during data collection	.
Closing procedure	User	Process for receiving the results and issuing the payments	15 days

Table 3.16: Researcher - Data collection tasks

start the test system configuration, as described in Section 3.3.4. The test will be considered completed when all requests for correction by the researcher have been responded to and/or implemented. At this point, the study code can be shared and the data collection can be started with the respective monitoring (see Section 3.3.7) and data collection support in case of problems or errors. Afterward, the technical leader can conduct the closing procedures and send the dataset to the researcher or start the data preparation process (see Section 3.4).

Name	Type	Description	Timing
Configure system	Manual	Process for configuring iLog tasks and sensors	15 days
Test configuration	Manual	Process for testing the back-end and front-end of the study configuration	7 days
Share study code	Send	Process for sending the study code	1 day
Data collection	Service	Process for collecting the data	.
Support data collection	Manual	Processes for supporting the researcher and fixing errors	.
Closing procedures	Manual	Process for dumping the data, storing them in a safe repository, and sharing a copy with the researcher	15 days

Table 3.17: iLog Platform - Collection tasks

Name	Type	Direction	Description
Study start	Timer Boundary	Interrupting	When iLog start to collect data
Study End	Timer Boundary	Interrupting	When iLog end to collect data
Error in configuration	Error	Interrupting	When errors in iLog configuration are detected during the study test
Problems in data collection	Error	Non-Interrupting	When problems arise during data collection, such as missing notification or server crash
Inclusive Gateway 1	Inclusive	Converging	iLog Platform-Gateway dealing with the technical problems that may arise during the data collection

Table 3.18: Data collection Gateways

3.4 Data preparation

According to Chapter 2 and Section 2.4, data preparation is a fundamental aspect to ensure the replicability of results and correct application of legal principles. To this end, this section will present a description of the types of (i) input information in the data preparation process, the actual process of (ii) data extraction, (iii) transformation, and (iv) cleansing of the data; data consolidation from the point of view of (v) privacy operations; and finally the creation of the (vi) documentation necessary to facilitate the replicability of the results. The section will conclude with some considerations on (vii) storage procedures and (viii) data quality, as well as on (ix) the process's BPMN.

3.4.1 Input data

In addition to the raw data from data collection applications (e.g., iLog), data preparation requires information useful for contextualizing the study and the variables collected. The information ranges from instructions for study design and configuration of the data collection application to privacy aspects, as detailed below.

Study design The document (which can be redacted according to the template reported in Appendix A.1) will be useful for drafting the final documentation, also considering some copyright information, like:

- *Author name*: The authors of the study, the individuals who conducted the study and are responsible for the results;
- *Copyright holders*: The copyright holders, the individuals or organization that holds the copyright for the study;
- *Copyright year*: The copyright year, the year in which the study was conducted;
- *Creator*: The Creator is the person who created the dataset documentation.

By including this information in the Catalog webpage (see Section 3.5, the individuals who access the dataset can verify the data's authenticity and ownership. Moreover, this information is essential for properly citing and crediting the dataset.

iLog question specification This documentation (which can be redacted according to the template reported in Appendix A.2) is used to describe the research protocol and facilitate data preparation in case of doubts or inconsistencies in the dataset. In particular, the following information is necessary:

- *Study name*: This parameter specifies the name of the study, which is going to be used to name the study directory;

- *Study ID*: This parameter uniquely identifies the study, enabling the extraction of data specific to that study;
- *Starting date of the study*: The parameter specifies the start date of the study, ensuring that only relevant data is extracted;
- *Ending date of the study*: The parameter specifies the end date of the study, ensuring that only relevant data is extracted;
- *Timezone of the study*: The timezone is required to successfully convert the dates from the server dates to the study timezone and have consistent detections.

Privacy material Like Ethical committee approval, informative, privacy statement (see Appendixes B.1, B.2, and B.3) and other relevant privacy material (e.g., the DPIA B.4, but also exchange of emails with the IRB or the legal department). This documentation must be stored and will be used in the audit case.

Additional materials Like flyers and instructions for using iLog shared with participants, but also additional notes of problems detected during data collection, which can be attached to the documentation to be shared with external parties or used as a reference for future experiments.

The input datasets Specifying the dataset names used to identify the relevant data to be extracted from the server.

3.4.2 Extraction

The extraction consists of retrieving the raw data and the user IDs from the iLog server. It is divided into three main parts: (i) Study setup, (ii) User extraction, and (iii) Sensor extraction. These parts provide the initial input for subsequent data cleaning and transformation steps, as described below.

Study setup Stores the study information in a .ini configuration file for the needed scripts. To ensure the reproducibility and repeatability of the study, it is essential to have a record of its setup. This information is stored in a .ini configuration file, which serves as input for the needed scripts. The configuration file is created using a script that parses the information given as arguments and creates a folder for the study directory. The script generates a .ini file inside the folder containing the study configuration. By storing the study information in this way, the setup can be easily accessed and modified, ensuring consistency and accuracy throughout the data preparation process.

Users exctraction Retrieves the user IDs for the participants. It is a process that involves downloading the user IDs from the iLog server. The user IDs are saved in a dictionary in `json` format. Then, each user is assigned a numerical index in ascending order. Finally, the updated user dictionary is saved as a `json` file, which will be used in subsequent steps of the data preparation pipeline.

Sensors exctraction The goal is to download each dataset specified as a parameter from the start timestamp to the end timestamp defined as study properties. The parameters for each dataset depend on the dataset type and have been previously decided based on the frequency of the dataset readings and usage.

Once these input parameters are defined, a Python script automatically downloads the data formatted as a `json` that adheres to a particular schema.

The schema contains two top-level fields: `studyId` and `properties`. The `studyId` field includes a unique identifier for the study to which the dataset belongs. The `properties` array field contains a single element, with critical `datasetName` as the value of the list of rows of the dataset. Each row is formatted as a `json` object composed of several key-value properties.

3.4.3 Transformation

After the download of the datasets, the next step is to transform them into a more efficient format for storage and analysis. The Data Transformation to `parquet` section explains the different steps involved in changing the `json` files to `parquet` format. The process is different for sensors and contributions: for the sensors, the object columns are split into multiple columns, and then the entire `json` file is converted to a single `parquet` file. For contributions, missing columns and content are added, and unused columns are removed before the conversion. In both cases, the new `parquet` files are concatenated, and the original `json` files are deleted to reduce disk space usage. Finally, a password-protected `zip` backup is created.

Split object column String, integer, and float values can be directly converted into a columnar format. However, object values require splitting into multiple columns. The script uses a hard-coded list of possible objects in the datasets to map the values to the corresponding columns. First, it selects the object columns by comparing the column names (which are always unique). Then, it creates new columns based on the possible values and maps them to the corresponding ones.

Add missing columns and map values Converting the contributions to `parquet` requires a more challenging approach, as it involves a manual procedure.

For the contribution questions, the manual procedure consists of identifying a rule to assign a corresponding tag for each first question that identifies the type of question and implementing it in Python code. This is necessary because the tag is not stored in the `json` file. Instead, all the information associated with that type of question is stored in the column `question`.

For the contribution answers, the manual procedure consists of (i) specifying the answers columns with the corresponding labels and type of question and (ii) identifying a pattern to assign the corresponding answers and the correct format.

Once the necessary modifications are made, the script has to be further modified to map the answer values to responses, identifying a pattern for the assignment. In each row, multiple answers correspond to a specific type of question for a user.

This information is in a complex array called `answer`, which needs to be analyzed to identify the associated responses. Thanks to the input documentation (see Section 3.4.1, it is possible to identify a unique pattern.

Remove useless and redundant columns Finally, some columns are redundant or useless for the analysis; thus, they are removed from the temporary dictionary. For instance, the `question` and `title` columns are deleted, as they contain redundant information not required for the analysis.

Convert `json` dictionaries to `parquet` At this process stage, the modified datasets are in a Python dictionary format and need to be transformed to `parquet` for simplified management and reduced space. This conversion is directly handled by the `dask.dataframe` library merges the different dictionaries (previously split by user and study chunks) corresponding to the same dataset into a single one.

Create password-protected `zip` backup After converting from dictionary to `parquet` format, a `zip` backup of the non-anonymized datasets is created to be stored in the SH for security reasons. This backup is password-protected to ensure the privacy of the participant's data. The backup process is automated, and the corresponding password is stored in a separate `txt` file saved in Safe Haven.

The non-anonymized datasets are saved following their original paths, which preserves their internal organization and makes it easier to locate a specific dataset if necessary. The backup process ensures that the non-anonymized datasets are securely stored in case of data loss or corruption.

3.4.4 Cleaning

The contributions are cleaned by removing duplicates based on the tuple of columns `studyid`, `userid`, and `instanceid`. The dates are converted to DateTime, and the timezone is changed to where the study occurred.

Concatenate by type contributions The script used in this step works by selecting the relevant datasets for each study (such as time diaries, tasks, etc.) and then concatenating the contributions by type (answer, question, or confirmation) into a single file. This helps simplify the data management process and allows for easy analysis of the contributions. Since the preparation procedure differs, the concatenated files are stored in their respective directories (Answers, Questions, and Confirmations).

Clean contributions The contributions are cleaned by removing the duplicate rows. To identify duplicates, the script checks the values in the following columns: `studyid`, `userid`, and `instanceid`. If a row has the same values in all three columns as another row, it is considered a duplicate. All the duplicate rows are deleted, and only the first row is kept. This ensures that each contribution is unique and reduces redundancy in the dataset.

Convert dates to DateTime and change timezone During the conversion of the datasets, an important step is to convert the dates from the integer format (YYYYMMDDhhmmssfff) to a DateTime format (YYYY-MM-DD hh:mm:ss.fff) and change the timezone from the one used by the Data Collection server (ETC/GMT+1) to the timezone of the study location. This step ensures that the DateTime information is consistent and usable for further analysis. The conversion can be done using the `pytz` and DateTime libraries in Python.

3.4.5 Privacy operations

The anonymization procedure dealt with the possible sources of possible identification, namely (i) Personal Data Anonymization, (ii) Network Anonymization, and (iii) GPS Anonymization.

Personal Data Anonymization All personal information, i.e., `email address`, `home address`, `name and surname`, must be removed from each of the three types of datasets (online questionnaire, time diaries, and sensors), still making sure that the same unique identifier would be assigned to the same person across all three datasets.

Network Anonymization Considering network connection there are three possible sources of re-identification, namely: (i) `WiFi-event`, which shows the WiFi network the smartphone is connected to; (ii) `cellular-network`, which shows the roaming network; and (iii) `WiFi-networks-event`, which shows the WiFi networks that are available in the environment. The relevant columns have been anonymized for each sensor file using unique identifiers. A hash function was applied to the WiFi network name, with a function that cannot be reversed (the SHA-256 cryptographic function is used to perform the hash).

GPS Anonymization The location visited by the participants is one of the most sensitive data. If joined with timing information, the location may quickly lead to re-identification, e.g., the place where a person spends most of their night hours is their home. Therefore, our solution is to make the spatio-temporal information more ambiguous. There are many ways of doing this, all having different consequences for the usability of the dataset for research. For example, eliminating GPS signals at night would ensure good personal data protection. Still, it would not allow a person's movements to be correctly tracked, especially if they are not home. Considering the main GPS uses, namely the tracking of movements and the identification of places visited, two anonymization methods are proposed as follows:

1. *Round Down* The GPS precision is truncated so that it becomes anonymous but in a way that is still useful for certain scientific purposes.
2. *Point of Interest (POI)* Instead of providing the entire stream of GPS position, the POI approach returns only those locations where the user has spent more than a certain amount of time. This dataset adds a POI tag to the stream if latitude and longitude do not change for one minute. For each POI, the elapsed time in seconds is also added. GPS longitude and latitude readings are removed. The POI is selected to identify a general location (suburb, city, region) and the closest relevant places (bar, restaurant, lake, etc).

The output of these procedures is two datasets called *RoundDown* and *POI*, each containing all the other sensors. For privacy reasons, only one of the two datasets can be downloaded by the same research institution since the union of the two would quickly lead to re-identification.

3.4.6 Data documentation

Once the data preparation is complete, the practitioner can proceed with documenting the dataset and data collection. Preparing the documentation will be helpful for internal management, audit purposes, and external purposes of publication and communication with other researchers.

As described in the 3.4.1 section, the documentation for internal and audit purposes concerns privacy documents and all the material produced during data collection. To facilitate access, these must be saved in the Core Repository (see Section 3.4.7) following the file system nomenclature.

However, three types of documentation must be created for external purposes: the metadata, the codebook, and the technical report.

As regards the metadata, an extended description can be found in the following Section 3.5 and aims to index the dataset, encouraging a concise description and facilitating access through search functions. The codebook must instead contain an accurate description of the variables used, considering the Acronym of the variable present in the dataset, the description of the variable, the Variable class (numeric, string, DateTime, boolean, ...), Unique values or number of users, Missing values. A template for the codebook is available in Appendix D.1, but it can be generated automatically using the Python and R packages described in Section 2.4.

Finally, the technical report is a document that should describe all the aspects that led to the dataset collection and its results. In particular, the technical report should highlight (i) the purpose of the data collection, which defines the boundaries in which data are applicable (it answers the question "What is the problem addressed?"), (ii) the state-of-the-art methodology and datasets, highlighting gaps and pointing at similar data available (it answers the question "Why is it important to generate a new dataset?"); (iii) a description of the data collection methodology adopted ("How the data were collected?"); and, (iv) the results, namely an outline of the type of data collected (e.g., demographics, list of data collected, examples of bivariate statistics) and their potential reuse.

A template for the technical report, drawn up based on [170], is available in the Appendix D.2.

3.4.7 Storage

The outputs of the data preparation process are two datasets that vary in their respective level of anonymization. The fully anonymized dataset, namely a dataset that does not contain any personal information, that is, neither direct identifiers (e.g., name, home address) nor indirect identifiers (e.g., the stream of GPS positions), can be saved in a protected repository which respects access control policies, but which does not necessarily have to be disconnected from the network. The person in charge of the repository (and the owner of the data) will have to guarantee correct access to the information, not least by ensuring that the repository contains not only the datasets but also all the documentation necessary for their correct reuse (e.g., description of variables, privacy documents, etc.).

The data preparation procedure also includes saving data that is not entirely anonymized. This can be the raw data or the output of intermediate preparations. Despite the high risk of managing this data, their retrieval is necessary when errors are found in the prepared and anonymized dataset or if new anonymization procedures are developed. To ensure the security of these datasets, they should be stored in a separate safe haven or encrypted repository. The dataset access should be limited to data owners only, and the maintenance should be limited in time, which, by privacy procedures, could last a few years after the conclusion of data collection. Also, in this case, the correct dataset maintenance requires saving the relevant documentation to facilitate future reuse.

3.4.8 Quality

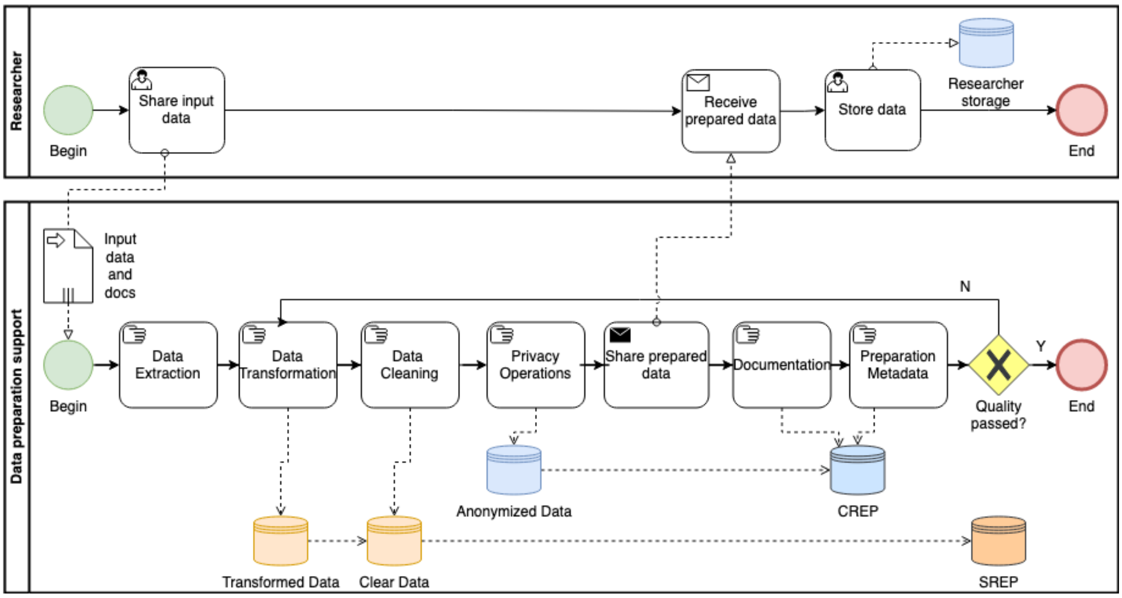
The entire data preparation process has been defined to ensure a minimum quality of the datasets, with a view to their consolidation and future reuse. Therefore, data quality should be derived from data consolidation procedures, such as the enrichment of variables with labels, data coding, and so on. Further enrichment of data quality can take place from missing data imputation procedures, such as using GPS location data to validate or infer the `Home` location label provided by the study participant. In any case, a fundamental aspect of consolidating data quality is testing or validation. The preparation procedure generally requires documentation such as the technical report and the codebook to be created. Reviewing these documents facilitates not only reuse but also control for inconsistencies that may have occurred during the data collection or preparation phase. Through the codebook, for example, it is possible to view missing data patterns and evaluate their consistency. However, it also checks that the data types have been correctly imputed and that all users are mapped to all datasets.

Finally, it is also possible to discover inconsistencies during the data distribution phase (as in Section 3.5), in which other practitioners may send feedback on the data quality or even additional scripts to generate new features.

3.4.9 Data preparation BPMN

The sub-processes of the overall service of the data preparation pipeline show the steps and data storage used for preparing the collected data and moving it to the final repository. The BPMN process is defined in Figure 3.5 and detailed in the tables 3.19,3.20,3.21,3.22,3.23 has been defined. The BPMN considers the case in which a researcher has participated in the previous phases of data collection and wishes to use the preparation service. The data preparation templates and support materials are reported in Appendix D.

Figure 3.5: Data preparation phase BPMN



Name	Type	Contained in	Description
Researcher	Pool	-	The user of the system that requests the data preparation execution
Data preparation	Pool	-	The main service for supporting the researcher in the data preparation process

Table 3.19: Data preparation pools and lanes

Name	Type	Definition	Description
Begin.Researcher	Start	Start request for data preparation	When the researcher requests for data to be prepared
End.Researcher	End	Storing prepared data	When the researcher stores the prepared data
Begin.Preparation	Start	Start data preparation	When the support receives the request, the information and the data to be prepared
End.Preparation	End	End data preparation	When the results of the prepared data are documented and stored in the repository

Table 3.20: Data preparation events

Name	Type	Description	Timing
Share input data	User	The researcher prepare the data and documentation for data preparation	5 days
Receive prepared data	User	Process for receiving the prepared data and checking them	5 days
Store data	User	Process for storing prepared data	1 day

Table 3.21: Researcher - Data preparation tasks

The Researcher Lane (see Figure 3.5 and Table 3.21) shows the activities carried out by the researcher. In particular, the researcher provides all the information necessary for data preparation (see 3.4.1) and then takes care of receiving the prepared data and storing it in a secure repository or safe haven, in the case of non-anonymous data (see 3.4.7).

The Data Preparation Support Lane (see Figure 3.5 and Table 3.22) shows the activities conducted by the technical support, who will have to deal with all the procedures described in this section from data extraction, to the transformation and

Name	Type	Description	Timing
Data Extraction	Manual	Process for extracting the data and converting them to parquet	3 days
Data Transformation	Manual	Process for converting `json` strings into tabular data	5 days
Data Cleaning	Manual	Process for labelling, merging and removing duplicated data	5 days
Privacy Operations	Manual	Process for pseudonymizing and anonymizing the data	5 days
Share prepared data	Send	Process for sending the prepared data to the researcher	1 day
Documentation	Manual	Process for documenting the data through a technical report and a codebook	10 days
Preparation metadata	Manual	Process for creating the metadata of the dataset preparation	5 days

Table 3.22: Data preparation support - Data preparation tasks

Name	Type	Direction	Description
Exclusive Gateway 1	Exclusive	Diverging	Data preparation support Gateway for assessing the data quality

Table 3.23: Data preparation Gateways

cleansing of the data; up to anonymization and the creation of the documentation which will be saved in the appropriate repository.

3.5 Data distribution

The distribution phase follows the data preparation phase described in Section 3.4. This ensures that the data is of the correct quality and sustainable from a privacy point of view. As from Section 2.5, although there are several catalogs and repositories available, the current methodology focuses on the LivePeople Catalog [172], developed by the KnowDive group in the context of the Horizon 2020 WeNet Project - The internet of us [29]. The LivePeople Catalog is designed to distribute Big Thick Data and contains features, such as specific metadata and privacy documentation, not present in other catalogs. The LivePeople Catalog is a systematic and organized information collection presented in a structured format. It serves as a reference or inventory that details a specific set of resources and allows the publication and download of the resources (i.e., experiment data collections). The new release has been implemented in JKAN. JKAN is based on GitHub and Jekyll; it allows the publication of a catalog based on a GitHub repository, making it more accessible to third parties.

The Catalog is publicly accessible, and the Appendix E presents a user journey to navigate the Catalog and its primary functions.

The main activities that concern the distribution process are (i) uploading and documenting, (ii) searching, and (iii) downloading data from the Catalog. The last section will present the (iv) BPMN.

3.5.1 Upload

The upload phase involves identifying and describing all the information necessary for the publication of the dataset, enabling its search through tags and keywords and facilitating its reuse. In particular, the essential information can be articulated into a set of (i) metadata, which should consider different (ii) data types and the relevant (iii) documentation of the dataset.

3.5.1.1 Distribution metadata

Metadata is a synthetic, findable and machine-readable description of the main features of the collected dataset. In general, much of the information regarding Big Thick Data is similar to that of other quantitative data and identifiable in the set of metadata proposed by [150], even if particular attention must be paid to the types of data (as described in the next section). Considering the search aspects of the dataset, as well as the different outputs of the data preparation, it is possible to distinguish the metadata categories into three parts. The first is *Title and Notes*, which represents the name of the dataset and a brief description with keywords that facilitate both research and understanding for non-experts. The second is the *Resources*, i.e., the entire set of downloadable materials regarding the dataset and data collection, to allow a more detailed analysis of the dataset's content and subsequent preparation and reuse. Finally, there are the actual *Metadata*, which contain the summary of the main characteristics of the dataset and allow it to be framed in terms of time and space (when and where the data collection took place and for how long) as well as for the features that highlight its composition and quality (e.g., the number of people who participated in the data collection) as well as aspects regarding copyright and privacy (e.g., the authors of the dataset, but also the rules for ownership and distribution). The following tables describe the metadata available in the LivePeople Catalog.

Field	Value type	Description
title	textual	The title of the resource
notes	textual	A description of the dataset, containing: (i) the project name and objectives; (ii) the type of data collected; (iii) suggested reuse.

Table 3.24: Title and Notes

3.5.1.2 Data type

The data saved in the Repository are of two types: interaction data, namely data collected via the user's interaction, and sensor data, passively collected from the user device(s).

Based on Android Developers [190], the categories of sensors are:

1. *Connectivity*: Package with Objects representing connections with other devices. This category includes the following: Bluetooth Low-Energy, Bluetooth Normal, Cellular Network, Wifi, and Wifi Networks.

2. *Environment*: These sensors measure various Environment parameters. The Android platform provides four sensors monitoring various environmental properties, such as relative ambient humidity, illuminance, ambient pressure, and ambient temperature near an Android-powered device. All four environment sensors are hardware-based and available only if a manufacturer has built them into a device. Environment sensors are not always available on devices except for the light sensor, which most manufacturers use to control screen brightness. This category includes the following: Ambient Temperature, Audio, Light, Pressure, and Relative Humidity.

3. *External Device*: All other sensors collected from external devices related to the leading smartphone (smartwatches and others). This category includes the following: Bluetooth accelerometer, Bluetooth gyroscope, Bluetooth heart rate, and Bluetooth magnetic field.

4. *Motion sensors*: These sensors measure acceleration and rotational forces along three axes. This category includes the following: Accelerometer, Accelerometer Uncalibrated, Activities, Gravity, Gyroscope, Gyroscope Uncalibrated, Linear Acceleration, Rotation vector, Step Counter, and Step Detector.

Field	**Value type**	**Description**
download-request-name	textual	The name of the document to request the dataset
download-request-URL	URL	The link to the downloadable document to request the dataset
download-request-format	textual	The format of the resource
technical_report-name	textual	A report describing the data collection (design, collection, preparation, main results)
technical_report-URL	URL	The link to the downloadable technical report
technical_report-format	textual	The format of the resource
codebook-name	textual	A codebook describes the contents of a data collection. A well-documented code- book contains information intended to be complete and self-explanatory for each variable in a dataset.
codebook-URL	URL	The link to the downloadable codebook
codebook-format	textual	The format of the resource
additional_material-name	textual	(optional) the materials used during the research (e.g., focus group tracks, interviews, and questionnaires)
additional_material-URL	URL	The link to the downloadable additional material
additional_material-format	categorical	The format of the resource

Table 3.25: Resources

5. *Position sensors*: These sensors measure a device's physical position. This category includes orientation sensors and magnetometers. This category consists of the following: Game Rotation Vector, Geomagnetic Rotation Vector, Location, Magnetic Field, Magnetic Field Uncalibrated, Orientation, and Proximity.

6. *App usage sensors*: These sensors identify how people use and interact with social media and applications. This category includes Headset Plug, Music, Notification, and Application.

7. *Device usage sensors*: These sensors identify how people use and interact with their devices. This category includes the following: Airplane Mode, Battery Charge, Battery Level, Doze, Ring Mode, Screen, Touch, and User Presence.

Field	Value type	Description
License	URL	The link to the license web page
dataset name	textual	The name of the dataset
location	textual	The name of the city (or province) in which the dataset has been collected
start date	textual	The beginning of the data collection (i.e., the oldest date in the dataset)
end date	textual	The end of the data collection (i.e., the most recent date in the dataset)
dataset type	categorical	The type of dataset
sensor type	categorical	The type of sensor (only if dataset type == Sensor)
size	float	The size of the dataset in MegaByte (MB)
dataset format	textual	The format of the dataset
other available format	textual	Other format of the dataset available
number of participants	integer	The number of participants (unique id) in the dataset
language	textual	The language of the dataset (only if dataset type==Synchronic interaction | dataset type==Diachronic interaction)
collection name	textual	The name of the collection that collected the dataset
project URL	URL	A link to the webpage of the project
organization	textual	The name of the institution that collected the dataset
domain	categorical	The domain of the dataset
5-stars	integer	Ranking based on Tim Berners-Lee's 5-star deployment scheme for Open Data
publication date	textual	The date in which the dataset has been released
identifier	textual	alphanumeric code identifying the resource
download request	textual	email of the institution responsible for validating the download request and sharing the data

Table 3.26: Metadata

3.5.1.3 Documentation

The additional documentation of the datasets is the output of the data preparation process (see Section 3.4) and aims to make accessible both aspects regarding the method with which they were generated and relevant information on the datasets' content. In this sense, documentation ultimately enables the reproducibility of studies and facilitates the reuse of the collected data. An additional purpose of the documentation is to facilitate legal audit, but often, the necessary materials regarding ethical compliance and privacy are not distributed to the general audience. The list below reports what, based on the state of the art (see Section 2.5), are the documents usually shared, or:

1. the technical report (see Appendix D.2),
2. (optional) the codebook (see Appendix D.1),
3. (optional) the materials used during the research (e.g., focus group tracks, interviews, and questionnaires).
4. (optional) relevant publications describing the dataset.

Note how the technical report is, in our opinion, the only essential document, as it contains all the fundamental information to reproduce the investigation and contextualize the data collected, thus guiding the practitioner to their contextualized reuse (i.e., not "out of context" as stated by [5] and reported in Chapter 1). This can be accompanied by all the materials used to conduct the study to make its replication even easier. These materials, together with the codebook, are also particularly useful in the data reuse phase, where the practitioner can easily reconstruct the labels associated with the dataset variables and their respective values, as well as view some exploratory analyses (in the cases of more structured codebooks). Finally, sharing scientific articles that have used the dataset will facilitate its understanding and reuse by a wider audience. This latest documentation cannot be the product of the preparation phase and will have to be updated over time.

3.5.2 Search

According to the FAIR principle [161], in addition to being documented and available on the web, datasets must be able to be found. This happens through various queries the end user will carry out based on their search purposes. The queries will filter superfluous information, returning only the subset of datasets of interest. Clearly, the greater the diversity of users you intend to reach, the lower the possibility of enabling domain-specific queries. For example, an experienced user of intensive longitudinal surveys in psychology will know precisely what to expect when searching for "ESM" data, but the acronym will be unusual or of little interest to other domains. For this reason, the LivePeople Catalog has been enabled with basic and as general search functions as possible, considering:

- the location in which the dataset was collected
- the acronym of the project that collected the dataset
- the dataset type (see Section 3.5.1.2)

Location is one of the critical components of the situational context, functions as a vehicle for many aspects of knowledge, and is a proxy for cultural aspects. The project acronym instead filters all the datasets collected following a specific research methodology. Finally, the type of dataset facilitates research for analysis purposes. Further aspects that may be considered are, for example, the research topic and the methodology adopted (for which a standard and interdisciplinary dictionary is necessary), as well as the duration of the studies or the size of the sample. Future implementations must consider the datasets' compositional aspects, as described in Section 3.5.3.1.

3.5.3 Download

Big Thick Data is exceptionally varied and, therefore, eminently multi-purpose as it concerns different aspects of a person's daily life and lends itself to multiple types of reuse. Furthermore, their variety and being person-centred pose a considerable risk for de-anonymization, particularly indirect. Since datasets are modular, it is difficult to predict all combinations that will lead to the identification of the subjects. For this reason, the download phase involves the definition of potential (i) data combinations and a (ii) procedure for their sharing, as reported in the following sections.

3.5.3.1 Type of datasets available for download

Since Big Thick Data is particularly diverse, it may have different uses and, thus, different combinations of datasets. Three types of dataset combinations are considered according to the spatial and temporal dimensions of the data collection:

1. *Basic*: (also called single location) a dataset composed of all the participants in one study, in one location, at one point in time. The composition of a Basic dataset is described in Section 3.5.1.2

2. *Multiple*: (also called multiple locations) a dataset composed of all the participants in one study, in multiple locations, at one point in time

3. *Composite*: (also called longitudinal) a dataset composed of all the participants that participated in different studies, in (potentially) multiple locations, at different points in time.

In other words, the user interested in studying people's behaviour in a single place will request a Basic dataset, while to carry out a comparative study, he will require a Multiple datasets. A composite dataset will be necessary if the user is interested in the evolution of behaviours over time (in the exact location or multiple locations).

This division of the datasets is proposed here mainly for technical reasons, as the different composition complexity will require different merge functions and a longer preparation duration. Note, however, that as the complexity of the data increases, the potential risk related to privacy and copyright also increases, not only because the datasets could belong to different owners but also because the amount of information can easily lead to de-anonymization. For example, considering the location, identifying a person's home may not entail a high risk if the dataset was collected several years before the analysis, as the person may have changed home in the meantime. On the contrary, if the same domicile is identified years later (as can happen in a Multiple dataset), the risks associated with deanonymization are more significant.

3.5.3.2 Data download request

Finally, it is necessary to conduct an application procedure to download the dataset since it is potentially de-anonymizable (see [97]). The request is based on drafting a proposal describing the data's purpose and methods of use and signing a contract regulating its use. Some relevant licensing conditions are: (i) the datasets may only be used for research purposes; (ii) redistribution of the datasets is forbidden; (iii) once downloaded, the datasets cannot be made public (e.g., on a website). See Appendix B.6 for a description of the download procedure and accompanying documentation based on the EUROSTAT approach to microdata management [100].

Once approved, the data can be shared using a Distribution Repository (DREP). The DREP is created to share collected information according to the requested download and following GDPR. Indeed, DREP is kept for the time necessary to download the data. To facilitate data retrieval, DREP includes a folder for each data set provided to a single institution.

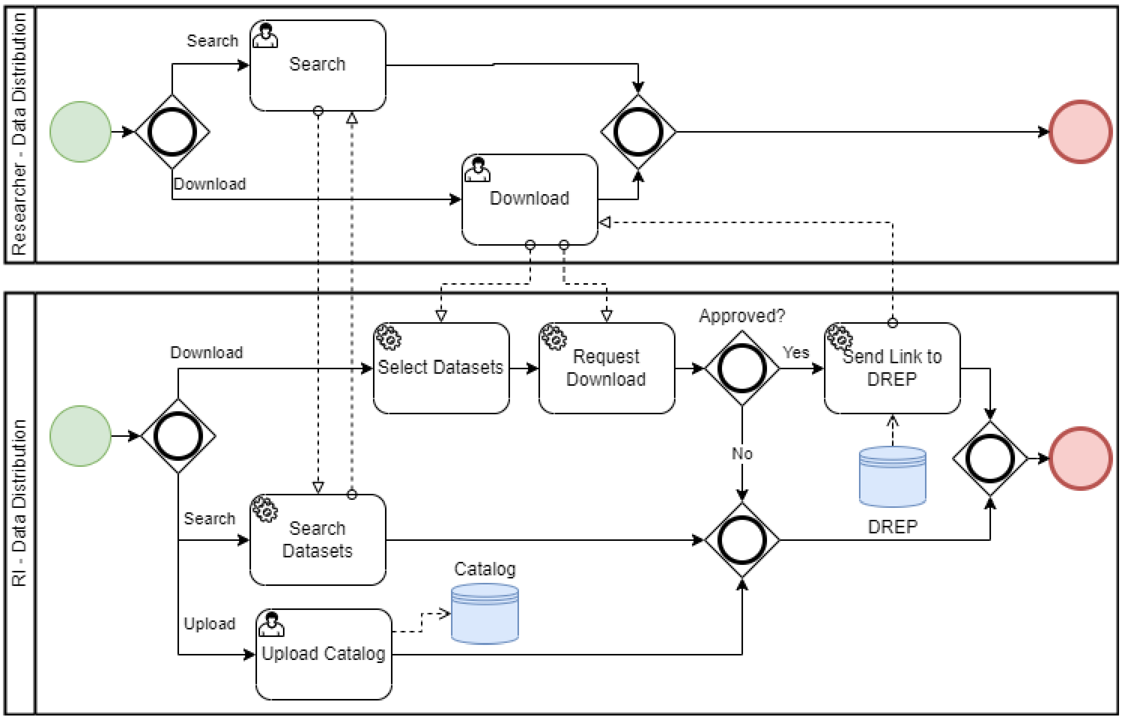
3.5.4 Data distribution BPMN and support material

The sub-processes of the overall service of Data Distributions show the steps and data storage used for searching or downloading the data collected in the previous processes. The BPMN process is defined in Figure 3.6 and detailed in the tables 3.27,3.28,3.29,3.30,3.31 has been defined. This process is associated with various templates and privacy materials helpful in this phase, which can be found in Appendix B.6, and the description of the user journey described in Appendix E.

Name	Type	Contained in	Description
Researcher	Pool	-	The user of the system that requests the data to be uploaded, searched, or downloaded
Data distribution	Pool	-	The main service for supporting the researcher in the data distribution process

Table 3.27: Data distribution pools and lanes

Figure 3.6: Data distribution phase BPMN



Name	Type	Definition	Description
Begin.Researcher	Start	Start request for data distribution	When the researcher requests for data to be uploaded, searched, or downloaded
End.Researcher	End	Upload, Search or Download	When the researcher has concluded the upload, search or download process
Begin.Distribution	Start	Start data distribution	When the support receives the request to upload, download, or search the datasets
End.Distribution	End	End data distribution	When the service of upload, search or download is finalized

Table 3.28: Data distribution events

The Researcher Lane (see Figure 3.6 and Table 3.29) shows the main activities conducted by the researcher. Ideally, the researcher will be able to use the search function to find the datasets of interest and make a request to download the data via the appropriate documentation downloadable from the web page.

Name	Type	Description	Timing
Upload	User	The researcher prepare the data and documentation for data distribution	5 days
Search	User	Process for searching the datasets within the Catalog	.
Download	User	Process for requesting and downloading the data	15 days

Table 3.29: Researcher - Data distribution tasks

Name	Type	Description	Timing
Upload Catalog	Manual	Process for uploading metadata and documentation on the Catalog the data and converting them to parquet	5 days
Search datasets	Service	Service for searching the data	.
Select datasets	Service	Service for selecting the data	.
Request download	Service	Service and documents for requesting the data	10 days
Send link to DREP	Service	Service for uploading the data on DREP and share the link with the researcher	1 day

Table 3.30: Data distribution support - Data distribution tasks

Name	Type	Direction	Description
Inclusive Gateway 1	Inclusive	Converging	Researcher Gateway for choosing the activities to be done (upload, download, search)
Inclusive Gateway 2	Inclusive	Converging	Data Distribution Gateway dealing with the request on the activities to be done (upload, download, search)
Inclusive Gateway 3	Inclusive	Converging	Data Distribution Gateway for approving or redefining the request for data download
Inclusive Gateway 4	Inclusive	Converging	Data Distribution Gateway dealing with finalization of the search, upload, or download procedures

Table 3.31: Data distribution Gateways

The Support Lane (see Figure 3.6 and Table 3.30) describes the activities conducted within the LivePeople Catalog, based on a set of services offered, from the Upload function to the Search and Download function. Before uploading, the dataset must be validated by a technical support expert to ensure the quality and reliability of the data. Furthermore, once the download request has been received, an expert will have to validate it and compose the set of datasets (based on what is described in Section 3.5.3.1) and then upload them to a repository accessible to the network called DREP, from which the researcher will have 24 hours to download it.

4

Case Study I: Diversity and Big Thick Data

Contents

This study [1] is part of the WeNet project (see Section 1.4), which aimed to study and foster the diversity of people mediated by technologies. The study is an example of a *Multipurpose* study design, that is, an approach to the daily life of students considering the main aspects of the context and their articulation over time in

[1] The current chapter is an adaptation technical report [174]

terms of routines and social practices. In other words, it is an example of a Big Thick Data collection that considers both annotations about the Personal Context and sensor data in the Object context. The results and their reuse bring multiple insights about the validity and reliability of the iLog methodology.

4.1 Topic and objectives

The study aimed to investigate the diversity of students' daily routines. Diversity is modelled based on earlier work in the social sciences [30, 191–198]. Along this line of thought, *surface* (i.e., objectual) and *deep* (i.e., personal) diversity fall under the broader area of theory of *social practices*, when studied at the group level, and of *behavioural routines* when studied at the individual level [199–209].

Social practices are modelled in terms of three components [210] as follows:

- *Material*, namely the material objects, such as a car or a membership, which allow the exploitation of a certain practice,
- *Competence*, namely the knowledge, skills, and abilities that enable a certain practice, and
- *Meaning*, namely a set of elements that give meaning to one or more social practices. It primarily captures the cultural component present in a specific society, as expressed by a subject, while, at the same time, motivating this subject to perform that particular practice.

Thus, for instance, travelling by public transport may be motivated by personal attention to the environment. In the same way, being careful about the climate justifies (and gives meaning to) the fact that one uses a bicycle, sorts garbage, becomes vegetarian, and so on. The theory of social practices identifies three critical components used to define behavioural routines: material, competence, and meaning. These components can also be recombined to give rise to new and different behavioural routines. No matter how generated, behavioural routines become social practices if recognised at the community level. [211–213].

This study has developed indicators and tools that allow the ment of measurement of diversity across the three dimensions underlying social practices (competence, material, and meaning) and how they are organised and performed in daily routines while maintaining a level of comparability across university students from different countries and cultural communities.

4.2 Measurement and tools

The study was based on three types of data collection tools, namely (i) closed-ended questionnaires (synchronic), (ii) time diaries, and (iii) sensor data (the last two being diachronic).

4.2.1 Questionnaires

Diversity is a complex, multidimensional, and multi-layered phenomenon. In other words, it is a latent concept that cannot be captured as a whole with a single measuring instrument. Its analysis requires decomposing diversity into elementary parts that can be measured and reconstructed. The study focuses on a few specific subsets of diversity, as seen from the topics addressed in the three questionnaires administered to the students.

- The first questionnaire was administered to the whole population to collect broad general information related mainly to surface diversity, cultural consumption and leisure (deep diversity), and some dimensions pertaining to social relations (online and offline).
- The second questionnaire was administered only to iLog participants and was mainly devoted to finding deep diversity information. This questionnaire was mainly focused on exploring specific social practices, such as moving, cooking, grocery shopping, and physical activities.
- The third and last questionnaire was administered only to iLog participants and was mainly devoted to finding deep diversity information. This questionnaire explored the user's experience with the app and testing a multiple intelligence scale.

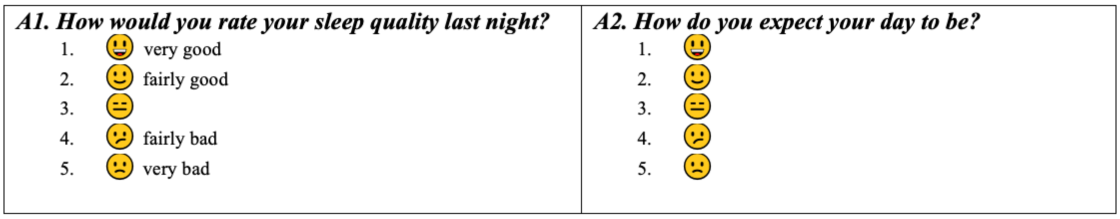
All three questionnaires gathered information related to material, competence, and meaning. In particular, four standard scales were used as a proxy for meaning and one as a proxy for competence. Regarding meaning, the following scales were used:

- two scales about personality, namely the Big Five Inventory [49] in the main questionnaire, and a Jungian scale on personality types [214–218] in the second questionnaire);
- two scales about values, namely the Basic Values Survey [219] in the main questionnaire and the Human Values Survey [79, 220] in the second questionnaire

Concerning competence, the multiple intelligence scale [221] was administered in the third questionnaire. Each question and scale can provide elementary information on specific diversity characteristics. Their combination, in turn, can be used as a complex measure of diversity in specific social practices.

4.2.2 Time diaries

Time Use Diary (TUD) is an intensive longitudinal survey (ILS) approach that observes how individuals spend their time. TUDs measure the frequency and duration of human activities, behaviours, and experiences, offering a detailed view of social behaviour. In a diary study, data are self-reported activity sequences in time episodes ranging from a few days to even a month or longer with a regular time interval. This data type is usually collected via a self-completed time diary [64] that allows registering (at fixed time intervals) the sequence of an individual's activities. In this sense, they are an excellent tool for measuring people's context and how this evolves (see Section 1.2). Indeed, the time diary approach allows one to observe people's perspectives on their daily lives based on where they are, who they are with, what they are doing and how they feel at different times of the day.



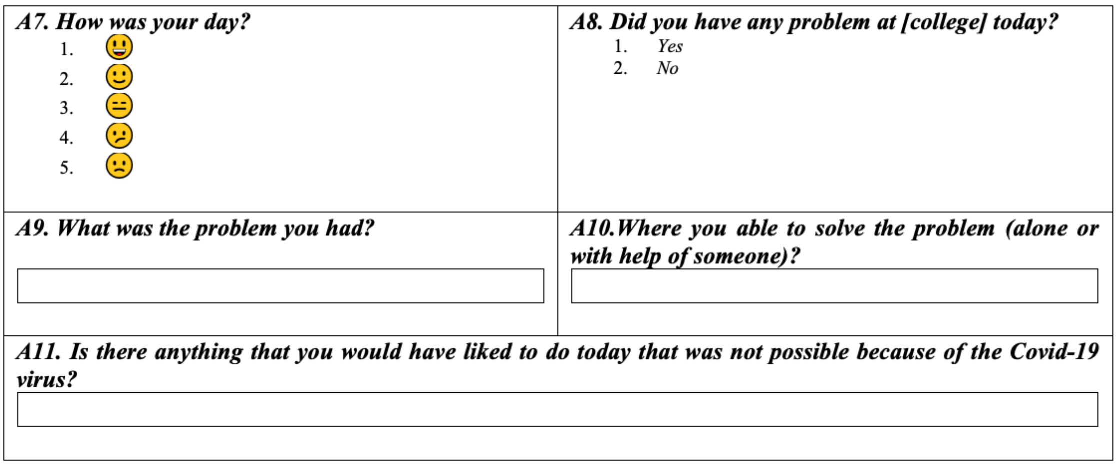
A1. How would you rate your sleep quality last night?

1. 😃 very good
2. 🙂 fairly good
3. 😑
4. 😕 fairly bad
5. 🙁 very bad

A2. How do you expect your day to be?

1. 😃
2. 🙂
3. 😑
4. 😕
5. 🙁

Figure 4.1: Morning questions sent using iLog



A7. How was your day?

1. 😃
2. 🙂
3. 😑
4. 😕
5. 🙁

A8. Did you have any problem at [college] today?

1. *Yes*
2. *No*

A9. What was the problem you had?

A10.Where you able to solve the problem (alone or with help of someone)?

A11. Is there anything that you would have liked to do today that was not possible because of the Covid-19 virus?

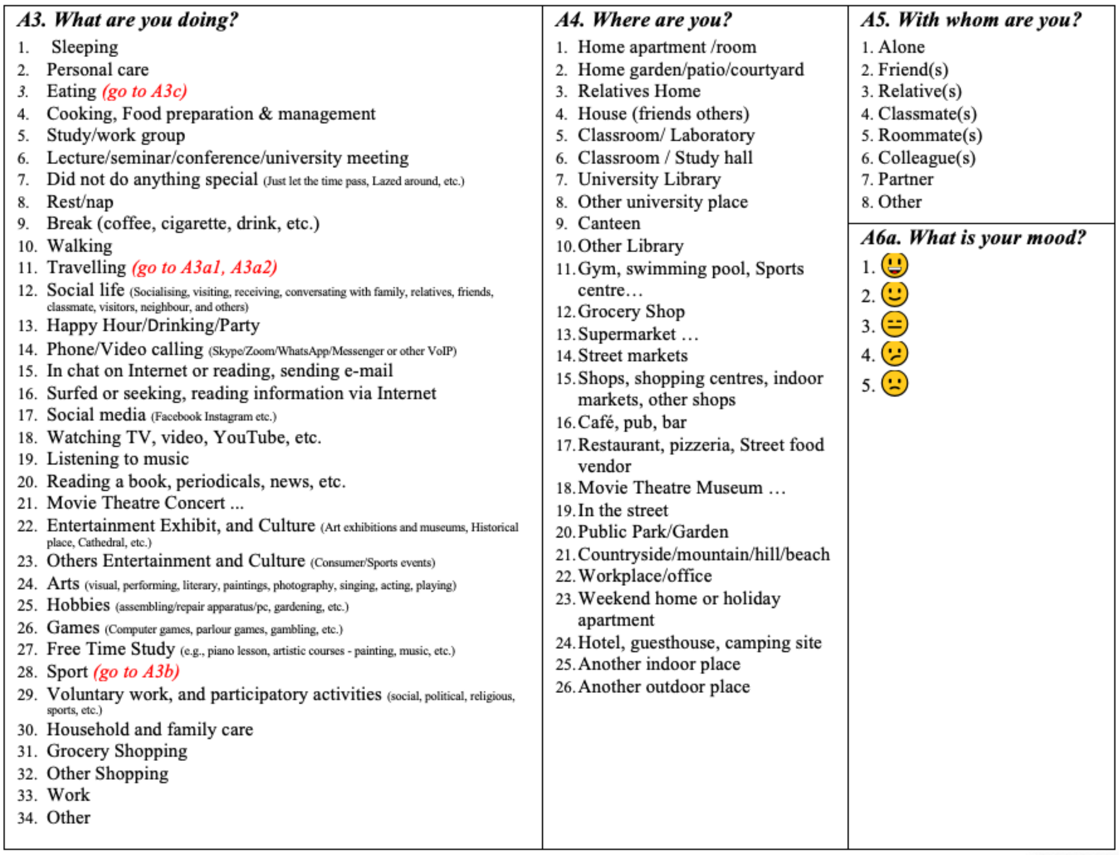
Figure 4.2: Evening questions sent using iLog

The ILS is composed of three different time diaries with different timings and different objectives, as follows:

- The first diary collects information about the day's beginning and end. Every time, at 08:00 AM (Figure 4.1), the subject received two qualitative questions about the sleep quality and the day's expectations. At 10:00 PM, (Figure 4.2) the subjects were asked (a) to rate their day, (b) if they had any problems during the day, and (c) how they solved them, and, finally, they received a (d) question about the COVID-19 pandemic.
- The second is a standard time diary (Figure 4.3) with special sections on three main activities. Every half hour for the first two weeks and every hour for the second two weeks, the participants received a notification on their smartphone with four questions as follows:
  - their activity "What are you doing?" providing the participant with 34 answer categories such as sleeping, eating, working, etc.;

- the current location "Where are you?" providing the participant with 26 categories such as home, workplace, university, restaurant, etc.;
- the persons being with the participants at the time of the question "Who is with you?" providing the participant with eight categories such as "Alone", "With my partner", "With friends", etc; and
- their mood "What is your mood?" providing the participant with a scale of 5 levels ranging from happy to sad.

If the subject claimed to be "eating," "travelling," or "doing sport," four different in-depth questions were asked for further information (Figure 4.4). Specifically



A3. What are you doing?

1. Sleeping
2. Personal care
3. Eating *(go to A3c)*
4. Cooking, Food preparation & management
5. Study/work group
6. Lecture/seminar/conference/university meeting
7. Did not do anything special (Just let the time pass, Lazed around, etc.)
8. Rest/nap
9. Break (coffee, cigarette, drink, etc.)
10. Walking
11. Travelling *(go to A3a1, A3a2)*
12. Social life (Socialising, visiting, receiving, conversating with family, relatives, friends, classmate, visitors, neighbour, and others)
13. Happy Hour/Drinking/Party
14. Phone/Video calling (Skype/Zoom/WhatsApp/Messenger or other VoIP)
15. In chat on Internet or reading, sending e-mail
16. Surfed or seeking, reading information via Internet
17. Social media (Facebook Instagram etc.)
18. Watching TV, video, YouTube, etc.
19. Listening to music
20. Reading a book, periodicals, news, etc.
21. Movie Theatre Concert ...
22. Entertainment Exhibit, and Culture (Art exhibitions and museums, Historical place, Cathedral, etc.)
23. Others Entertainment and Culture (Consumer/Sports events)
24. Arts (visual, performing, literary, paintings, photography, singing, acting, playing)
25. Hobbies (assembling/repair apparatus/pc, gardening, etc.)
26. Games (Computer games, parlour games, gambling, etc.)
27. Free Time Study (e.g., piano lesson, artistic courses - painting, music, etc.)
28. Sport *(go to A3b)*
29. Voluntary work, and participatory activities (social, political, religious, sports, etc.)
30. Household and family care
31. Grocery Shopping
32. Other Shopping
33. Work
34. Other

A4. Where are you?

1. Home apartment /room
2. Home garden/patio/courtyard
3. Relatives Home
4. House (friends others)
5. Classroom/ Laboratory
6. Classroom / Study hall
7. University Library
8. Other university place
9. Canteen
10. Other Library
11. Gym, swimming pool, Sports centre…
12. Grocery Shop
13. Supermarket …
14. Street markets
15. Shops, shopping centres, indoor markets, other shops
16. Café, pub, bar
17. Restaurant, pizzeria, Street food vendor
18. Movie Theatre Museum …
19. In the street
20. Public Park/Garden
21. Countryside/mountain/hill/beach
22. Workplace/office
23. Weekend home or holiday apartment
24. Hotel, guesthouse, camping site
25. Another indoor place
26. Another outdoor place

A5. With whom are you?

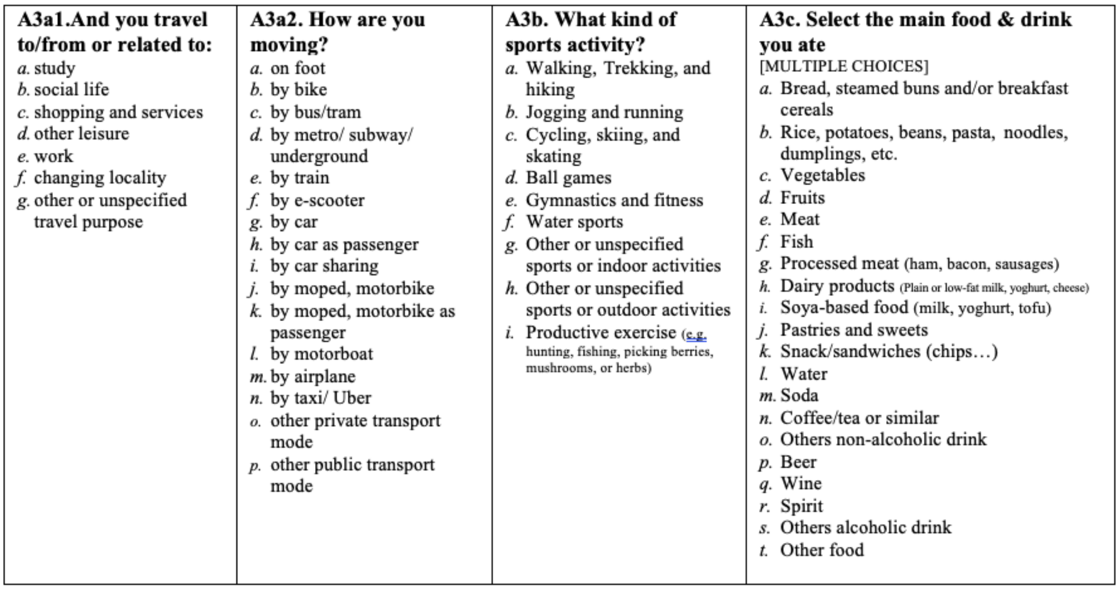
1. Alone
2. Friend(s)
3. Relative(s)
4. Classmate(s)
5. Roommate(s)
6. Colleague(s)
7. Partner
8. Other

A6a. What is your mood?

1. 😃
2. 🙂
3. 😑
4. 😕
5. ☹

Figure 4.3: Standard time diary with the questions sent every 30 minutes using iLog

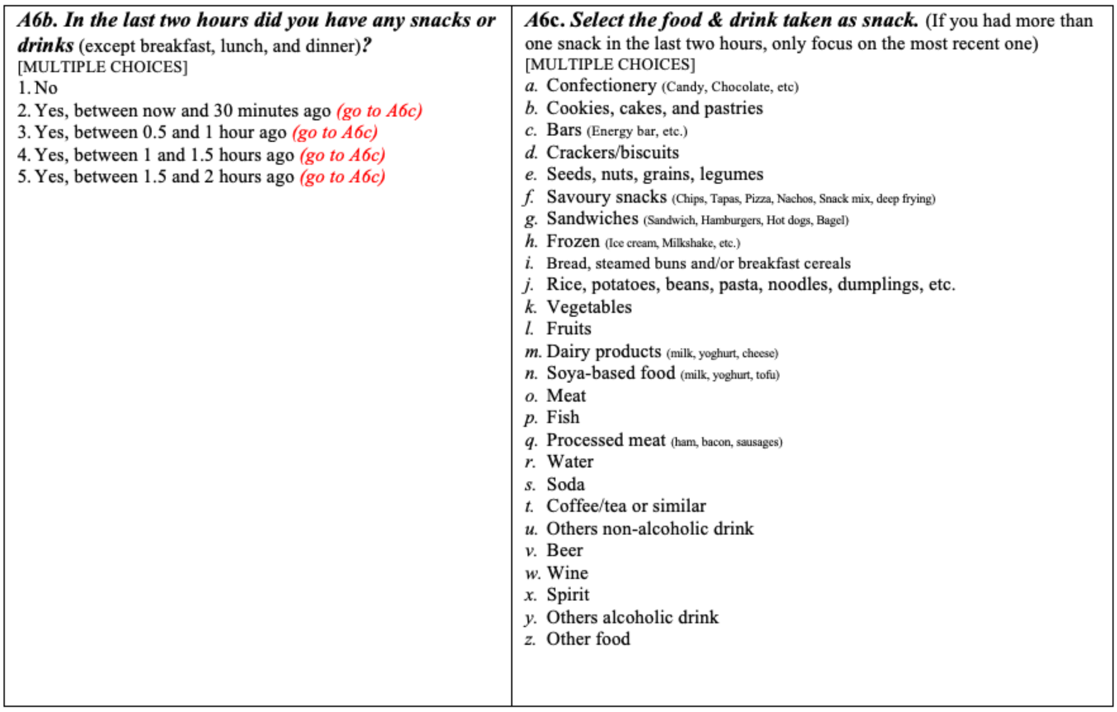
- when eating, the subject had to report foods and drinks selecting them from 20 categories, such as rice, potatoes, meat, beer, etc. (adapted from [222]);
- when doing sport, the subject had to state the type of sport, selecting them from 9 categories, such as jogging and running, water sports, etc;
- when travelling, the subject had to state (a) the reason for the travel within seven categories, such as study, social life, etc., and (b) the means



A3a1.And you travel to/from or related to:	**A3a2. How are you moving?**	**A3b. What kind of sports activity?**	**A3c. Select the main food & drink you ate**
a. study *b.* social life *c.* shopping and services *d.* other leisure *e.* work *f.* changing locality *g.* other or unspecified travel purpose	*a.* on foot *b.* by bike *c.* by bus/tram *d.* by metro/ subway/ underground *e.* by train *f.* by e-scooter *g.* by car *h.* by car as passenger *i.* by car sharing *j.* by moped, motorbike *k.* by moped, motorbike as passenger *l.* by motorboat *m.* by airplane *n.* by taxi/ Uber *o.* other private transport mode *p.* other public transport mode	*a.* Walking, Trekking, and hiking *b.* Jogging and running *c.* Cycling, skiing, and skating *d.* Ball games *e.* Gymnastics and fitness *f.* Water sports *g.* Other or unspecified sports or indoor activities *h.* Other or unspecified sports or outdoor activities *i.* Productive exercise (e.g. hunting, fishing, picking berries, mushrooms, or herbs)	[MULTIPLE CHOICES] *a.* Bread, steamed buns and/or breakfast cereals *b.* Rice, potatoes, beans, pasta, noodles, dumplings, etc. *c.* Vegetables *d.* Fruits *e.* Meat *f.* Fish *g.* Processed meat (ham, bacon, sausages) *h.* Dairy products (Plain or low-fat milk, yoghurt, cheese) *i.* Soya-based food (milk, yoghurt, tofu) *j.* Pastries and sweets *k.* Snack/sandwiches (chips…) *l.* Water *m.* Soda *n.* Coffee/tea or similar *o.* Others non-alcoholic drink *p.* Beer *q.* Wine *r.* Spirit *s.* Others alcoholic drink *t.* Other food

Figure 4.4: In-depth questions that appear when certain options are selected in the question "What are you doing?"

of transport within 16 categories, such as car, bus, etc.



A6b. In the last two hours did you have any snacks or drinks (except breakfast, lunch, and dinner)***?***	***A6c. Select the food & drink taken as snack.*** (If you had more than one snack in the last two hours, only focus on the most recent one)
[MULTIPLE CHOICES] 1. No 2. Yes, between now and 30 minutes ago *(go to A6c)* 3. Yes, between 0.5 and 1 hour ago *(go to A6c)* 4. Yes, between 1 and 1.5 hours ago *(go to A6c)* 5. Yes, between 1.5 and 2 hours ago *(go to A6c)*	[MULTIPLE CHOICES] *a.* Confectionery (Candy, Chocolate, etc) *b.* Cookies, cakes, and pastries *c.* Bars (Energy bar, etc.) *d.* Crackers/biscuits *e.* Seeds, nuts, grains, legumes *f.* Savoury snacks (Chips, Tapas, Pizza, Nachos, Snack mix, deep frying) *g.* Sandwiches (Sandwich, Hamburgers, Hot dogs, Bagel) *h.* Frozen (Ice cream, Milkshake, etc.) *i.* Bread, steamed buns and/or breakfast cereals *j.* Rice, potatoes, beans, pasta, noodles, dumplings, etc. *k.* Vegetables *l.* Fruits *m.* Dairy products (milk, yoghurt, cheese) *n.* Soya-based food (milk, yoghurt, tofu) *o.* Meat *p.* Fish *q.* Processed meat (ham, bacon, sausages) *r.* Water *s.* Soda *t.* Coffee/tea or similar *u.* Others non-alcoholic drink *v.* Beer *w.* Wine *x.* Spirit *y.* Others alcoholic drink *z.* Other food

Figure 4.5: Additional questions related to food and drinks

- In the third time diary (Figure 4.5), the subjects received additional questions about food and drinks. These questions were asked every two hours outside the main meal hours.

4.2.3 Sensor data

All sensor data listed in Appendix C.1 were collected as frequently as possible to ensure the dataset reuse across multiple domains.

4.3 Ethics and privacy

All the study activities and results at each site comply with academic and national ethical privacy-protecting laws and guidelines. Additionally, for non-European Studies, the activities and results have been developed to be compliant with those of a selected European country, as requested by the European Commission. The Italian legislation was selected as the reference. The details are described in [30, 213]

4.4 Incentives strategy

The incentive strategy was aimed at maintaining active participation during the second and third phases of the investigation, where the dropout rates were higher and had a greater impact on the results. The incentives were designed based on the effort required by the different tasks, as described below:

1. Payments for completing at least the 85% of:
   - the 1st two weeks of the ILS;
   - the 2nd two weeks of the ILS;
2. Daily prizes (random extraction)
3. Final prizes (random extraction), for:
   - the 1st two weeks of the ILS;
   - the 2nd two weeks of the ILS.

Table 4.1 shows how the remuneration was adjusted according to the basket of goods that can be purchased in each country.

4.5 Study protocol

The whole data collection process was identically applied in all the pilot sites. Various roles (see Section 3.1.6) were identified, including local Study Leaders in charge of questionnaire translation and administration, recruitment and field supervisors, and a local Research Leader appointed as data controller. The role of Technology Leader was mainly conducted by UNITN, which supported all the different tasks and phases. These organisational details and the ethical and legal aspects are described in [30].

	Payments		Daily	Final Prizes	
	1st Weeks	2nd Weeks	Prizes	1st Weeks	2nd Weeks
AAU	150 kr	150 kr	5 of 40 kr	3 of 800 kr	3 of 1200 kr
LSE	0	0	0	£150 (1/50)	£150 (1/50)
NUM	10k MNT	10k MNT	5k MNT	100k MNT	150k MNT
UC	25k GS	25k GS	10 vouchers	1 restaurant voucher	1 restaurant voucher
UNITN	20 €	20 €	5 of 5 €	3 of 100 €	3 of 150 €
JLU	100 rmb	100 rmb	1 of 20 rmb	3 of 88 rmb	3 of 88 rmb
IPICYT	0	0	0	0	0
AMRITA	0	0	0	0	0

Table 4.1: Incentives

The data collection process lasted approximately 13 weeks, involving participants from different countries (see Figure 4.7 and Table 4.2). The process was articulated in the following phases (see Figure 4.6):

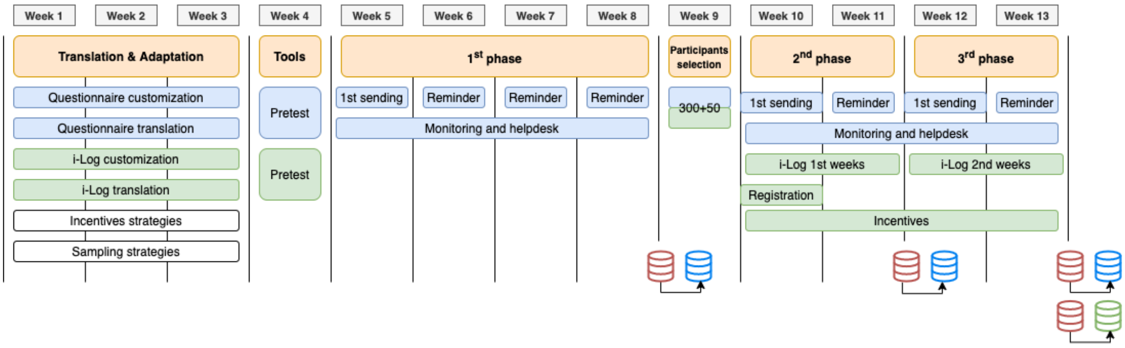


Figure 4.6: Steps and phases of the data collection process.

1. *Translation and Adaptation.* In this phase, each site received the English version of the questionnaires and the app, including the time diaries and the list of sensors to be collected. In coordination with all the partners, these tools were evaluated and adapted to the specific context (e.g., invitation letters, type and amount of incentives for the participants of iLog, privacy, and ethics documentation). Some countries made minimal changes to better adapt the questionnaire to the local situation or academic organisation. Concerning the standard scales mentioned above, the translations were completed by a forward translator from the original English version and then validated via panel and back-translation processes by independent translators.

2. *Tools.* After translation and adaptation, the tools were tested locally. A first test was conducted to check and validate the translations and evaluate the tools' usability. A second test was conducted by sending the questionnaires to

a small sample of participants, both project partners and students from the various universities. As far as questionnaires were concerned, approximately 30 participants were involved. This test was also used to ascertain the completion times. Concerning iLog, a two-week validation test was conducted.

3. *First phase.* This was the first of the three phases of the data collection. This phase started by sending an email containing the description of the study, the invitation to the first questionnaire, and information on the second part of the study. This invitation was then reiterated through 4 weekly reminders to all students who had not completed the first questionnaire. Some changes concerned the first questionnaires at JLU and IPICYT. In JLU, the first questionnaire was shown on social channels and various WeChat groups, involving about 5000 students; in this case, the reminders were not sent directly to the participants but posted on the social channels. In the case of IPICYT, the recruitment took place through direct involvement of the participants, for which the reminders were made by voice and through messages on the WhatsApp group created specifically for the study.

4. *Participants selection.* At this stage, a subset of the eligible participants was selected to participate in the second part of the study. The requirements were two: having consented to the processing of personal data and being in possession of a smartphone compatible with the app. In the case of IPICYT, this phase occurred before the first questionnaire was sent.

5. *Second phase.* This phase started by sending the second questionnaire to the selected subset of participants, followed by a reminder after one week. When sending the second questionnaire, an email with instructions on downloading iLog was sent, accompanied by a short specification manual.

6. *Third phase.* The final questionnaire was sent during this last phase, followed by a reminder one week later. It is worth noticing that, during this phase, the frequency of administration of time diaries via iLog was reduced.

7. *Closing the study.* At the end of the ILS, a last email was sent with the steps to follow before uninstalling the app and a last reminder, where needed, to fill in the second and third questionnaires.

4.5.1 Monitoring

To facilitate the iLog ILS monitoring and identify possible problems, daily reports were produced containing (1) the number of notifications each participant responded to and (2) the amount of data collected by the individual sensors. Using this information, the local field supervisors could contact the inactive participants every three days and support them as needed. A further element of contact was the daily sending of the results of the daily prize (see below the description of incentives).

4.6 Results

As described in Figure 4.7 and reported in Table 4.2, the Diversity pilot involved more than 20,000 students in 8 countries. Of these, 350 participants were selected, balancing the sample based on gender and department, as different faculties have different lesson times, which can affect students' daily behaviour. If there were not 350 participants from the first survey, all available participants were invited. Of all the participants invited, 757 have installed the iLog app, responding to approximately 216,000 TUD notifications, reaching more than 216k situational context collected every half an hour. NUM and UNITN reached the highest number of participants, 228 and 267 respectively. At the same time, in the other cases, more than 40 students participated in the data collection, except for AAU (26 participants) and IPICYT (21 participants).

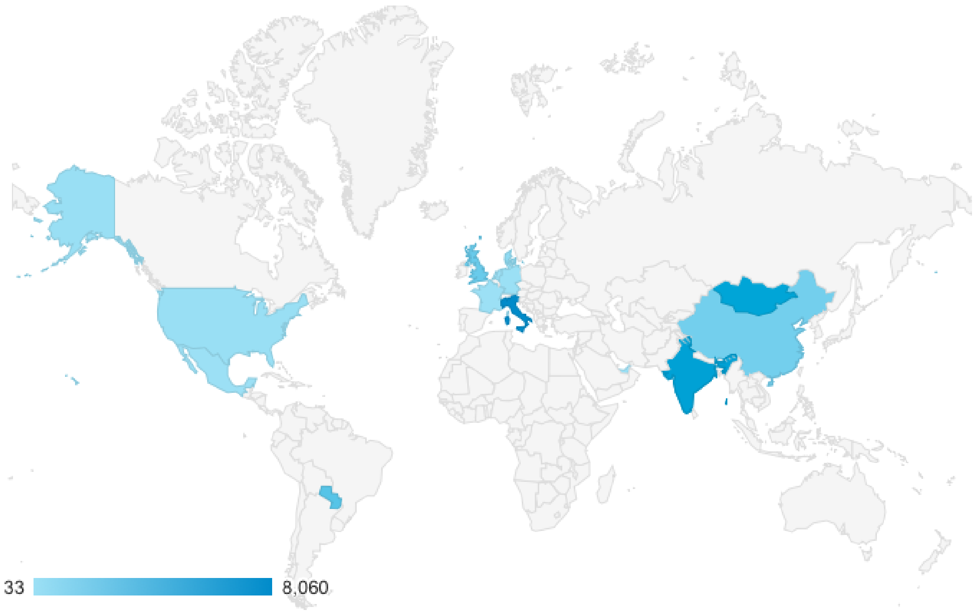


Figure 4.7: Questionnaire and ILS access from countries

The Tables 4.3, 4.4, and 4.5 report the average of the daily responses per participant, divided into the two parts of the ILS (48 possible daily responses for the first two weeks and 24 for the second two). The tables are also divided into quintiles.

As the tables demonstrate, in all pilots (except AMRITA), 40% of participants provided an average of 20 responses per day or approximately 10 hours of annotated sensors. The average daily response rate increases significantly in UNITN, JLU and NUM.

Table 4.6 reports the percentage of participants who provided data from the sensors, divided by each sensor. In most cases, at least 35% of the participants provided data from all the sensors except AMRITA and in some cases where specific sensors

Site	1st QU	2nd QU	3rd QU	iLog
AAU	412	16	15	27
LSE	1980	143	45	86
NUM	3972	214	152	224
UC	1342	33	25	42
UNITN	5692	287	215	238
JLU	989	136	82	54
IPICYT	88	0	21	21
AMRITA	6598	256	71	65
Total	21073	1085	626	757

Table 4.2: Participants per pilot site during the three waves and the iLog data collection.

	AAU			LSE			UNITN		
quint	**P.**	**1°w**	**2°w**	**P.**	**1°w**	**2°w**	**P.**	**1°w**	**2°w**
1	6	0,6	0,0	18	0,0	0,0	54	0,9	0,0
2	5	4,7	0,0	18	0,9	0,0	54	15,9	2,4
3	5	17,8	2,1	17	7,0	0,0	53	35,7	15,7
4	5	40,2	18,3	17	27,2	3,3	53	44,1	22,1
5	5	45,5	23,4	17	44,6	21,2	53	46,6	24,1
Total	26			87			267		

Table 4.3: Average daily responses from participants for each country (part 1)

	AMRITA			IPICYT		JLU		
quint	**P.**	**1°w**	**2°w**	**P.**	**M1**	**P.**	**1°w**	**2°w**
1	13	0,0	0,0	5	2,7	9	1,1	0,0
2	13	0,0	0,0	4	6,6	9	16,9	5,2
3	12	0,2	0,0	4	11,8	9	37,9	16,6
4	12	1,8	0,0	4	27,2	9	42,9	20,5
5	12	16,5	4,8	4	38,0	9	47,0	23,6
Total	62			21		45		

Table 4.4: Average daily responses from participants for each country (part 2)

	NUM			UC		
quint	**P.**	**1°w**	**2°w**	**P.**	**1°w**	**2°w**
1	46	0,6	0,0	9	0,0	0,0
2	46	7,7	0,0	9	0,4	0,0
3	46	22,9	3,4	9	5,6	0,2
4	45	37,9	17,6	8	17,1	6,8
5	45	45,0	22,9	8	39,3	21,1
Total	228			43		

Table 4.5: Average daily responses from participants for each country (part 3)

did not work for the detection. The most dramatic cases are highlighted in red. However, it will be possible to notice that most cases are well above the threshold described, with multiple peaks of participation exceeding 85%.

Dataset	**AAU**	**AMR**	**IPI**	**JLU**	**LSE**	**NUM**	**UC**	**UNI**
Accelerometer	80,8	32,3	95,2	84,4	77,0	76,3	69,8	86,9
Activities	65,4	30,6	95,2	37,8	67,8	75,0	67,4	77,9
Airplane Mode	11,5	9,7	33,3	11,1	19,5	18,4	18,6	36,0
Application	80,8	32,3	95,2	88,9	77,0	77,2	69,8	87,6
Battery Log	80,8	32,3	95,2	88,9	75,9	76,3	62,8	86,9
Batterycharge	80,8	32,3	95,2	88,9	77,0	77,2	69,8	87,6
Bluetooth LTE	57,7	17,7	90,5	66,7	55,2	23,7	41,9	60,7
Bluetooth	57,7	19,4	90,5	66,7	56,3	28,1	41,9	64,0
Cellular Network	80,8	24,2	95,2	88,9	74,7	65,8	62,8	93,6
Doze	76,9	25,8	95,2	66,7	72,4	69,7	53,5	80,5
Gyroscope	73,1	0,0	0,0	0,0	77,0	0,0	0,0	0,0
Headset Plug	42,3	14,5	61,9	75,6	41,4	66,2	37,2	62,9
Light	80,8	29,0	90,5	88,9	77,0	70,6	60,5	86,1
Location	80,8	25,8	95,2	88,9	72,4	63,2	62,8	82,8
Magnetic Field	73,1	29,0	85,7	8,9	75,9	64,9	53,5	82,8
Music	38,5	11,3	61,9	11,1	36,8	30,7	34,9	54,7
Notification	61,5	27,4	90,5	68,9	60,9	65,4	51,2	68,5
Pressure	23,1	4,8	33,3	20,0	35,6	28,5	18,6	16,5
Proximity	80,8	32,3	95,2	88,9	77,0	76,3	67,4	86,5
Ring Mode	61,5	24,2	76,2	57,8	60,9	73,7	58,1	77,9
Screen	80,8	32,3	95,2	88,9	77,0	77,2	69,8	87,6
Step Counter	65,4	29,0	90,5	88,9	64,4	62,7	48,8	69,3
Step Detector	38,5	29,0	85,7	57,8	58,6	54,8	39,5	47,9
Touch	65,4	19,4	95,2	84,4	62,1	73,2	62,8	75,3
User Presence	80,8	32,3	95,2	88,9	77,0	77,2	69,8	87,6
Wifi	80,8	27,4	95,2	80,0	75,9	73,7	62,8	86,1
Wifi Networks	80,8	24,2	95,2	88,9	73,6	63,6	65,1	83,5
N.	**26**	**62**	**21**	**45**	**87**	**228**	**43**	**267**

Table 4.6: Average percentage of participants that provided sensors data for each country

4.7 Data preparation

The data preparation followed the steps indicated in Section 3.4.

Regarding the *Personal Data Anonymization*, all personal information, i.e., *email address*, *home address*, *name and surname*, has been removed from each of the three types of datasets (online questionnaire, time diaries, and sensors), still making sure that the same unique identifier would be assigned to the same person across all three datasets.

Concerning *Network Anonymization*, there are three possible sources of re-identification, namely: (i) *WiFi-event*, which shows the WiFi network the smartphone is connected to; (ii) *cellular-network*, which shows the roaming network; and (iii) *WiFi-networks-event*, which shows the WiFi networks that are available in the environment. The relevant columns have been anonymised using unique identifiers. A hash function was applied to the WiFi network name, with a function that cannot be reversed (the SHA-256 cryptographic function is used to perform the hash).

Concerning *GPS Anonymization*, the main problem is that the position of a person, in particular, if joined with the specific time and day, leads very quickly to re-identification, in particular when a person is in places that are not too crowded (e.g., outside cities) or when collected for an extended period. The only solution is to make the spatio-temporal information more ambiguous. There are many ways of doing this, all having different consequences for the usability of the dataset for research. The GPS information of this data set has been anonymised in two different ways, as follows:

1. *Round Down* Here, the idea is that precision is deliberately truncated from the location sensor so that it becomes anonymous but in a way that is still useful for certain scientific purposes. Furthermore, the dates associated with each GPS point are truncated;

2. *Point of Interest (POI).* Here, the idea is to collect *only* those points where the user has spent more than a certain amount of time. In this dataset, a POI tag is added to the stream if latitude and longitude do not change for one minute. For each POI, the elapsed time in seconds is also added. GPS longitude and latitude readings are removed. The POI is selected to identify a general location (suburb, city, region) and the closest relevant places (bar, restaurant, lake, etc).

The procedure output is two datasets called *RoundDown* and *POI*, each containing all the other sensors. For privacy reasons, only one of the two datasets can be downloaded by the same research institution, as merging the data contained in the two would quickly lead to re-identification.

4.8 Data distribution

The dataset is available in .csv format and PARQUET.

The dataset's primary entry point documentation can be found in the dataset Catalog at [172].

This website contains information including:

1. The technical report of the data collection

2. The dataset metadata and codebook (summary statistics)

3. A sample of the dataset
4. Links to all articles that have been published through the dataset
5. The procedure to request the dataset

Concerning the documentation to request the dataset, a license must be signed before downloading the dataset to fully comply with GDPR. Some relevant licensing conditions are: (i) the datasets may only be used for research purposes; (ii) redistribution of the datasets is forbidden; (iii) once downloaded, the datasets cannot be made public (e.g., on a website).

4.9 Roles

The order of names is by contribution of the Institution and, inside each Institution, by contribution of the individuals. As such, the order of names does not necessarily reflect the importance of the contribution of the single individuals. The roles of the authors, presented by their initials, are as follows:

- *Study management*: F.G., I.B., A.D.G., Matteo Busso;
- *Study design*: F.G., I.B., A.D.G., Matteo Busso, R.C.A., G.V., D.G.P., L.M.;
- *Technical support*: M.R., M.Z., C.G., Matteo Busso;
- *Data Collection*: Matteo Busso, R.C.A., M.R., A.D.G, P.K., A.G., A.C., G.G., S.S, M.B., L.C., A.H., J.L.Z., H.X., D.S., S.D., C.N., S.R.C., A.R.M.;
- *Data Preparation and correction*: R.C.A., C.G., I.B.,Matteo Busso, D.G.P., L.M.

4.10 Discussion

Below are some considerations regarding the validity and reliability of the results and methodological considerations.

4.10.1 Validity

Numerous interdisciplinary publications have addressed the data collected, demonstrating the validity of the constructs used and the choice of standard measurements from the social sciences. Below is a description of the main articles based on the datasets collected.

Mobile social media usage [223]. This work was conducted with a previous version of the Diversity1 dataset (see [109], involving the sensor data called `Running Applications`, *WHAT* annotations, and questionnaire data. These variables were used to analyse the logs of social media apps and compare them to students' credits and grades. The results show a negative pattern of social media usage that significantly impacts academic activities. The main information used in this dataset was the objective data collected about the user interaction with the social apps, compared with the subjective description of users about their interactions with social media. Once more, and coherently with what is known in the state of the art, objective and subjective information was not aligned.

Dealing with people providing wrong annotations [62]. This work was conducted with a previous version of the Diversity1 dataset (see [109]. The data used in this paper are both sensors and time diaries. The sensors are hardware (i.g., gravity and temperature) and software (e.g., screen status, incoming calls). In this work, the authors propose redesigning a sceptical learning algorithm centred around Gaussian Processes (GPs) tested on this dataset. The results show the new algorithm works well at varying noise levels and when new classes are observed. This algorithm recognised user-provided feedback's inconsistency in subjective and objective labels.

Predicting human behaviour [61]. This work was conducted with a previous version of the Diversity1 dataset (see [109]. This study investigates the role played by four contextual dimensions based on the data about `Events` (Timestamp and Time Diaries), `Location` (GPS Position), and `Person` 's social ties (Time Diaries), on the predictability of individuals' behaviours. The analysis shows that any selected modalities are more predictable when the other modalities are available. In certain cases, the availability of these modalities is particularly crucial as their values are nearly impossible to guess. Furthermore, it shows that subjectivity, modelled in this dataset by the labels provided at run-time, substantially impacts predictability since, in the location recognition experiment, the authors found that subjective location annotations convey more information about activity and social ties than the information derived from GPS.

Learning the Personal Context [21]. This is very early work in recognising the objective and subjective context, with the final goal of analysing their mutual consistency. Sensors and Time Diaries are used in this work, which aims to design an ontological model representing the personal context within a learning process that integrates with machine learning. The situational model used by the dataset presented here was first presented in this paper.

Predicting personality from patterns of behaviour [224] This work was conducted with a previous version of the Diversity1 dataset (see [109]. The study examines how individuals' personality dimensions can be predicted based on behavioural information collected via sensor and log data. The study could be replicated with our dataset. It could benefit from the annotations regarding *WHAT WHERE* and *WHOM*, as well as the sensors, to predict the psycho-social traits collected in the questionnaire.

Generalization and Personalization of Mobile Sensing-Based Mood Inference Models [225] This work focused on *Mood* inference based on *Time Diaries* and s*sensor data* to assess the effect of geographical diversity on mood inference models. This is particularly relevant as the generalisation of the models is a hot topic in machine learning research. The study shows how the predictive models perform better in cross-cultural settings: continent-specific models outperform multi-country models.

Complex daily activities, country-level diversity, and smartphone sensing [226] This work focused on human behaviour prediction based on *Time Diaries* and *sensor data* to assess the effect of geographical diversity on mood inference models. The paper shows that despite the generic multi-country approach, the country-specific approach performed better, highlighting that the Big Thick Data approach helps advance human behaviour understanding through smartphones and machine learning.

4.10.2 Reliability

Starting from the tables described previously, it is possible to see how, from the point of view of participation in the questionnaires, the response rate was high. Regarding time diaries, participation was particularly unbalanced in different countries, where only UNITN and NUM reached exceptionally high numbers (well above similar studies). The dropout rate was exceptionally high, with peaks in the case of AMRITA, and the completion rate did not exceed 40% in many cases. From the point of view of ESM approaches, a much higher completion rate is generally expected (equal to 90%) even if the quantity of annotations provided in ESM is usually smaller and specific to some sub-portions of the context. From the point of view of other possible applications, in the previous section, it was demonstrated how the dataset is still (highly) performing. As regards the sensors, aside from AMRITA, the participants from the other countries provided a high number of sensors, in line with state-of-the-art (see [226]).

4.10.3 Methodological considerations

Despite the exceptional results achieved in some cases of the data collection (taking into account that it was carried out during the COVID-19 pandemic, during which it was difficult to get in touch with the participants, if not only via email), it is possible to pose some critical reflections on the progress of the study in order to contextualise the disparity in the data.

Firstly, as highlighted by [227], some problems were encountered in adapting the studies during the preparation phases of the questionnaires and ILS. The presence of experts in different disciplines made communication complex, disadvantaging those who did not have a background in longitudinal questionnaires and ILS design.

In addition to the adaptation problem, it was difficult to obtain approval from the ethics committees of the various institutes, which in some cases took more than nine months.

Furthermore, the disparity in the proposed incentives has undoubtedly affected the success of data collection in some contexts. Leaving aside the cases in which there were no incentives, these results demonstrate what [78] stated, namely that payment in gifts or prize draws is less functional than the definition of a fixed salary for the participants.

A final problematic aspect concerned the management and preparation of data, both during and after data collection. Leaving aside the second, it was clear that the time between the first questionnaire sent and the actual start of the study with iLog was excessive, increasing the dispersion of participants.

5

Case Studies II: Verticalizing Big Thick Data in interactions

Contents

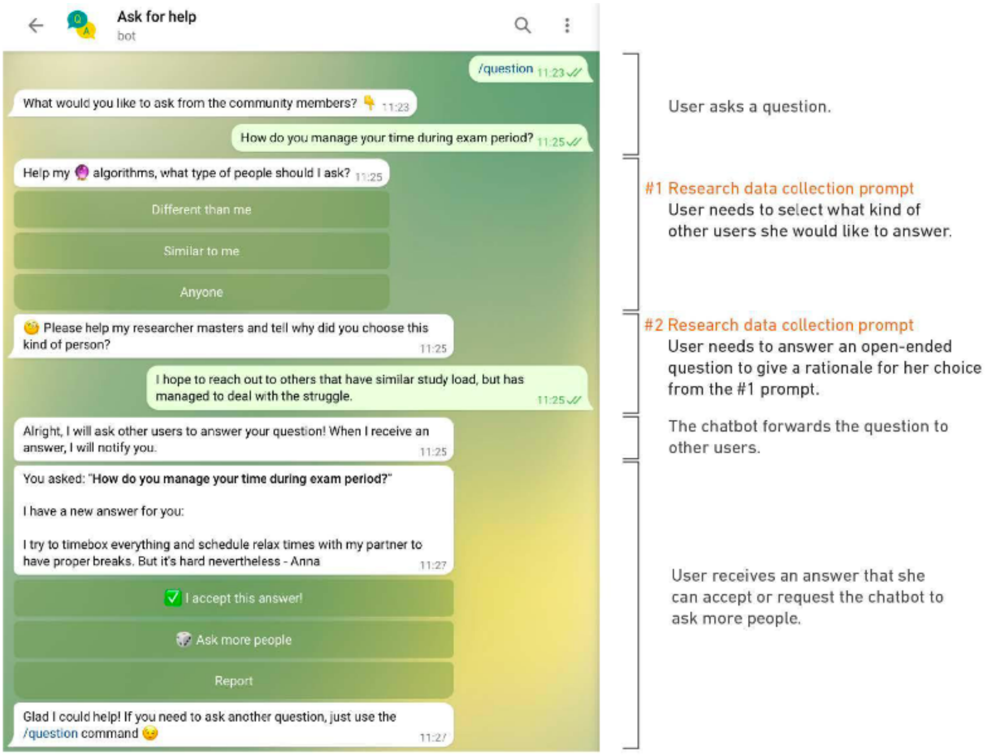
As anticipated in Chapter 1, Big Thick Data is understood as a stream of situational context concerning multiple occasions in a person's life that can be explored at different levels of granularity. Likewise, there are numerous possible study configurations, each of which can lead to varying levels of validity and reliability of the results. This study [1] verticalize the observation of social interaction within the *Personal context* through a chat application to enable the exchange of messages

[1] This chapter is an adaptation of [175].

between students of the same university but belonging to different formal and informal groups. This study is also part of the WeNet project (see Section 1.4), which aims to leverage the diversity of people mediated by technologies.

5.1 Topic and objectives

The WeNet Ask4Help - chat application (see Figure 5.1 aims to demonstrate and test different diversity-aware technologies being developed within the WeNet Project and to enable the communities at the diverse WeNet pilot locations to benefit from the diversity of their members mutually. In this context, a community question-answering (Q&A) chat application was designed, where the users interact with their peers through a chat application that connects them based on their diversity.



/question 11:23

What would you like to ask from the community members? 👇 11:23

How do you manage your time during exam period? 11:25

User asks a question.

Help my 🔮 algorithms, what type of people should I ask? 11:25

#1 Research data collection prompt
User needs to select what kind of other users she would like to answer.

😊 Please help my researcher masters and tell why did you choose this kind of person? 11:25

I hope to reach out to others that have similar study load, but has managed to deal with the struggle. 11:25

#2 Research data collection prompt
User needs to answer an open-ended question to give a rationale for her choice from the #1 prompt.

Alright, I will ask other users to answer your question! When I receive an answer, I will notify you. 11:25

The chatbot forwards the question to other users.

You asked: "How do you manage your time during exam period?"

I have a new answer for you:

I try to timebox everything and schedule relax times with my partner to have proper breaks. But it's hard nevertheless - Anna 11:27

Glad I could help! If you need to ask another question, just use the /question command 😉 11:27

User receives an answer that she can accept or request the chatbot to ask more people.

Figure 5.1: The Chat Application mediated the interaction between users and through research prompts (1 and 2 highlighted)

This study was conducted involving students from the following universities: Aalborg University (AAU) (DK), London School of Economics and Political Science (LSE) (UK), the National University of Mongolia (NUM) (MN), the Universidad Católica "Nuestra Señora de la Asunción" (UC) (PY), and the University of Trento (UNITN) (IT).

The experience with the chat application has been the basis for several studies. In particular [228] they focused on the aspect of incentives, particularly relevant for studies with Big Thick Data and intensive longitudinal surveys, where survey abandonment or failure to complete is a rather significant problem. According to [228], the prevalence of online participation in citizen science projects has grown, but a small group of highly engaged participants contributes to the majority of tasks. In contrast, most participants only perform a few tasks. Previous research has explored the motivations behind participant engagement, highlighting factors such as personal interest, the desire to learn, volunteering, and contributing to science. The study aims to increase the quantity and quality of contributions by understanding user interests, social norms, task importance, and user reputation.

Further information on design choices and data collection insights can be found in [59].

5.2 Measurements and tools

This study has developed indicators and tools that allow the measure of diversity across the three dimensions underlying social practices (competence, material, and meaning) and also to measure how they are organized and performed in interaction while, at the same time, maintaining a level of comparability across university students from different country and cultural communities. The process was articulated as a four-stage data collection, as follows:

1. The first stage is called Diversity 1, and it is described in Chapter 4. It is composed of three close-ended questionnaires and a smartphone app that collected observations on social practices and students' daily routines;
2. The second stage, called Ask4Help Pilot, is administered via two main channels:
    - A chat application that allows participants to exchange questions and answers;
    - A smartphone application that collected data from 9 smartphone sensors.
3. A close-ended questionnaire to evaluate the chat application features;

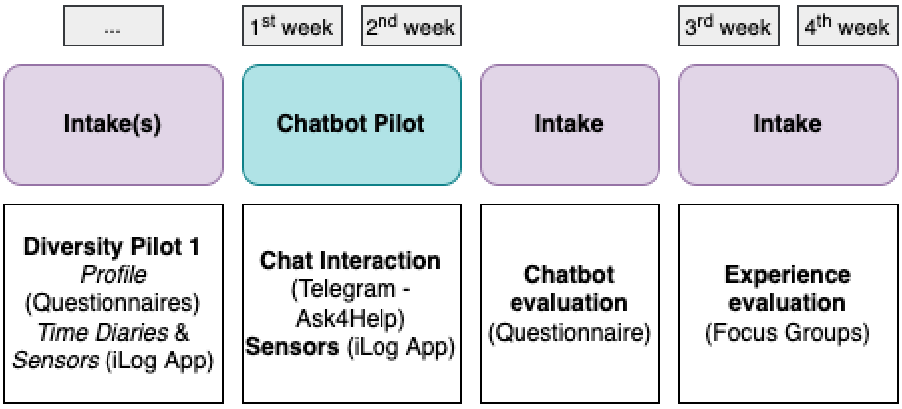


Figure 5.2: Chat Application study protocol

As described in Figure 5.2, the first stage (two months) was dedicated to Diversity Pilot 1 (see Chapter 4). From this, participants were selected and invited to download the chat application and iLog app to participate in the two weeks of data collection. An evaluation questionnaire was sent right at the end of the data collection. During all the data collection, a help desk was active and ready to support students in all the problems that were arising.

5.2.1 Data collection tools

The questionnaires were managed with the LimeSurvey [118] platform. An invitation to participate in the online survey was sent through LimeSurvey to the email addresses of students enrolled at the various universities.

The chat data were collected through a Telegram chat application called Ask4Help, designed by Service Design Lab and developed by U-Hopper Srl and the coordinating partner, the University of Trento and the Ben-Gurion University of the Negev. This chat application was assessed by the JRC's Market Creation Potential indicator framework as having a "Noteworthy" level of Market Creation Potential.[2]

The sensor data were collected via `iLog app` [124], which collected the information remaining active in the background without interfering with the use of the phone and chat application.

5.2.2 Sample design

The sample strategy was to involve 50 students from each pilot site, with the following requirements:

1. Having participated in the Diversity Pilot 1 questionnaires;
2. Have an Android smartphone;
3. Having installed iLog;
4. Having consented to be contacted for WeNet inquiries.

The selected students were contacted by email, and, in the case of LSE and UC, they participated in an online event in which the study and the chat application were presented.

For various pilot sites, the selection criteria were relaxed, involving people who had not participated in Diversity Pilot 1 (who were, in any case, administered a summary version of the three questionnaires) and people in possession of smartphones with iOS or other systems. The latter participants could not install the iLog app during the chat application pilot.

[2] https://www.innoradar.eu/innovation/37259

5.3 Incentives strategy

The strategy of involving the participants concerned tangible and intangible incentives. Regarding the material incentives, a fixed payment has been defined for all those who have installed the chat application. Table 5.1 reports the fees for each site. In addition, AAU has made seven prizes of 500kr available to be drawn by lot among the active participants.

Table 5.1: Incentives per pilot site

	Compensation
AAU	150 kr
LSE	10 £
NUM	20.000 MNT
UC	35.000 Gs
UNITN	10 €

In addition to monetary incentives, two different types of intangible incentives were provided. First, a set of badges with the following characteristics has been produced:

1. Quantity badges, or based on the number of contributions to:
    - Incentivise asking questions
    - Incentivise answering questions
2. Quality badges to reflect the participant's judgment on the contribution. The badges were given when the requester accepted the answer.
3. Reputation, to showcasing users achievements

The complete set of badges can be found in Table 5.2.

Secondly, a set of messages was sent to:

1. overcome "question posting anxiety";
2. incentives a better performance - asking or answering more questions.

The complete set of messages can be found in Table 5.3.

5.4 Results

The final dataset contains data about *only* 186 participants. Table 5.4 shows the selection and participation of students in the various stages of the survey. As can be seen, participants from the previous study were invited (only those who completed more than 75% of the survey, where possible). To address the inevitable abandonment of the survey, additional participants were invited at all sites except UNITN - which had a sufficient number of participants in the previous study.

Although over 50 people have agreed to participate in all pilot sites except LSE, many have not installed the app. The highest number of defections occurred in UC, with only 22 participants, while at LSE, 47 participants downloaded the chat application, the highest number in the survey. Regarding the final questionnaire, 39 participants per site completed it, except AAU (29) and UC (21).

The following paragraphs are aimed at describing the data from the three different sources, namely (i) questionnaires, (ii) the chat application, and (iii) sensors.

5.4.1 The questionnaires

Since the survey was conceived as a continuation of Diversity Pilot 1 [174], the observations of most of the participants can be deduced from the previous questionnaires. Tables 5.5 describes the main characteristics, from which it can be deduced that in all pilots, most of the participants were women, with a peak at LSE where they were 89.4%. Most participants were enrolled in a Bachelor's (at NUM and UC, they were the only participants), except at LSE, where 57.4% did the Master's. In addition, most were enrolled in a hard sciences degree program. In particular, at NUM where 71.4% were enrolled in one of these courses.

In addition to the previous questionnaires, an evaluation questionnaire was carried out on the chat application and its elements. The questionnaire contains quantitative assessments and textual comments regarding the following:

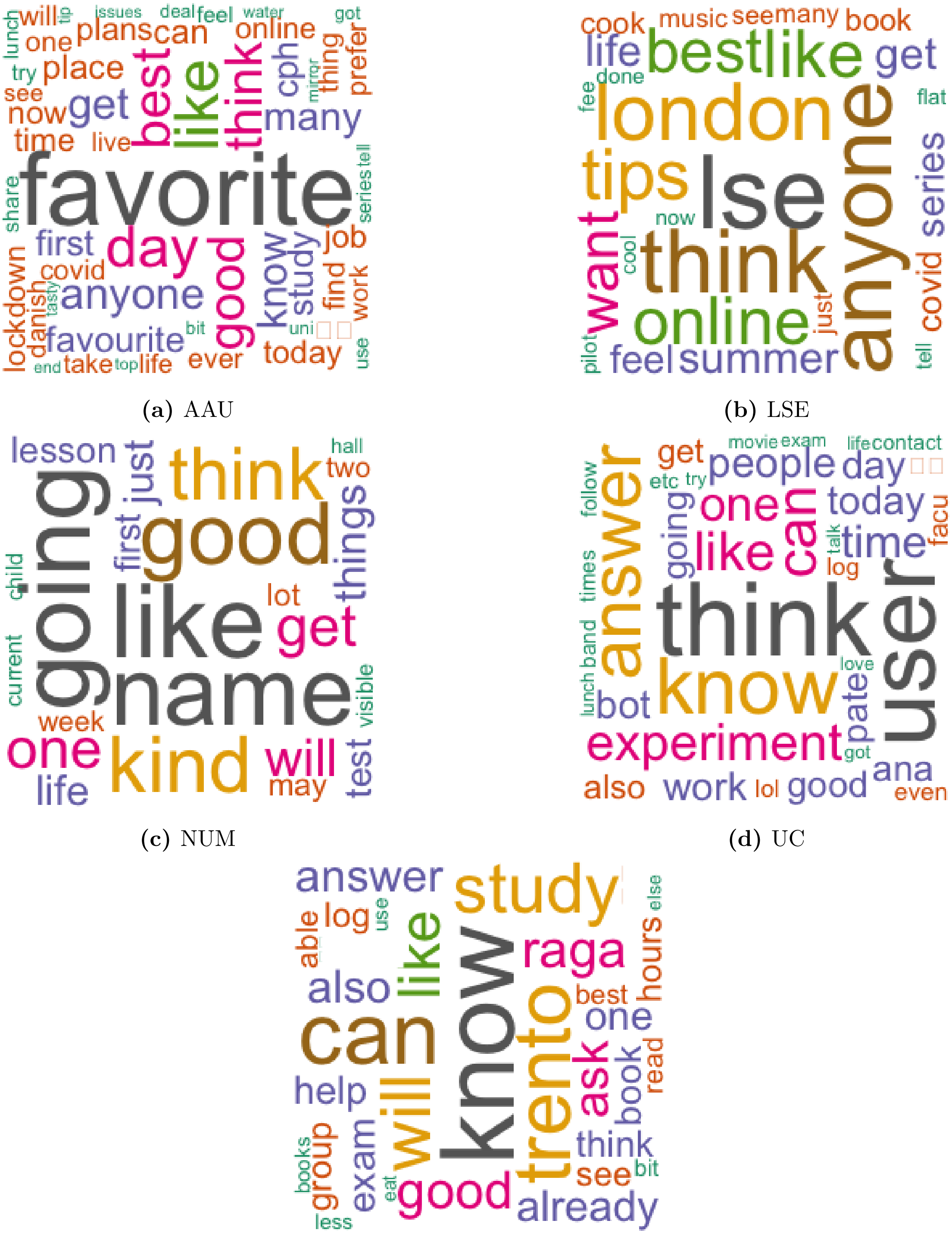
1. the user experience
2. experience with badges and messages used as incentives
3. the general experience with the chat application and its features.

This questionnaire can support the analyses and observations deriving from the study of the chat application data.

5.4.2 Chat application data

The chat application was based on a question-and-answer mechanism on any topic of interest to the participating student community. Tables 5.6 and 5.7 describe the average and total of the questions and answers asked at each site. From the tables, it can be seen that there were very active participants, particularly at NUM, where one participant asked up to 56 questions, and another gave 143 answers. For this reason, NUM was the site with the most significant interactions, i.e., 589 questions and 3389 answers.

Figures 5.3 and 5.4 present the most common words for each pilot used during the study. As for the questions, the common term was "favorite" in AAU, "lse" for LSE, "like", "think" and "know", respectively for NUM, UC and UNITN. As for the answers, the most common words were "like" for AAU, "going" for NUM, and "yes" for the other three pilots.



(a) AAU

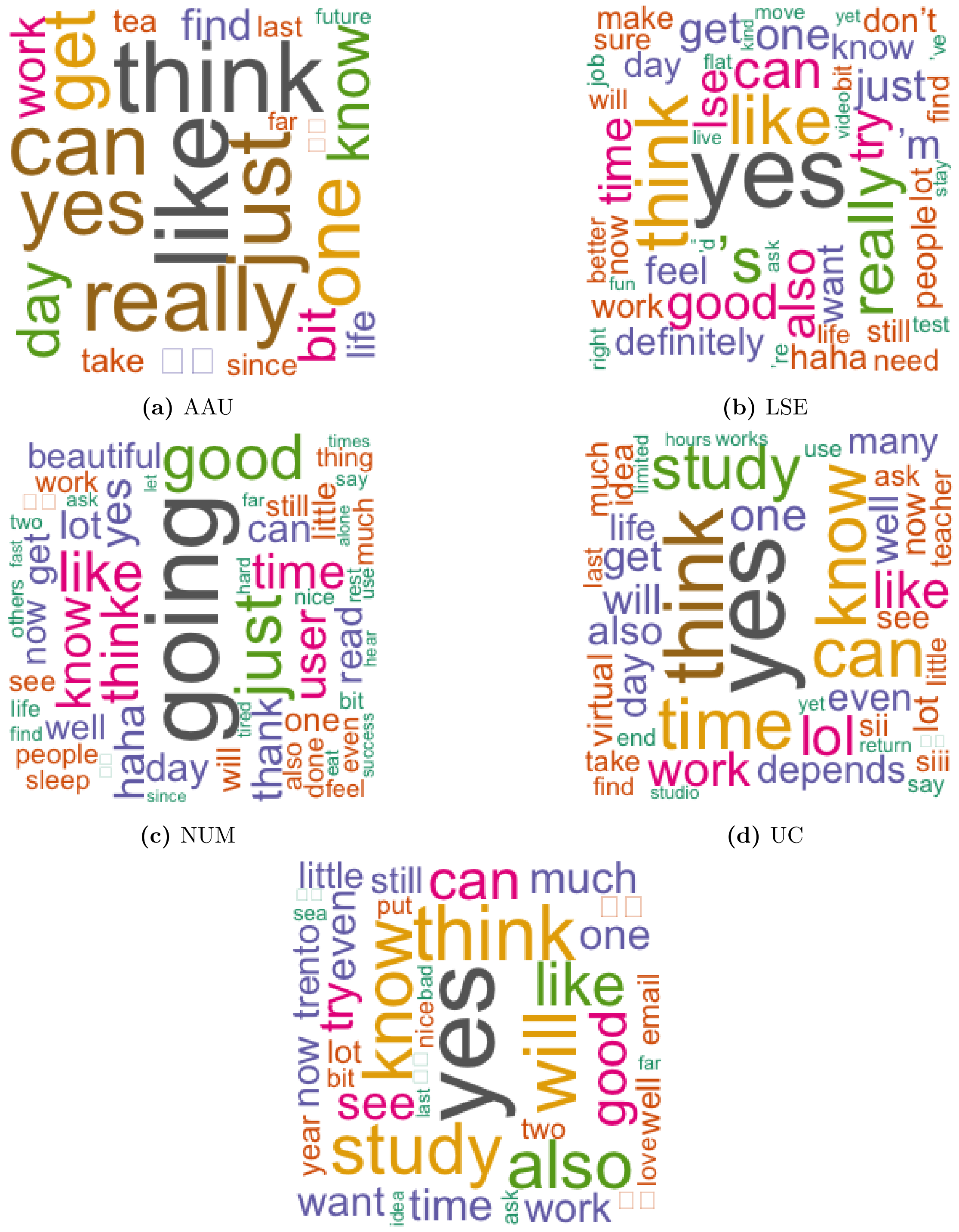
(b) LSE

(c) NUM

(d) UC

(e) UNITN

Figure 5.3: Most frequent words in questions per pilot site



(a) AAU

(b) LSE

(c) NUM

(d) UC

(e) UNITN

Figure 5.4: Most frequent words in answers per pilot site

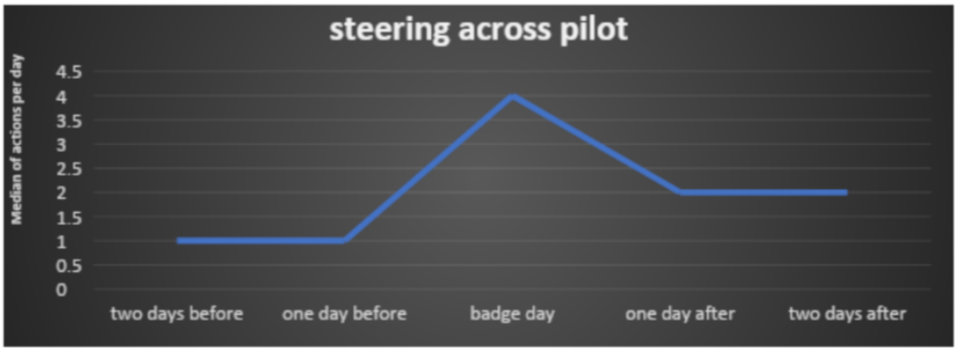


Figure 5.5: Overall badges steering effect

The chat application dataset also contains different parameters regarding the timing and the types of questions that have been asked. In addition, the participant could define who to ask the question and the reason for accepting one answer above the others.

5.4.3 Incentives data

Regarding incentives [228], as shown in Figure 5.5, there is an evident steering effect [3], as demonstrated by the changes in user behaviours. Users exhibit an uptick in activity leading up to the badge allocation day, followed by a decline in actions after the badges are granted. Furthermore, the post-granting decrease stabilizes at a new activity level, surpassing the level observed before badge allocation.

5.5 Roles

The roles of the authors, presented by their initials, are as follows:

- *Experiment management*: F.G., I.B., A.D.G.;
- *Experiment design*: F.G., I.B., A.D.G., M.B., R.C.A., G.G., M.B., D.G.P., L.M.;
- *Technical support*: M.R., M.Z., C.G., M.B., C.C.;
- *Data Collection*: M.B., R.C.A., M.R., A.D.G, P.K.,A.G., A.C., G.G., M.B., L.C., A.H.;
- *Data Preparation and correction*: R.C.A., C.G., M.B., C.C., P.K.

[3] The steering effect is an increase in the rate of participation associated with the badge achievement date [229]

5.6 Discussion

Below are some considerations regarding the validity and reliability of the results and methodological considerations.

5.6.1 Validity

The data collected have been the subject of several interdisciplinary publications demonstrating the validity of the constructs used. The exciting aspect of these studies is to note how they explore a component of Big, Thick Data that comes closest to a practical wild situation. Unlike the previous study, where the observation of the context was structured through questions spaced over time, the streaming of information in this study is (almost) spontaneous, leading to more qualitative representations of the context. Below is a description of the main articles based on the datasets collected.

Exploring diversity perceptions in a community through a Q&A chatbot [59] The paper focuses on the chat application's design to explore how a diversity-aware approach could help students' daily lives. In particular, the main focus was on creating a community of students based on exchanging messages, sharing interests, curiosity, or help requests. The findings show how students aggregated in sub-communities according to their interests after exploring other students' diversity.

Analyzing User Experience of the Chatbot SOS TUTORÍA UC [230] 'SOS TUTOR'IA UC' is a responsive web application facilitating academic assistance through external tutoring. The study focuses on the development and validation of the application, highlighting the incorporation of *Personal context* features in recommending students for tutoring. The goal is to help students connect with knowledgeable peers with different, similar, or indifferent personality traits for academic support.

A Diversity-Aware Application for Tutor Recommendation [231] The study focuses on developing and validating the 'SOS TUTOR'IA UC' application, emphasizing the incorporation of personality traits from the Big Five model in recommending students for academic tutoring. The recommendation system testing, again based on *Personal contexts* traits, indicated positive results with room for improvement, particularly in satisfying personality requirements and providing options to diversify recommendations by gender. Participants emphasized the importance of considering personality traits in tutor selection, expressing a need for additional information and more detailed profiles to enhance the matching process.

5.6.2 Reliability

Although the approach has favoured several scientific publications, demonstrating its intrinsic validity, both the duration of the study and the reduced participation of the students (see Table 5.4) have affected the reliability of the results, in particular when considering machine learning approaches for learning the social context of the participants. In addition, many participants were female, a common trend in surveys [232]. However, it is possible to integrate this dataset with the one collected subsequently (i.e., [176]), thus reaching more participants. Since the interactions cover different aspects of daily life, not only related to university but also to sport, interests in literature and cinema, and shared knowledge of frequented places, it is a remarkable case study for qualitative analyses.

5.6.3 Methodological considerations

Starting from the results, it is possible to draw some considerations on the research protocol, the design of the instruments, and the use of incentives.

The data collection took place in the context of the COVID-19 pandemic, which certainly affected the aspects of participant engagement, forcing the researchers to rely only on email communications. Another element that disadvantaged the sample size right from the start was using two separate applications, which made the registration procedure more confusing for participants.

From the point of view of the survey design, it has, as mentioned previously, the advantage of approaching the interaction methods typical of daily life, in line with the [15] approach. The interactions between participants were based on natural language and were not pre-coded. The extension of the survey, combined with more structured support technologies, could encourage greater communication as well as interest on the part of the participants.

The results showed that non-monetary incentives were exceptionally functional in soliciting active contributions. This effect could also be mediated by the presence of monetary incentives, which generally ensure more lasting participation in intensive studies. In addition, other surveys (e.g., with a Multipurpose design) must be adapted to avoid soliciting participants too much.

Name	Type	Description	Message
First question	Quantity: counting first question asked by user	After first question asked	Congratulations! You just earned the First Question badge! Way to go!
Curious level 1	Quantity: counting sum of questions asked by user	After 5 questions asked, user gets the badge	Congratulations! You are Level 1 Curious! Stay hungry and keep asking questions!
Curious level 2	Quantity: counting sum of questions asked by user	After 10 questions asked, user gets badge	Congratulations! You are now Level 2 Curious! Amazing interest in the world, keep asking questions!
First answer	Quantity: counting first answer given by user	After first answer given	Congratulations! You just earned the First Answer badge! Way to go!
Helper level 1	Quantity: counting sum of answers given by user	After 5 answers, user gets badge	Congratulations! You are now Level 1 Helper! Keep sharing your knowledge!
Helper level 2	Quantity: counting sum of answers given by user	After 10 answers given, user gets badge	Congratulations! You are now Level 2 Helper! Keep sharing!
First Good Answer	Quality: counting first accepted answer by user	After first accepted answer	Congratulations! You just earned your First Good Answer badge! Well done!
Good Answers level 1	Quality: counting first accepted answer by user	After 3 accepted answers	Congratulations, your answers are appreciated! You just earned your Good Answers Level 1 badge!
Good Answers level 2	Quality: counting sum of accepted answers by user	After 5 accepted answers	You keep giving great answers! You just got Good Answers Level 2! Congratulations!

Table 5.2: Intangible incentives: Badges

Table 5.3: Intangible incentives: Messages

Message	When?
You are x question/s away from a new badge! Type: question to ask the community!	User is messaged when there are no questions after a certain time frame, and x question is missing for leveling up.
You are x answer/s away from a new badge!	User is messaged when the user needs x answer to level up and has been inactive for a period.
You haven't asked a question yet. You can get help from the community with your questions. Type: question to ask the community!	User is messaged when they have not asked any questions yet.
You haven't asked a question recently. Anything you are wondering about that the community may know?	User is messaged when no questions have been asked after a certain time frame.
There are open questions to answer. Type: answer for the list! (by querying getTasks in the task manager)	User is messaged when there are open questions to answer and the user hasn't answered a question for a while.
Help the community by answering open questions or by asking new questions!	User is encouraged to contribute to the community
Any question is a good question! Type: question to ask the community.	User is encouraged to overcome "question posting anxiety"

Table 5.4: Sample selection and participation

	Completed D1	Invited from D1	Additional invitation	Agree to participate	Installed Ask4Help	Completed the exit survey
AAU	24	all	up to 2628	55	34	29
LSE	76	all	up to 500	47	47	39
NUM	213	>75% (N=76)	up to 350	68	39	39
UC	28	all	up to 500	58	22	21
UNITN	238	>75% (N=110)	.	53	46	39

Table 5.5: Sample descriptive statistics

	AAU	LSE	NUM	UC	UNITN
Gender					
Male	22.2	10.6	25.6	47.6	31.8
Female	77.1	89.4	74.4	52.4	68.2
Degree					
BSc		36.2	100.0	100.0	50.0
MSc		57.4	.	.	38.6
Other		6.4	.	.	.
Department					
Hard Sciences			56.4	71.4	38.6
Social Sciences			25.6	9.5	36.4
Humanities			17.9	14.3	25.0
Total	100 (N=35)	100 (N=47)	100 (N=39)	100 (N=21)	100 (N=44)

Table 5.6: Number of questions per user and total number of questions

	AAU	LSE	NUM	UC	UNITN
mean	11.5	6.1	17.8	11.5	3.8
median	11	4	17	11	2
sd	7.7	8.0	13.5	9.2	4.4
max	27	40	56	34	20
min	1	1	1	1	1
Total	402	257	589	230	379

Table 5.7: Number of answers per user and total number of answers

	AAU	LSE	NUM	UC	UNITN
mean	43.1	16.9	86.7	31.2	19.4
median	32	13	88	31	15
sd	38.9	17.8	77.6	21.9	19.6
max	145	82	293	90	72
min	1	1	1	2	1
Total	1638	762	3384	655	1218

6

Case Study III: Applying the iLog methodology with external researchers

Contents

Ultimately, the iLog methodology described in Chapter 3 has the aim of enabling the production and reproduction of context-based studies, intended as a quantitative implementation of the notion of Big Thick Data (see Chapter 1). In this sense, the methodology's validation can only occur through its autonomous reuse by researchers with different backgrounds. This Chapter presents two data collections conducted by researchers selected among the applicants to the WeNet Open Calls (see Section 1.4, i.e., a set of challenges aimed at testing the technologies and methodologies developed within the European project.

While in the data collection described in the previous Chapter 4, the writer had an active role in all phases of the process, in the case of the Open Calls, the researchers were invited to autonomously apply both the methodology and the services developed, with minimal support. In particular, the researchers of the Open Calls had access to the materials described in Appendix A, B (except the documentation regarding the data distribution) and to the codebook and technical report template (Appendix D) and were invited to independently conduct the data collection, following the iLog methodology phases.

The writer supported the researchers in the operational management of the different phases. The support was limited to providing advice on the design of the study and questionnaires and in the preparation of recruitment and monitoring. Furthermore, the writer managed the data preparation and anonymization aspects, supported by the KnowDive research group of the University of Trento.

Below is the full version of the reports written by the two participants in the Open Calls [1], namely the researchers from the FPT University (Vietnam) with a background in Computer Science (Section 6.1) and the researchers from the University of Thessaly (Greece), with a background in Economics and Behavioral Studies (Section 6.2).

6.1 VietF&D: Collecting Hand-To-Mouth Activities from Surveys and Mobile Sensing Data

Young college adults take full responsibility for their daily eating, drinking, and lifestyle practices, especially when transitioning into student life. In this critical period, it is challenged for students to have healthy food choices which are potentially influenced by personal reasons (e.g., self-discipline), social communities (e.g., eating habits from their family), physical environment (e.g., availability and accessibility of food in campuses), and macro environment (e.g., advertisement). In developing countries like Vietnam, many ordinary people, including students who move to big modern cities to study at universities in rural areas, do not consider adopting healthy eating habits. Moreover, the possibility of living away from family will give students more freedom to hang out and possibly drink alcohol with friends. In Vietnam, there are not many strict regulations on buying alcohol, e.g., being able to buy an alcohol-related drink at the supermarket after 7 PM, which gives students more chances to be drunk, causing harmful or uncontrolled behaviors of students. To our knowledge, no study about eating and drinking activities has been conducted in Vietnam using ubicomp data with quantitative methods. Researchers used multiple traditional methods, e.g., surveys and paper-and-questionnaires, but they had many limitations. Due to these reasons, using a smartphone to collect ubicomp data and food and drink log data is an essential preliminary step for

[1] The reports have been revised and corrected to make them more intelligible for the reader, taking care to keep the structure and contents unchanged.

researchers to conduct a study about the hand-to-mouth behavior of students and extract higher implications about nutrition, behavioral science, and psychology.

Many commercial apps, e.g., MyFitnessPal or Apple Health, enable users to log their food intake as a diary and provide possible statistical reports after a period. However, these apps do not collect mobile sensor data and extract possible patterns connected to food and drink consumption. In the scope of WeNet, iLog was used to collect smartphone sensors and diary logs or participants' surveys.

Compared to the previous project using other applications, our proposal will collect mobile sensing data (e.g., accelerometer, locations, etc.), especially logs of food and drink by young students. The project's novelty is the collection of diverse data in different places to better understand the self-perception of food and beverages among young adults across regions in Vietnam (i.e., Hanoi, Da Nang, Ho Chi Minh City). In addition, the data we collected from students in a developing country such as Vietnam could be an attractive value that complements the current diversity-awareness datasets in WeNet.

The primary goals of this study are to identify the diversity of healthy food behavior and alcohol drinking among young adults, along with passive sensing through mobile phone sensors. We then leverage collected data (i.e., food and drink surveys with eating and drinking places, social context, etc.) to get insights into food and drink consumption from young people in multiple geographical locations in Vietnam. Indeed, passive sensing data is used by previous works for various purposes of inferences, e.g., stress, mood, activities, sociability, food types, and heavy drinking. Most prior work used data collected from Europe and America. In contrast, this project will collect food and drink diaries and ubicomp sensing data of young people in a developing country like Vietnam.

Interestingly, Vietnam is a tropical country with different food categories, various food and drink consumption styles, and daily life compared to other countries in Europe, America, and Asia. Hence, it promises to get new patterns by using sensing data and activities with food and drink logs that potentially contribute to the diversity purpose of this project to WeNet. We hypothesize that mobile sensing features and eating and drinking patterns could be used to make students think about their hand-to-mouth behavior after joining our data collection campaign.

6.1.1 Project execution

The data collection was designed to replicate the methodology developed in the WeNet Project and validate the replicability and usability of the services proposed by the project. In particular, the study design was based on the methodology and investigation protocol of the Diversity 1 project (see Chapter 4), deepening the aspects related to the consumption of food and drinks. The data collection was based on the iLog app, developed by the KnowDive group at the Department of

Information Engineering and Computer Science of the University of Trento. It is widely used in the context of WeNet.

The app was configured to (a) collect all the data from the smartphone's sensors and (b) send notifications at different times of the day regarding food and drink consumption. In particular:

Smartphone sensor To facilitate data reuse in different research fields, all data from the smartphone sensors were collected. The sensors of particular interest for this investigation were app usage (indicating app usage behavior), accelerometer (implying daily activity levels of users), battery events (showing general phone usage of users), screen events (indicating the activeness of phone being used), and location (indicating latitude/longitude position of participants). This sensor information will be collected throughout the day.

Food and drink notifications Regarding the food log, we asked participants to fill out a food survey with food categories (e.g., cereal, fish, vegetable, etc.), social context (e.g., eating with friends, family, etc.), eating location (e.g., home, restaurant, etc.), activities during eating (e.g., watching TV, playing a game, etc.), participants' mood and stress level at eating time, food consumption amount (1-5 Likert scale), and especially rate of healthy level (1-5 Likert scale). To get a food log, iLog sent notifications three times a day in eating hour frames that are particularly common in Vietnam, i.e., 6:00-9:00 for breakfast, 11:00-13:00 for lunch, and 6:00-9:00 for dinner. Regarding collecting drinking activities, the log focused on drink types (e.g., beer/cider, wine/champagne, liquor, etc.), drink containers (glass, can, shot, etc.), drink size (small, medium, significant), surrounding environment with people (friends, spouse, family, etc.), drinking places (bar, restaurant, private places, etc.), and drinking motives.

During data collection, active monitoring was conducted, sending notifications and reminder emails to participants to reduce non-response and dropout rates.

6.1.1.1 Sample design and recruitment strategy

The data collection targeted 18-25-year-old students enrolled in Vietnam's various FPT study faculties. An email was sent to 10K+ students at FPT University and students from other universities to invite potential participants. The email introduced the study's goal, methods, privacy operations, and incentives. In detail, we explain what data we collect, how their information will be anonymized, how they get feedback at the end of the data collection campaign, and how their data is used for future studies. The email invited interested participants to contact the organizers of the study. Once the responses had been received, the organizers sent a short survey, requesting phone numbers, email addresses, ages, and genders and inviting the students to download the iLog app to join the study. Since email response rates were meager, two additional steps were followed to recruit more participants. Therefore, the recruitment campaign was held as follows:

1. Step 1: June 06 –12. Academic Affairs of FPT University emailed 10K+ students in Ho Chi Minh City, detailing the project and the steps to joining the data collection campaign. In addition, a flyer has been posted at the library and entrance lobby of the FPT University campus in Ho Chi Minh City. The flyer and a project description were also posted on the Facebook pages of FPT student clubs in Hanoi, Da Nang, and Can Tho.

2. Step 2: June 13-17. To broaden the diffusion of the invitation, lectures from various faculties were contacted. In particular, two lecturers at the University of Science (Vietnam National University - Ho Chi Minh City), one lecturer at Nha Trang University, one at Van Lang University, and additional lecturers at FPT University in HCMC were invited to spread the invitation to their students.

3. Step 3: June 20-24. The 94 participants who had already registered for the study were encouraged to invite their colleagues to join the study.

6.1.1.2 Incentives strategy

The Incentive strategy consisted of withdrawing a prize of one iPhone and six Apple Watches for all participants who answered at least 75% of the questions. Participants who invited at least two other colleagues had higher chances of receiving the prizes. The prize's goal was (i) to keep the participants' attention high and (ii) to encourage communication with the helpdesk.

6.1.2 Ethics and data protection

The project received authorization from local institutional authorities and advice from the FPT University's Research Ethics Committee and the Vietnamese government. In addition, it followed and implemented the WeNet legal documentation, considering the Data Protection Impact Assessment, Privacy statement, and Participant Consent forms, a Data Processing agreement among the WeNet Consortium, and a Declaration of Commitment agreement (the templates regarding data collection are available in the Appendix B). The project adapted to the EU's 2016 General Data Protection Regulation GDPR and the WeNet ethical and privacy guidelines.

Personal data and anonymization All the information collected relating to the identifier (e.g., name, identification number, home address, etc.), device address (e.g., internet protocol addresses, MAC addresses, etc.), and demography information (e.g., ethnic origin, religion, sexual orientation, etc.) were anonymized by technical support colleagues at the University of Trento according to the iLog data preparation procedures (see Section 3.4.

Transparency The email and privacy documentation shared with participants followed the WeNet templates, clearly explaining how and which data were collected and stored. In particular, the documentation explained how collected data will be held on the WeNet server for future research and under strict data protection under the WeNet Consortium. Potential risks were also addressed. Consent to participate was acquired by checking the acceptance after showing the project details in the privacy statement. During the data collection period, participants could retrieve their consent and request the deletion of their data at any time.

6.1.3 Results and achievements

The final dataset contains about 117 participants, with 99 active participants who sent at least one response per day for two weeks. The sampling strategy was balanced according to the main socio-demographic characteristics, namely gender, age, and departments in which the students were enrolled. However, due to typical problems in the recruitment phases (e.g., participation and dropout rates, lack of responses, selection bias), our participants are primarily males, 18-25 years old, and from the computer science and business administration department. The following paragraphs describe the data from the three different sources.

The questionnaire Every user joins data collection in 14 days, which we consider a data point. We received questions about messages, tasks, and time diaries for 2106 user days. The total number of time diary surveys is 1985, in which we received 1404 questions and 581 answers.

Sensors We collected sensor data, e.g., accelerometer, etc. Every user contributes data over multiple days. We also consider user-day to be a data point. We extracted accelerometer data and grouped them into eight activities. We have N=979 user-day and normalized the results. As a result, "Still" takes the highest distribution at 63% while "Tilting" and "In-Vehicle" take 11% and 5%, respectively. There, 18% of the data is "Unknown".

Application usage Regarding applications used during our data campaign, we found 734 applications used in 981 users. The top 4 daily applications are Facebook, YouTube, Chrome, and Zalo. Zalo is a local Vietnamese application that allows users to send messages, like WhatsApp.

Bluetooth We also extracted the Bluetooth data from 244 user days and implied the possible devices connected to participants' mobile phones. We found that 90% of Bluetooth connections are uncategorized, while 6% are related to the speaker and video display devices.

Location We collected 1.9M data points from 1764 users with the three providers, i.e., network, GPS, and passive location.

6.1.4 Feedback and afterthoughts

In this project, we used iLog, an application from the set of enablers of WeNet. Thanks to Matteo, we composed the questions and answers in English and Vietnamese. This setting makes it convenient to tune the application interface into the local language, i.e., Vietnamese. However, our idea was not implemented well under the current design of iLog. For example, the questions' answers will lead to the available states of some of the following questions. We cannot implement this with the current iLog application. We suggest that iLog should be more flexible and adaptable to many scenarios.

We re-designed the list of questions, which avoided the above problems. Matteo also helped us during the project, and we found the solution to these problems.

The WeNet Advisory Board has added reflection regarding ethical and privacy aspects. The use of information related to alcohol and food consumption, although already present in previous WeNet pilots, was particularly in-depth during this pilot. This could create problems in the reuse of data, as in addition to the risk of deanonymisation, there is the risk of sharing potentially sensitive data. Therefore, despite the approval obtained from FPT University, cautious use of this dataset and limited sharing is recommended.

6.2 A study about the diversity of students in social contribution activities

The project aimed to explore the specifics of the University of Thessaly (UTH) students involved in civic engagement activities. In particular, it aimed to shed light on their characteristics, values, behavior, and whether and how these influence participation in civic activities that affect the local or their broader community. The data that we collected concerned the profile of the students, their daily behavior, the activities in which they are involved, and civic matters.

Using a digital platform and the iLog application, we collected such data from students of different levels (undergraduates, postgraduates, doctoral students) studying at various departments of the UTH. All UTH students were invited (giving them incentives) to participate in our survey. Still, finally, those who remained committed and provided complete and valid data to the research were 100 students. Our sample was unbiased as it came from different academic disciplines and UTH departments in other cities. It was concerned with students of all ages and levels of study, with almost equal participation of men and women.

The research was organized in three stages:

1. 1st stage: Recruitment of participants and online survey using the Google Forms platform with questions about the data we want to collect, such as their profile and behavior, including participation in civic engagement activities.

2. 2nd stage: Use the iLog application to collect daily behavior data such as movement, location, and time diaries and create and send queries to users about civic engagement behavior (i.e., the subject of the study). The frequency of questions was every half an hour for the first two weeks and every hour for the next two weeks.

3. 3rd stage: Additional survey data were collected through Google Forms to detect deviations in the participants' behavior during the initial stage of the research and questions aimed at collecting students' experience from their participation in the study.

6.2.1 Project execution

During the implementation of the research, there was a deviation in the execution time of each stage due to unexpected issues that emerged. First, some finance and management issues with the UTH research committee needed to be solved before the research could start. Second, the approval process from the UTH Ethics Committee took longer than expected; due to the nature of the study and the sensitivity of the data about to be collected, the Committee required the project to provide further clarifications on several matters and more detailed documentation. The survey approval process and signing of the DPA documents were prerequisites for both sides to start data collection using questionnaires and the iLog application. In addition, additional time was needed to convince students to participate and to recruit the participants for the project. Finally, there was a delay in the finalization of the iLog application (due to technical reasons) that delayed the data collection and the 2nd stage. For the above reasons, the project could not conclude as scheduled by the end of May, and it was finally completed at the end of July. The extension to conduct the research was crucial to its thorough and successful completion.

The overall data collection process lasted eight weeks. The process was articulated as a two-stage data collection, as follows:

1. The first data collection stage, administered through two questionnaires to collect general data about the participants and their civic engagement

2. The second data collection stage, administered using the iLog application, allowed us to observe the student's daily routines and behavior.

As described in Figure 6.1, the first two weeks were dedicated to recruiting the participants. This involved sending our first questionnaire, which collected primary data concerning the students' profiles and engagement in civic matters. The 3rd week was dedicated to the main questionnaire of the research to the selected sample (collecting detailed data with the profile of the students, their values, their beliefs, their behavior, and the activities they are involved in general and about civic matters). The following month was entirely dedicated to the data collection through the iLog application installed on the students' smartphones. In the first two weeks, the frequency of the time diaries was every 30 min., and in the last two weeks was

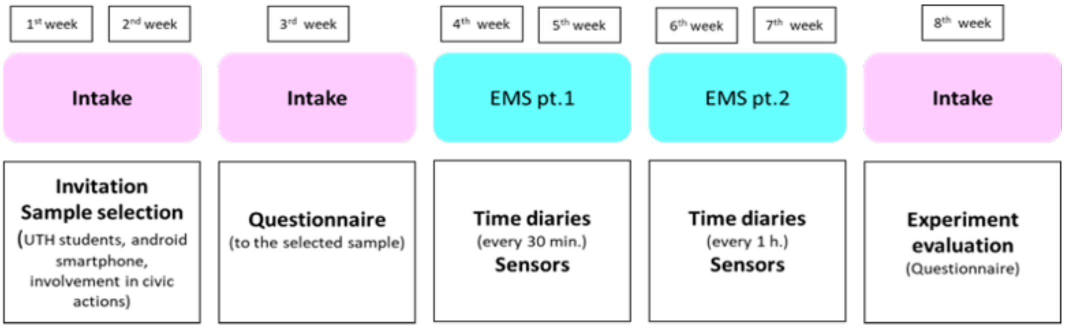


Figure 6.1: Data collection process

every 1 hour. Finally, the last week was dedicated to the final questionnaire to evaluate the research and the participants' experience. The sample was selected from the entire student population of the University of Thessaly. All students were sent an invitation to participate in the survey. After the first contact via email, we used a questionnaire expressing interest in participating in the research with which we could select students who meet specific criteria (they study in departments of the University of Thessaly and are involved in civic engagement activities).

We followed an incentive strategy to reach and engage the participants in the research. As the UTH Research Committee did not allow us to offer an amount of money as compensation for the participation, we decided to provide (by draw at the end of the project) many gift vouchers from a well-known electronics and stationery store. In particular, we decided to give the participants a total of 3000€, divided into 50 gift vouchers 50€ each and one laptop worth 500€. The participants who chose to participate in the research for the first two weeks had one entry into the draw, and those who decided to stay in the study for the additional two weeks had two entries into the draw. The prizes were aimed to keep the participants' attention high and encourage them to use the iLog application daily.

6.2.2 Ethics and privacy

Based on the UTH Ethics Committee requirements, the procedures were followed to secure the personal and sensitive data collected. The UTH is responsible for processing personal data subject to processing through the research conducted in its departments. No personal data collected for this purpose can be transmitted outside the EEA. If, in the future, it is required, within the responsibilities of the University and within the framework of its legal obligations or claims, to make a transfer of data to a third country or an international organization, this will be carried out under the conditions of legal and secure transfer provided by Regulation 679/2016 and national legislation. The UTH pledges to take appropriate organizational and technical measures for data safety, security, and protection from accidental or improper processing. Its specially authorized personnel, which processes personal data, has received the appropriate training and guidance, while the measures taken are reviewed and amended at regular intervals. Throughout the project, we had the guidance of the University's relevant personnel to ensure the security of the participant's personal data. At the beginning of the project, the person in charge of protecting and processing the personal data of the UTH was appointed. The research participants were fully informed about the purpose of the study, how to secure their personal data, and how they will be used by signing the respective form and giving consent for data to be extracted from their mobile phones through the iLog application. At the same time, participants were informed, and all relevant rights were guaranteed to be respected. These concerned the Right to access personal data, the Right to correct inaccurate personal data, the Right to delete and the Right to be forgotten, the Right to data portability, the Right to restrict processing, the Right to object to processing of their data, that someone: a) considers that a request has not been sufficiently and legally granted, or b) considers that the right to the protection of their personal data was infringed by any processing carried out. In the cases above, the person concerned has the right to file a complaint with the Personal Data Protection Authority.

The University of Thessaly's Ethics Committee accepted the documents the participants signed. The research involved the participants' personal data, so specific ethical and ethical concerns govern the project. These concerns involved the management of the participant's personal data and their complete anonymization, as well as the observance of confidentiality by the research team members. For this reason, all the necessary precautions were taken in the research to protect personal data following the principles of the UTH Ethics Committee. All data collected were immediately anonymized, and the data will be kept for as long as the research requires with all the necessary security procedures by the research team members. The personal data collected will be used exclusively for this research. With the sole exception of those cases where data retention is required by law, Personal Data will be deleted or at least anonymized by controllers and/or data processors, wherever they are stored, as soon as the Personal Data is no longer necessary for the specific purposes of the Project.

6.2.3 Results and achievements

The sample was selected from the entire student population of the University of Thessaly. All students were sent an invitation to participate in the survey. After the first contact via email, we used a questionnaire expressing interest in participating in the research with which we were able to select students who met specific criteria (they study in departments of the University of Thessaly and are involved in civic engagement activities). Initially, 310 people expressed interest in participating in the research, of which very few were not eligible. Due to the limitation of the application being available only on Android devices, several eligible participants were eventually unable to continue. At the same time, some withdrew from the research due to the kind and volume of daily that were about to be provided. Of those remaining, 163 people installed the application, of which 141 connected and used the application. Finally, only that stayed active and provided valid data were only 77. This smaller sample emerged after the clearance of the data and concerned participants who provided consistent and accurate data in all stages of the research. Table 6.1 below shows the characteristics of the sample, such as gender, age, and department of studies in which the students were enrolled, and table 6.2 shows personality traits (BFI-20).

	%
Gender	
Female	60%
Male	40%
Age	
18-22	51%
22-50	49%
Departments	
Engineering	16%
Humanities and Social Sciences	14%
Health Sciences	21%
Economics and Business Administration	20%
Agricultural Sciences	8%
Technology and School of Sciences	21%
Total	77

Table 6.1: Descriptive statistics of the participants

	mean	**sd**	**range**
Agreeableness	4.1	0.23	1-5
Conscientiousness	3.4	0.50	1-5
Extraversion	2.8	0.31	1-5
Neuroticism	2.9	0.32	1-5
Openness	3.9	0.13	1-5

Table 6.2: BFI-20 items distribution

6.2.4 Feedback and afterthoughts

For the implementation of this research, the iLog application (app), developed by the University of Trento team for research purposes, was used. Using this app, we collected data from survey participants about their daily behaviour, including engagement in civic activities.

For the application to be used by the students of the University of Thessaly, we had to upload the questions into the application and translate the questions and the entire menu into Greek. This was a time-consuming process, but with the guidance and help of the WeNet team and Matteo Busso in particular, the task was conducted appropriately and successfully. We singled out the application's advantages: its easy installation, ease of use due to its simple and understandable menu of options and settings, and low battery consumption.

However, some problems have been reported that constitute disadvantages of the app. For example, some users mentioned that they encountered problems installing the app, while others highlighted that the app's interface was not user-friendly enough. Some others reported that the app was crashing on several occasions, either stopping to function for some time or shutting down by itself. As a result, some participants dropped out, and others claimed that, in some cases, they did not receive the scheduled questions or received them with a delay of some hours. Fortunately, most reported app malfunction issues were resolved promptly with the guidance of the WeNet team.

Yet, it is essential to draw attention to a substantial drawback of the app. The app is available only to Android phones and not to users with iOS (Apple) devices and iPhones, something that cuts out a significant user base. In our case, a considerable number of students, although willing, were not able to participate in the research due to this issue.

Based on the above, we strongly recommend that the app's development team provide an iOS version of the iLog as soon as possible. We believe that this will increase the app's validity and boost participation greatly. In addition, another issue to be examined by the development team could be providing an upgrade of the app with a more user-friendly interface.

Concerning the whole research process, a significant issue that we encountered that considerably delayed the project was the ethics and personal data protection rules and the commitments undertaken in this regard. The process was lengthy, and it took quite some time to prepare the required documents and get approvals from all relevant parties. All these had to be done before the initiation of the actual research. This led to a considerable time loss, which was critical for the smooth conduct of the project. Finally, we would like to point out that this kind of research (where online surveys are involved and detailed daily data are sought to be collected) has quite a low response rate, especially those that collect data many times through the day (in our case, every 30 minutes).

Participants need incentives to engage in research projects like this, which require a substantial amount of their time daily. For this reason, it is paramount for the research team to develop and implement a strong motivation plan to keep the participants' interest alive.

In conclusion, we should highlight the importance of flexibility in managing projects like this. On this basis, we express our gratitude to our partners because they have approved the necessary amendments that provided us with additional time to properly prepare the documents required for the approval process from the ethics committee and the iLog application. The WeNet team's prompt response helped us resolve any issues that arose and conclude our research successfully.

6.3 Discussion

Both results are an essential contribution to the methodology set out in this thesis.

Firstly, the content of the reports shows how the researchers, despite coming from different backgrounds, could follow the various phases of the methodology. The researchers implemented the data collection design independently, effectively considering the management and planning aspects (all phases were executed within six months). Furthermore, the researchers followed the various ethical and GDPR protocols, obtaining the necessary approvals and implementing the documents for communication with the participants, and conducted the recruitment and monitoring phases of the study, obtaining good quality data and managing them for the publication of the report.

Secondly, the results show how the methodology is effectively replicable in different contexts. In both cases, more than 100 registered participants were in line with the participation rates of similar data collections. While dropout rates remain high, they have not differed significantly from WeNet's Diversity 1 pilot study. This feature will have to be addressed in future executions of the experiments (see Chapter 8 for further reflections on this aspect). Regarding the validity of the results, unfortunately, neither data collection has yet been published at the time of writing this thesis.

Finally, the proposed feedback is essential support for future implementations of the methodology and services, highlighting management aspects, such as the problems encountered in the approval by the ethics committee, which extended the timing of the study, but also factors related to services, such as the need to implement the configurability of the iLog app and its user interface. Further insights into this aspect will be provided in the last chapter of this thesis (see Chapter 8).

7

Case Study IV: Validation and ideation of iLog methodology services

Contents

This study concerns the results of the workshop for designing services connected to the iLog methodology. In theory, as an ideation workshop, it does not constitute a proper validation of the methodology. In practice, the workshop was held at the end of the WeNet Project (see Section 1.4), involving, among others, project members who had the opportunity to use the services and methodology in the project's data collections and contribute to their development. Furthermore, the workshop included an introduction to the services to facilitate discussion towards their implementation or the creation of new ones. For these reasons, participants had the opportunity to evaluate the current iLog methodology approach, thus not only proposing new services but also evaluating and identifying strengths and possible implementations of existing ones. In this sense, the workshop is presented here as a means for the validation of the iLog methodology.

7.1 Topic and objectives

Digitization [2][1] is exponentially increasing the production of data and driving relevant economic, social, and political changes [233]. The trend is often associated with datafication [234], namely the act of quantifying and computerising people's everyday life and that, according to [235], has the potential to shape both the ontologies and the methodologies of many disciplines.

However, as [5] would say, "bigger data are not always better data". Indeed, much useful data to model people's everyday lives is often missing [4], which leads to the problem of reducing people to their average, ignoring and disincentivizing their diversity, namely how similar or different a person's experience, competencies or traits with regards others [236].

Secondly, a growing body of literature is focusing on bias and bias management (see, e.g., the extensive work done by [85–87]) in scientific fields such as AI, health, and behavioural studies.

Finally, data management is a complex and multidisciplinary process, ranging from ethical and legal to social sciences and AI. Furthermore, it involves various phases, from collection to preparation up to distribution and reuse (see, e.g., [42]).

Although many strategies to mitigate the risks associated with non-diversity-aware approaches to data have already been proposed, such as explainability [83] or the report on Ethics guidelines for trustworthy AI [84], we believe that, following [5] suggestion, a broader (and radical) approach should be taken.

This is the reason why we suggest the development of an end-to-end research infrastructure (RI)[2] that enables trustworthy diversity-aware data. Furthermore, data is used in numerous fields, both for research and innovation. Being RIs pivotal for developing research areas as they consolidate both technologies and methodologies (considering, for example, standards or guidelines), we believe that an end-to-end RI is necessary, to support the researchers or developers within the whole data management process.

Therefore, the current workshop primarily aims to create new services to facilitate data management and data collection aspects to enable effective Hybrid Human-Artificial Intelligence approaches, following the WeNet approach (see Section 3).

The remainder of this chapter is organized as follows: Section 7.2 describes the methodology of the workshop, its organization, and the roles of the organizers. Section 7.3 presents the results divided for each workshop panel. Finally, Section 7.4 discusses the results and conclusions.

[1] This section is an adaptation of [82].

[2] According to [237], RIs are "facilities that provide resources and services for the research communities to conduct research and foster innovation in their fields".

7.2 Methodology and workshop organization

To guide the ideation process, the workshop was based on validating the services created in the context of WeNet, particularly those linked to the iLog methodology, considering the three main aspects: data collection, data preparation and analysis, and data distribution.

Considering data collection, different services were considered, such as (i) iLog App to collect sensor data and people feedback; (ii) Configuration service to configure the main aspects of iLog; (iii) Project webpage to advertise the study; and aspects of (iv) Participants engagement to invite students to download the app and participate in the study.

Concerning data preparation, it was considered (i) Transformation to convert Cassandra NoSQL collected data into a format suitable for storage (e.g., CSV, parquet); (ii) Anonymization for removing personal information from the dataset; (iii) Cleaning and formatting time stamp and variable label; and, (iv) Documentation, such as technical reports, additional material, and codebook templates.

Finally, as regards data distribution, the LivePeople Catalog was presented and its main features, i.e. (i) Search, which allows looking for datasets based on their salient characteristics, such as the type of data, the location and the year of data collection; (ii) Publish which allows uploading information about previously collected data, describing them based on a set of valuable metadata for search; and, (iii) Download which allows making a (GDPR compliant) request for a subset of data.

The workshop was organized around three discussion panels, one for each service (collection, preparation, and distribution), each organized into 4 phases interspersed with a moment of presentation of the results and joint discussion. Therefore, after a brief introduction in which the entire data collection and management process was presented, the participants were divided into three groups and faced the following phases:

1. **Case study:** the first phase was preparatory; this phase aimed to familiarize the participants with the main functions of each service to facilitate the recovery of the memory of past experiences and the focus on the activities

2. **Build on your own experience:** the second phase aimed to encourage critical reflection on the service, focusing the discussion on problematic and positive aspects, both of the services and the experiences linked to them

3. **Related work:** the phase aimed to reflect on examples of other platforms and services, used or not by the participants, to broaden their critical gaze and encourage the creation of services based not only on their own experience but also on that of others

4. **What's next:** the fourth and final phase was the founding one of the workshop and aimed to collect suggestions and ideas on possible services to

create or implement in LivePeople to support researchers in the data collection and management process

To facilitate the management of activities, the interaction between participants within the panels, and the discussion among all, a Miro Board[3] was created. Each panel, therefore, had access to its section on the board and the ability to navigate between the other panels. The ideas were transcribed in the form of Post-it notes attached to each of the relevant sections.

7.2.1 Roles

The workshop was designed, organized and managed by WeNet project members Matteo Busso (M.B.), Amalia De Götzen (A.D.G.) and Ronald Chenu Abente Acosta (R.C.A.), respectively as workshop proposer and RI expert, expert in participatory design and ideation workshop, and expert in management and RI. The roles of the authors, presented by their initials, are as follows:

- *Workshop design*: M. B., A.D.G.
- *Workshop management*: M. B.
- *Panel discussion*: M. B. (Coordinator, Facilitator for Group 1), R.C.A. (Facilitator Group 2), A.D.G (Facilitator, Group 3)
- *Notes and afterthoughts*: M. B., A.D.G, R.C.A.

The order of names reflects the importance of the contribution of the single individuals.

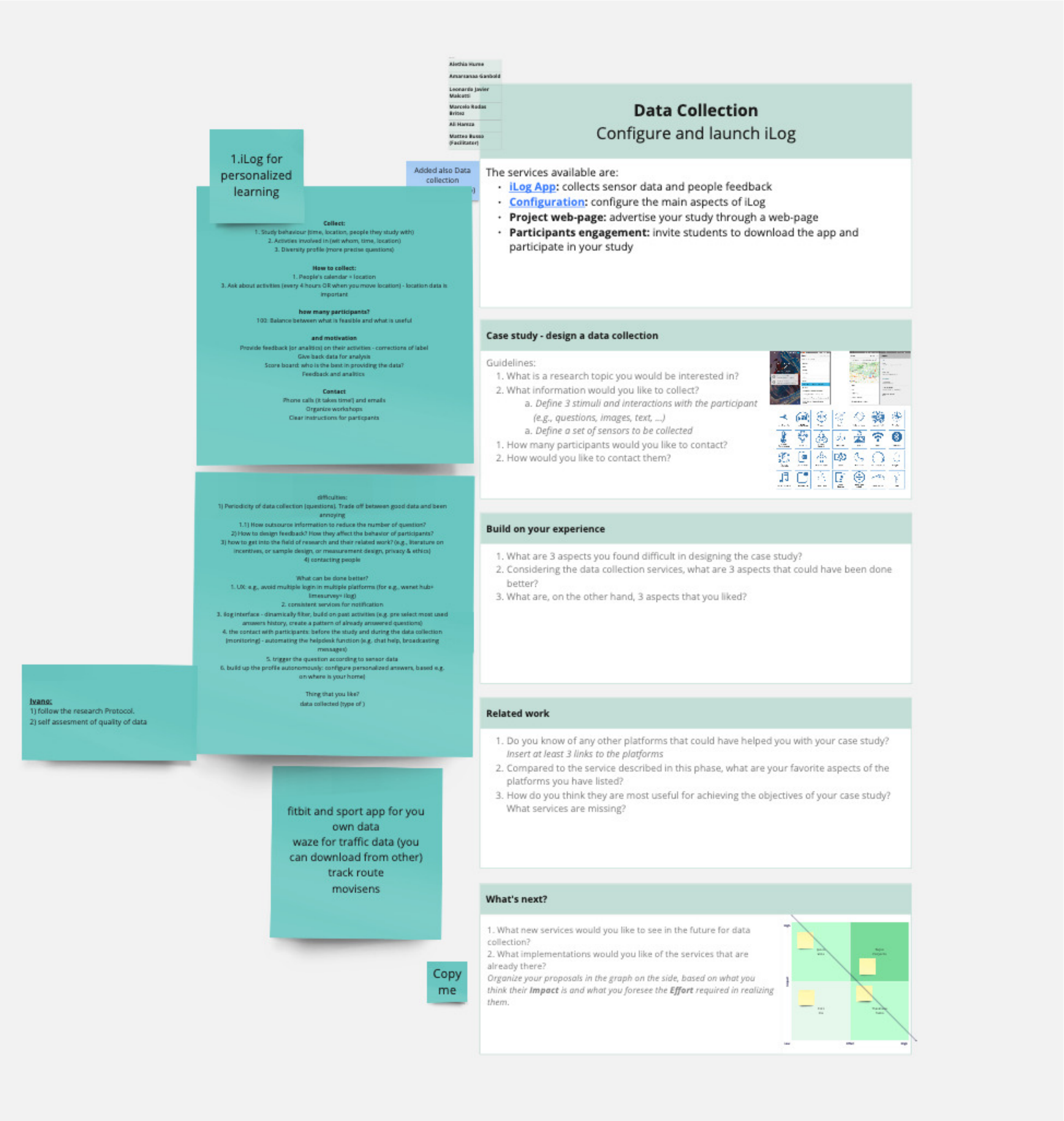
7.3 Results

The *Research Infrastructure Services Ideation Workshop* was held on 21st April in Trento in a hybrid version (both online and in presence) and lasted approximately 4 hours. The workshop involved 17 experts with different backgrounds, from the design and management of data collections to data analysis in computational social science and artificial intelligence, many of whom had already had various experiences within the WeNet Project. The workshop results are described below, divided by each of the panels.

7.3.1 Panel on data collection

During the first part of the workshop, Group 1 focused on data collection and began to reflect on a possible data collection scenario via the iLog app. In the scenario, it was decided to collect data on university students' study behaviour, considering when they study, the places they frequent during study hours, and the people they

[3] Link to the workshop Miro Board: https://miro.com/app/board/uXjVMUufM0c=/?share_link_id=669496392682



Alethia Hume
Amarsanaa Ganbold
Leonardo Javier Malcotti
Marcelo Rodas Britez
Ali Hanza
Matteo Busso (Facilitator)

Data Collection
Configure and launch iLog

The services available are:
- iLog App: collects sensor data and people feedback
- Configuration: configure the main aspects of iLog
- **Project web-page:** advertise your study through a web-page
- **Participants engagement:** invite students to download the app and participate in your study

1.iLog for personalized learning

Added also Data collection

Collect:
1. Study behaviour (time, location, people they study with)
2. Activities involved in (wit whom, time, location)
3. Diversity profile (more precise questions)

How to collect:
1. People's calendar + location
3. Ask about activities (every 4 hours OR when you move location) - location data is important

how many participants?
100: Balance between what is feasible and what is useful

and motivation
Provide feedback (or analitics) on their activities - corrections of label
Give back data for analysis
Score board: who is the best in providing the data?
Feedback and analitics

Contact
Phone calls (it takes time!) and emails
Organize workshops
Clear instructions for participants

difficulties:
1) Periodicity of data collection (questions). Trade off between good data and been annoying
1.1) How outsource information to reduce the number of question?
2) How to design feedback? How they affect the behavior of participants?
3) how to get into the field of research and their related work? (e.g., literature on incentives, or sample design, or measurement design, privacy & ethics)
4) contacting people

What can be done better?
1. UX: e.g., avoid multiple login in multiple platforms (for e.g., wenet hub+ limesurvey+ ilog)
2. consistent services for notification
3. ilog interface - dinamically filter, build on past activities (e.g. pre select most used answers history, create a pattern of already answered questions)
4. the contact with participants: before the study and during the data collection (monitoring) - automating the helpdesk function (e.g. chat help, broadcasting messages)
5. trigger the question according to sensor data
6. build up the profile autonomously: configure personalized answers, based e.g. on where is your home)

Thing that you like?
data collected (type of)

Ivano:
1) follow the research Protocol.
2) self assesment of quality of data

fitbit and sport app for you own data
waze for traffic data (you can download from other)
track route
movisens

Copy me

Case study - design a data collection

Guidelines:
1. What is a research topic you would be interested in?
2. What information would you like to collect?
   a. *Define 3 stimuli and interactions with the participant (e.g., questions, images, text, ...)*
   a. *Define a set of sensors to be collected*
1. How many participants would you like to contact?
2. How would you like to contact them?

Build on your experience

1. What are 3 aspects you found difficult in designing the case study?
2. Considering the data collection services, what are 3 aspects that could have been done better?
3. What are, on the other hand, 3 aspects that you liked?

Related work

1. Do you know of any other platforms that could have helped you with your case study? *Insert at least 3 links to the platforms*
2. Compared to the service described in this phase, what are your favorite aspects of the platforms you have listed?
3. How do you think they are most useful for achieving the objectives of your case study? What services are missing?

What's next?

1. What new services would you like to see in the future for data collection?
2. What implementations would you like of the services that are already there?
*Organize your proposals in the graph on the side, based on what you think their **Impact** is and what you foresee the **Effort** required in realizing them.*

Figure 7.1: Group 1 Miro Board - Data Collection

study with. To investigate the different behaviours in detail, Group 1 reasoned about using diversity-aware profiles to target the questions on the participants, for instance, based on the degree course they attend.

Group 1 then reflected on how to collect information, deciding to select as active data the questions asked every four hours regarding the activities conducted rather than the questions asked every time the participant changes location. As passive data, they considered the possibility of collecting information from participants' calendars and GPS, agreeing that location is a significant source of information.

The number of participants was selected at 100, as it is a good balance between what is feasible in terms of recruitment strategies (in many universities, recruitment takes place through direct and personal contact with participants), the cost of incentives, and what is helpful in terms of data analysis, considering possible participant dropouts. Still considering recruitment, Group 1 highlighted the best strategy is to contact participants via phone, even if this takes a long time. Another aspect that could ensure a good sample size is organizing workshops and providing clear instructions for participants through flyers and manuals.

Group 1 then reflected on the incentives needed to achieve good data quality and considered the possibility of providing feedback (or analytics) to the participants on their activities during data collection, implementing correction mechanisms for the answers provided by the participants, and making inferences about sensors. For example, if the participant usually studies in the library but reports that they are studying in the park, it would be possible to send a confirmatory question to ensure that the participant was not mistaken.

According to Group 1, an essential aspect of incentives would be the possibility of giving back to participants their data for personal analysis and creating a scoreboard in which participants can see who is the best at providing the data to encourage gamification aspects.

7.3.1.1 Considerations on iLog and data collection process

The discussion revolved around three main questions, namely:

1. What aspects did you find difficult in designing the case study?
2. Considering the data collection services, what are three aspects that could have been done better?
3. What are, on the other hand, three aspects that you liked?

The answers to the questions are presented in the following paragraphs.

Difficulties in designing studies with iLog Considering the design aspects of an experiment, Group 1 highlighted some critical points and difficulties. The first aspect concerns identifying the periodicity to send notifications or questions during data collection, as there is a trade-off between good data and annoying. In particular, the most complex aspects to consider concern the possibility of outsourcing information to reduce the number of questions.

Another particularly complex aspect concerns the definition of the feedback given to participants, for instance, through messages encouraging participation. In this field, it is complicated to define how these can affect participants' behaviour, leading them to change their response behaviour, for example, by favouring response set attitudes, in which participants randomly answer questions only to reach a higher number of contributions.

One of the most complex aspects of designing a survey with iLog concerns the state of the art or how to get into the field of research and related work. There is a vast and disparate literature regarding incentives, sample design, measurement design, privacy, and ethics, which involves different disciplinary fields.

A final, particularly complex aspect concerns the strategies for contacting people and recruitment. These vary depending on the local context and can require constant daily commitment, so a person dedicated only to this would be needed.

iLog features to be improved Regarding past experiences with iLog, Group 1 highlighted some aspects that could be improved. The first aspect that emerged concerns the user experience; it was particularly confusing to have different platforms to collect data in addition to iLog. Furthermore, regarding the graphical interface of iLog, one aspect that should be improved concerns the addition of dynamic filters, which allow building on past activities, e.g., pre-select most used answers history, create a pattern of already answered questions. In this way, the respondent burden would be reduced, and the redundancy of applications would be reduced.

A second aspect concerns a more consistent service with notifications, as many questions are lost and do not reach their destination on the participant's smartphone.

As a third point, Group 1 focused on aspects of contact with participants both before the study and during the data collection (monitoring). This aspect requires a significant amount of time and resources. Group 1 believes automating the helpdesk function via chat help or broadcasting messages directly to iLog would be helpful. In this way, some information would not be lost and would not have to be repeated to individual participants.

Positive aspects about iLog Finally, Group 1 focused on the quality of data and interactions via the app. In particular, it emerged that the function of triggering the question according to sensor data and allowing the participant to build up the profile autonomously, configuring personalized answers based, e.g., on where their home is, would be helpful. This way, the participants would not have to answer the same questions, but the location sensor would respond. In general, Group 1 was still satisfied with iLog, particularly with the quantity and quality of the data collected, which made it a unique app and allowed for multiple analyses of people's behaviour.

7.3.1.2 Relevant platforms and technologies

Although many members of Group 1 had performed support and data collection roles in intensive longitudinal investigations, particularly as part of the WeNet Project, they were unaware of any applications that provided a similar service to iLog, apart from Movisens. However, they were aware of sports apps and FitBit from which to download their data, as well as the Waze app for traffic data, which allows you to download other people's track routes.

7.3.1.3 What's next with iLog?

Considering the proposed case study, Group 1 prioritized the possibility of using sensors to trigger questions and validate their quality during data collection. In this way, reducing the respondent burden - asking questions only when necessary - and guaranteeing better data quality would be possible. A second important aspect is that of providing information as a means of encouraging participation. According to Group 1, if a participant has the opportunity to know their data and have feedback on the survey's progress, they would be stimulated to provide more and better quality information. A final aspect concerns the possibility of having iLog pre-configurations to launch the experiments. In this way, even people who are not experts in designing questions and incentives could still start data collection without starting from scratch.

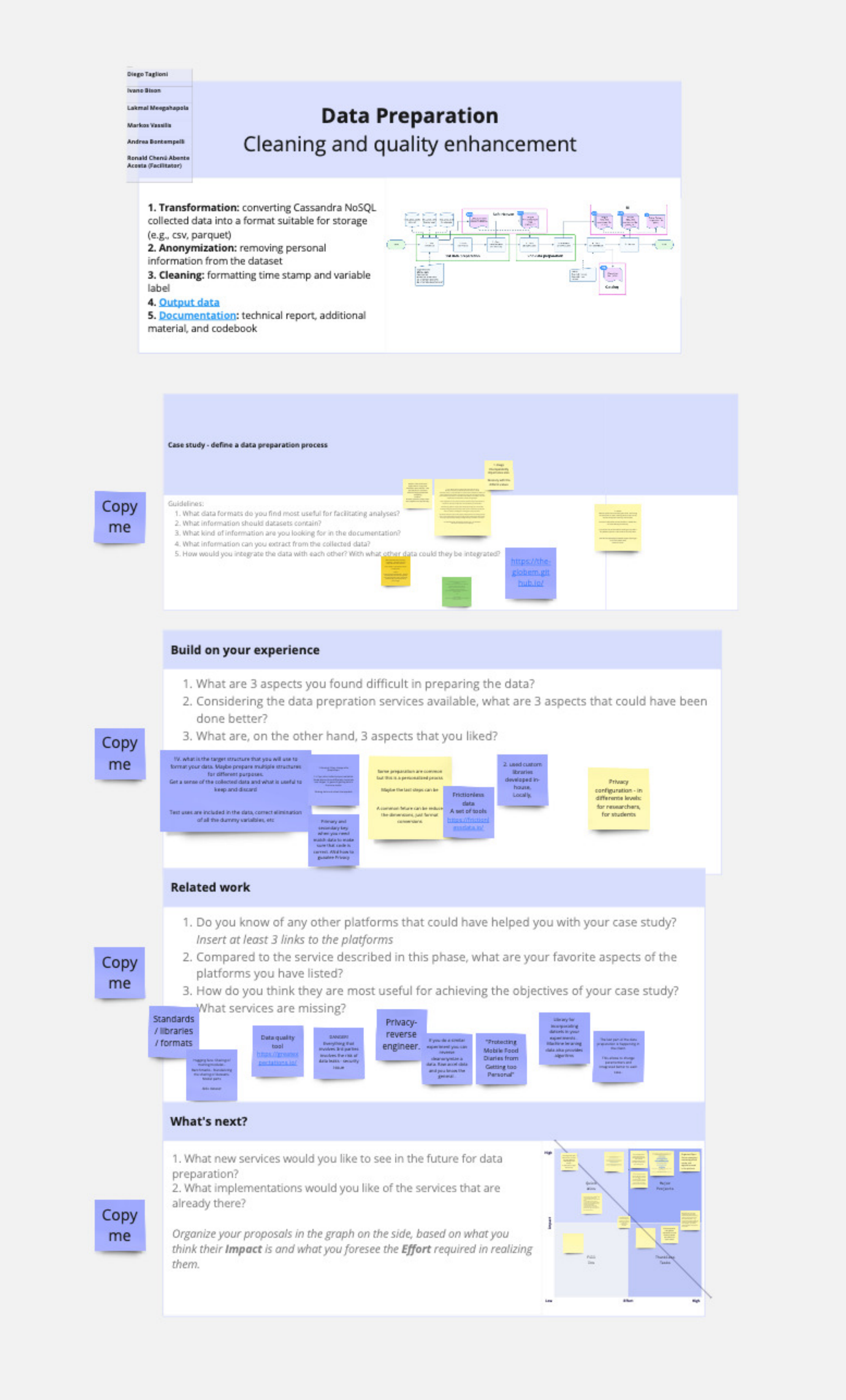
7.3.2 Panel on data preparation

During the first part of the workshop, Group 2 started a discussion on mutual knowledge regarding data preparation aspects. People have different opinions about cleaning data, as each cleaning operation inserts the opinion/models of the cleaner. For this reason, many researchers and analysts prefer to have access to data "as raw as possible". Another aspect to consider is the type of format with which the data is shared, so it is helpful to have a basic format that is as open and shared as possible so that each researcher can convert the data into a valid format.

A second aspect of comparison concerned the adoption of FIHR standards (for medical data) and measuring characteristics of the datasets such as Globem `https://the-globem.github.io/` or using libraries like `https://frictionlessdata.io/` from which to start with the data cleaning processes. In general, Group 2 participants usually adopt validation standards to test the emerging datasets, understand their quality, and have a good view of the data's clean and value. Finally, Group 2 addressed the problem of automating the data-cleaning process. Overall, there seems to be an irreducible personalized part of data preparation that is use-specific and can never be done externally. Chances for automation are at the beginning during the cleaning and reformatting process and at the end, e.g., via offering several data formats with different levels of anonymization to respect privacy.

7.3.2.1 Considerations on data preparation process

Group 2 highlighted how an essential aspect of the data preparation process is not so much in the datasets but in the documentation provided. In this sense, some aspects need to be improved, such as how the variables are named and coded, which must be punctual (i.e., present for each variable in the dataset), intuitive, and understandable by different scientific communities. Similarly, the webpage and the documentation should speak the same language as the target community since using wrong or outdated terminology may imply that the dataset was not generated



Diego Taglioni
Ivano Bison
Lakmal Meegahapola
Markos Vassilis
Andrea Bontempelli
Ronald Chenú Abente Acosta (Facilitator)

Data Preparation
Cleaning and quality enhancement

1. Transformation: converting Cassandra NoSQL collected data into a format suitable for storage (e.g., csv, parquet)
2. Anonymization: removing personal information from the dataset
3. Cleaning: formatting time stamp and variable label
4. Output data
5. Documentation: technical report, additional material, and codebook

Case study - define a data preparation process

Guidelines:
1. What data formats do you find most useful for facilitating analyses?
2. What information should datasets contain?
3. What kind of information are you looking for in the documentation?
4. What information can you extract from the collected data?
5. How would you integrate the data with each other? With what other data could they be integrated?

https://the-globem.github.io/

Copy me

Build on your experience

1. What are 3 aspects you found difficult in preparing the data?
2. Considering the data prepration services available, what are 3 aspects that could have been done better?
3. What are, on the other hand, 3 aspects that you liked?

IV. what is the target structure that you will use to format your data. Maybe prepare multiple structures for different purposes.
Get a sense of the collected data and what is useful to keep and discard

Test uses are included in the data, correct elimination of all the dummy variables, etc

Primary and secondary key when you need match data to make sure that code is correct. And how to guaratee Privacy

Some preparation are common but this is a personalized process.

Maybe the last steps can be

A common feture can be reduce the dimensions, just format conversions.

Frictionless data
A set of tools
https://frictionlessdata.io/

2. used custom libraries developed in-house.
Locally.

Privacy configuration - in differente levels: for researchers, for students

Copy me

Related work

1. Do you know of any other platforms that could have helped you with your case study? *Insert at least 3 links to the platforms*
2. Compared to the service described in this phase, what are your favorite aspects of the platforms you have listed?
3. How do you think they are most useful for achieving the objectives of your case study? What services are missing?

Standards / libraries / formats

Data quality tool
https://greatexpectations.io/

DANGER!
Everything that involves 3rd parties involves the risk of data leaks - security issue

Privacy-reverse engineer.

If you do a similar experiment you can reverse deanonymize a data. Raw accel data and you know the general...

"Protecting Mobile Food Diaries from Getting too Personal"

Library for incorporating datsets in your experiments. Machine learning data also provides algorithms

Copy me

What's next?

1. What new services would you like to see in the future for data preparation?
2. What implementations would you like of the services that are already there?

*Organize your proposals in the graph on the side, based on what you think their **Impact** is and what you foresee the **Effort** required in realizing them.*

Copy me

Figure 7.2: Group 2 Miro Board - Data Preparation

following precise approaches and is of poor quality. Regarding documentation, Group 2 believes having checksums and descriptive statistics is essential to facilitate access to the dataset.

Group 2 then reflected on the risks regarding the data preparation process, particularly that every new platform used increases the risks of data leaks. Hence, it needs to be compelling enough for people to overcome this fear. Furthermore, the deanonymization risk is now higher thanks to AI. Even in anonymized data, if you provide the data collection details, attackers can make a similar data collection campaign to train a model and deanonymize the data from the original experiment. An example that Group 2 reflected on is the paper [238], which demonstrates how inferences of sensitive attributes of users (gender, body mass index category) are possible using a combination of sensor data and self-reports.

7.3.2.2 Relevant data preparation platforms

Group 2 highlighted different platforms and approaches available to manage the data preparation process and increase the quality of the prepared data. Below is the list of identified platforms.

1. Great Expectations https://greatexpectations.io/ is a tool for ensuring data quality through verification by other analysts and researchers.

2. Hugging Face https://huggingface.co/, a platform for sharing training modules for AI research

3. UNESDOC https://unesdoc.unesco.org/ark:/48223/pf0000385841 which is a library for feature extraction to provide "plug and play" data for analysts and researchers to reduce time and data preparation costs drastically

4. RAPIDS https://www.rapids.science which stands for "Reproducible Analysis Pipeline for Data Streams" and is a platform to process smartphone and wearable data to extract behavioural features, visualize data, and create reproducible workflows for data analysis.

7.3.2.3 What's next with data preparation?

Group 2 highlighted possible future activities, dividing them between quick wins and major projects according to the effort to implement the new features. As quick wins, Group 2 listed the following:

1. Providing open source libraries for the beginning (format and easy data fixing depending on the domain) and the end of data preparation (target data)

2. Release the dataset documentation paper in a venue like Nature Scientific Data, NeurIPS Datasets, and Benchmarks. This increases the impact and visibility massively.

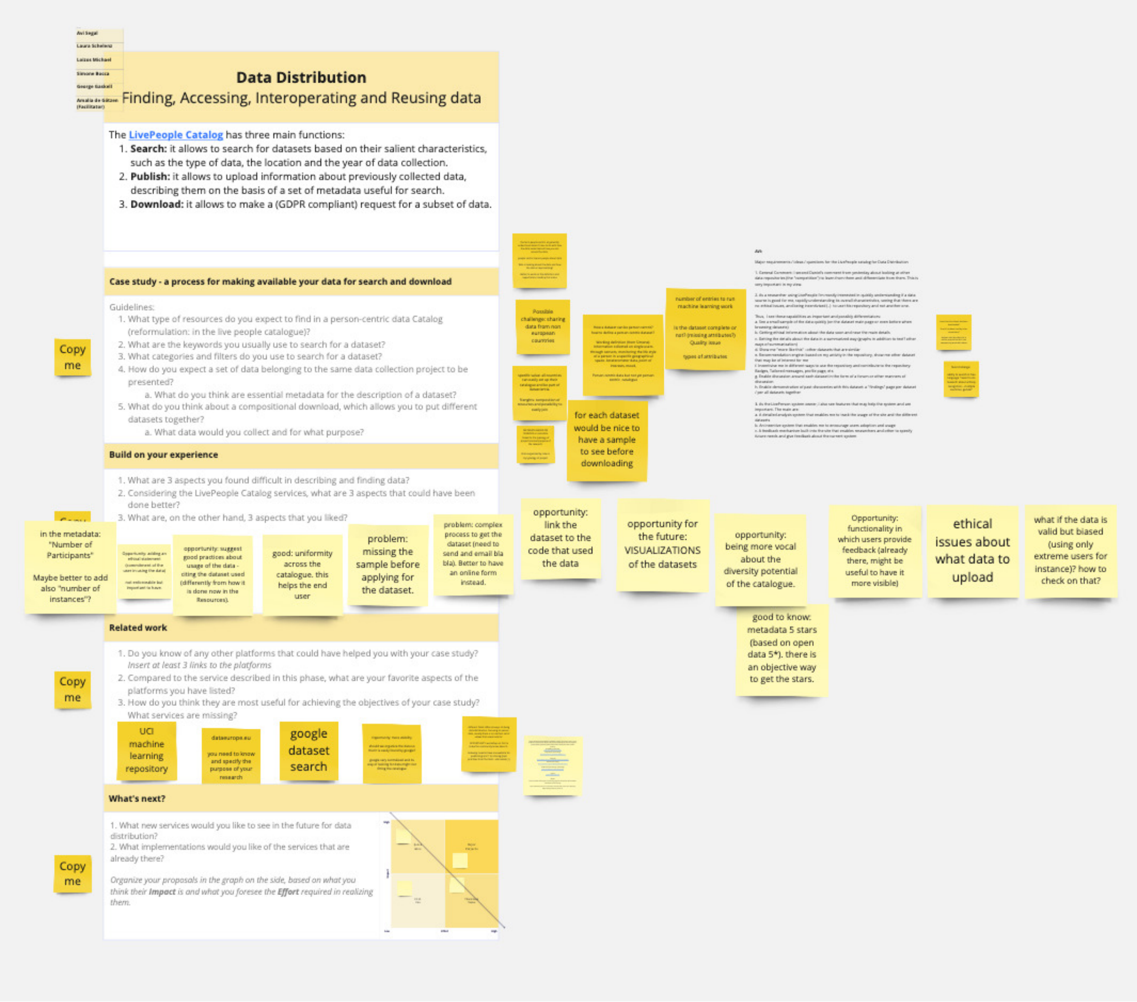
3. Organize workshops selecting the correct high-impact venues, considering the type of data being shared, such as Ubicomp and CHI, which are particularly interested in sensor data.
4. Provide clear data documentation considering:
    - The intended purpose of each offered dataset;
    - The way data has been cleaned;
    - The way any (meta-)analysis has been conducted, alongside any related published scripts and software;
    - What would consist of proper and improper use of our datasets.

As major projects, Group 2 listed the following:

1. Domain-specific study of the terminology and standards to be able to talk to each community in their terminology and standards
2. Privacy configurator: a tool for preparing datasets with different levels of privacy protection for other uses. This should be able to provide the same dataset with varying levels of noise to be used by students/researchers/companies
3. Creating a benchmark and a proper website for the particular dataset. We can have LivePeople for data sharing, but a project/dataset specific should be maintained to publicize/market datasets properly. This is standard practice in communities that work with sensor data: Ubicomp, ISWC, Mobisys, Mobicom, and IPSN. Below are some suggested platforms for data sharing and advertising.
    - GLOBEM for publicizing `https://the-globem.github.io/`,
    - Physionet `https://physionet.org/content/globem/1.1/` ,
    - LifeSnaps `https://www.nature.com/articles/s41597-022-01764-x`,
    - Zenodo `https://zenodo.org/record/6832186#.ZEJdf-xBznw`,
    - StudentLife `https://studentlife.cs.dartmouth.edu/`
    - eSense `https://www.esense.io/`
4. Organized Open Source repository containing all the scripts and algorithms used in the platform

7.3.3 Panel on data distribution

An initial discussion revolved around the term "people-centric" used to define the LivePeople Catalog. According to Group 3, the term people-centric, as generally understood, doesn't have to do with how the data looks but how you can access the data, so it is a matter of accessibility of the platform and also about who the



Avi Segal
Laura Schelenz
Loizos Michael
Simone Bocca
George Gaskell
Amalia de Götzen (Facilitator)

Data Distribution
Finding, Accessing, Interoperating and Reusing data

The LivePeople Catalog has three main functions:

1. **Search:** it allows to search for datasets based on their salient characteristics, such as the type of data, the location and the year of data collection.
2. **Publish:** it allows to upload information about previously collected data, describing them on the basis of a set of metadata useful for search.
3. **Download:** it allows to make a (GDPR compliant) request for a subset of data.

Case study - a process for making available your data for search and download

Guidelines:

1. What type of resources do you expect to find in a person-centric data Catalog (reformulation: in the live people catalogue)?
2. What are the keywords you usually use to search for a dataset?
3. What categories and filters do you use to search for a dataset?
4. How do you expect a set of data belonging to the same data collection project to be presented?
   a. What do you think are essential metadata for the description of a dataset?
5. What do you think about a compositional download, which allows you to put different datasets together?
   a. What data would you collect and for what purpose?

Build on your experience

1. What are 3 aspects you found difficult in describing and finding data?
2. Considering the LivePeople Catalog services, what are 3 aspects that could have been done better?
3. What are, on the other hand, 3 aspects that you liked?

Related work

1. Do you know of any other platforms that could have helped you with your case study? *Insert at least 3 links to the platforms*
2. Compared to the service described in this phase, what are your favorite aspects of the platforms you have listed?
3. How do you think they are most useful for achieving the objectives of your case study? What services are missing?

What's next?

1. What new services would you like to see in the future for data distribution?
2. What implementations would you like of the services that are already there?

*Organize your proposals in the graph on the side, based on what you think their **Impact** is and what you foresee the **Effort** required in realizing them.*

Copy me
Copy me
Copy me
Copy me
Possible challenge: sharing data from non european countries
number of entries to run machine learning work
is the dataset complete or not? (missing attributes?) Quality issue
types of attributes
for each dataset would be nice to have a sample to see before downloading
in the metadata: "Number of Participants" Maybe better to add also "number of instances"?
opportunity: suggest good practices about usage of the data - citing the dataset used (differently from how it is done now in the Resources).
good: uniformity across the catalogue. this helps the end user
problem: missing the sample before applying for the dataset.
problem: complex process to get the dataset (need to send and email bla bla). Better to have an online form instead.
opportunity: link the dataset to the code that used the data
opportunity for the future: VISUALIZATIONS of the datasets
opportunity: being more vocal about the diversity potential of the catalogue.
Opportunity: functionality in which users provide feedback (already there, might be useful to have it more visible)
ethical issues about what data to upload
what if the data is valid but biased (using only extreme users for instance)? how to check on that?
good to know: metadata 5 stars (based on open data 5*). there is an objective way to get the stars.
UCI machine learning repository
dataeuropa.eu
you need to know and specify the purpose of your research
google dataset search
Impact
Effort

Figure 7.3: Group 3 Miro Board - Data Distribution

data in the platform represents. Once a working definition of "person-centric" data was set up as "information collected on single users through sensors, monitoring the lifestyle of a person in a specific geographical space," Group 3 agreed that the Catalogue is about person-centric data. Still, it is not yet a person-centric catalog. Another positive characteristic of the platform is the uniformity across the Catalog, which helps the end user.

As a researcher using LivePeople, Group 3 was primarily interested in quickly understanding if a data source is good for me, rapidly understanding its overall characteristics, seeing that there are no ethical issues, and being incentivized to use this repository and not another one. Thus, Group 3 sees these capabilities as essential and possibly differentiators:

1. See a small sample of the data quickly (on the dataset main page or even before when browsing datasets)
2. Getting ethical information about the data soon and near the main details

3. Getting the details about the data in a summarized way (graphs in addition to text? other ways of summarization)
4. Show me "more like this": other similar datasets
5. Recommendation engine: based on my activity in the repository, show me another dataset that may be of interest to me
6. Incentivise me in different ways to use the repository and contribute to the repository: Badges, Tailored messages, profile page, etc.
7. Enable discussion around each dataset in the form of a forum or other manners of discussion
8. Enable demonstration of past discoveries with this dataset: a "findings" page per dataset / per all datasets together

As the LivePerson system owner, Group 3 also sees features that may help the system and are important. The main are:

1. A detailed analysis system that enables me to track the usage of the site and the different datasets
2. An incentive system that enables me to encourage users' adoption and usage
3. A feedback mechanism built into the site that enables researchers and others to specify future needs and give feedback about the current system

7.3.3.1 Considerations on data distribution process

During the second activity, Group 3 focused on defining the main problems and proposing opportunities and solutions.

- Problem: missing the sample before applying for the dataset
- Opportunity: add the sample description
- Problem: complex process to get the dataset (e.g., need to send an email). It is better to have an online form instead.
- Opportunities
  - suggests good practices about data usage - citing the dataset used (differently from how it is done now in the Resources).
  - links the dataset to the code that used the data
  - provides dataset visualizations to allow a better exploration of the features
  - being more vocal about the diversity potential of the Catalogue.
  - Add functionality in which users provide feedback (already there; it might be useful to have it more visible)

- Problem: what if the data is valid but biased (using only extreme users, for instance)? How do you check on that?
- Opportunity: adding an ethical statement (the user's commitment to using the data) is not enforceable but important to have.

A final good note is that using the 5 stars principle to denote the quality of the data is particularly appreciated since there is an objective way to get the stars.

Relevant data distribution platforms Group 3 suggested having a look at the UCI machine learning repository and to dataeurope.eu even though within these repositories, you need to know and specify the purpose of your research.

7.3.3.2 What's next with LivePeople Catalog?

According to Group 3, the most relevant aspect to consider in the future is the advertisement of the LivePeople Catalog. A good opportunity for more visibility is to organize the data so that it is easily found by Google even if it is very centralized, and its way of looking for data might not fit the Catalogue. A workaround could be to connect the LivePeople Repository to Google dataset search, in line with some best practices on indexing.

7.4 Discussion and conclusion

Section 7.1 shows how important it is to create RIs that drive the creation of technologies driven by social innovation. Following the WeNet approach, three essential components of RI have been identified which concern data collection, preparation and distribution services. Each of these services was presented and discussed within three different panels.

The results show how in general the approach considers the salient aspects of an RI and already presents a set of services useful for the purpose. Most of the services presented were considered sufficient to conduct the data collection and management process, even if various implementations were proposed.

In particular, regarding data collection, Group 1 highlighted that iLog and the design methodology are unique and very useful for the study of human behaviour. In any case, to facilitate the configuration of the app it is necessary to design and develop software that helps interface with the app, facilitating its deployment even for non-experts. A secondary implementation aspect concerns the improvement of the user interface, making it more intuitive and engaging for participants.

As for data preparation, Group 2 argued that the general pipeline for the production of raw data compliant with the GDPR is robust concerning different disciplinary approaches, but it is complex to be able to create specialized datasets for individual disciplines. It is therefore advisable to define a second phase of data preparation and feature extraction adaptable to the needs of individuals, making the dataset prepared in the first phase available to everyone. Group 2 also highlighted that some tools and standards can facilitate this second preparation.

As for the data distribution, Group 3 found the Catalog particularly useful, proposing some specific implementations regarding data description, search and download. Some metadata fields have been suggested as additional (for example, the number of rows contained in datasets) and the use of tagging and aggregation of datasets so that they are more meaningful and intuitive for the end user should be considered. Finally, the download procedures should be simplified, for example through the use of online forms and the possibility of selecting different datasets in which the user is interested.

In conclusion, the workshop contributed to both the validation and the creation of services connected to the iLog methodology, paving the way for future developments.

And you may ask yourself, "Well, how did I get here?"
Letting the days go by, let the water hold me down
Letting the days go by, water flowing underground
Into the blue again, after the money's gone
Once in a lifetime, water flowing underground

— Talking Heads *Once in a lifetime*

8 Conclusion

Contents

The thesis introduced a new methodology for designing, collecting, managing and distributing complex data to model people's perspectives embedded in their context. Starting with Big Data, the thesis showed how, despite the abundance of information, it lacks crucial variables for modelling personal context. Big Data are often "thin"; namely, they extend their volume over a few variables, often collected by sensors, in which the person's point of view on fundamental aspects such as the places where they are located, their interactions, their events, emotions and feelings are not represented. In contrast, various existing approaches, especially from social science disciplines like psychology and sociology, often collect "thick" data, which is full of personal information but is scarce in quantity or limited in duration and space coverage. The distinction between thick and thin data can be assimilated to that between annotations and sensors, where the former is information self-reported by people who bring their point of view on the context and are helpful in framing and providing meaning to the latter. On the other hand, sensors, thanks to their pervasiveness (but not invasiveness), allow us to observe people in all moments of their daily lives, providing vital information on their behaviour that is difficult to report verbally (at least not with the same frequency with which the sensors collect them).

Alongside [15], the thesis advocated for the creation of new data that is both dense and extensive, i.e. Big Thick Data (see also Section 1.1). Such data pose new challenges for their management, modelling and organization, which can be achieved with the help of Knowledge Graphs built to represent the person's point of view at every moment of their everyday life, i.e. their Situational Context [239] (see Section 1.2), allowing for scalability, native integration with annotations, and data reuse.

Although this approach provides theoretical and ontological guidance to enable research and innovations that integrate the perception of context with the information streams from the devices surrounding us, it says little about epistemology. In particular, quantifying Big Thick Data poses several methodological challenges towards collecting valid and reliable data, managing data quality and privacy, and enabling data reuse. In other words, to be considered valid, quantifying Big Thick Data via Situational Context needs to be verified empirically through observations, i.e., through collecting valid and reliable data. Furthermore, the reader will have noticed how the concept of Big Thick Data is eminently interdisciplinary. The interactions with the person and of the person with his own self and the environment surrounding him have been the object of study of the social sciences for more than a century. The collection of data from sensors, their modelling, and the design and implementation of the devices themselves are addressed by engineering and IT. Ethical considerations and privacy regulations to manage interaction with the person and the processing of personal data to prevent damage and respect rights are a matter of study by philosophers and jurists.

Therefore, Chapter 2 explored current approaches to the design, collection and management of data from an interdisciplinary perspective. It also considered aspects relating to their distribution and reuse, motivated by the fact that a complex and interdisciplinary theory requires multiple evaluations by experts from different sectors to be considered valid. Finally, the Chapter focused on tools and support materials for executing all methodology phases, evaluating when practitioners with different domain expertise can replicate this. In this sense, Section 2.1 showed how there is a consolidated approach to social investigations which involves the consideration of different methodological procedures (see, e.g., [27]) aimed at limiting the errors that occur in the operationalization and data collection phases (see, e.g., [28]) ultimately to generate valid and reusable data. Of course, since Big Thick Data are data about people, Section 2.2 showed how there are ethical and privacy principles and guidelines for collecting and managing this data, including its distribution.

Similarly, extensive literature has focused on data preparation and sharing (see, e.g., [42]), as from Sections 2.4 and 2.5. Alongside methodologies, many disciplines have created their own platforms to support practitioners, with tools ranging from consultancy to education to technologies for data collection, preparation and databases for data sharing. Yet, none of the disciplines consider a comprehensive approach to collecting self-reported contextual annotations and sensors in streaming, that is, those data fundamental for exploring the situational context. Furthermore, materials and supporting technologies often remain confined to their respective disciplinary fields, effectively preventing reuse by non-experts.

In summary, despite the extensive methodological literature and the platforms developed in various disciplinary fields, no current end-to-end methodology can effectively handle Big-Thick Data.

Consequently, Chapter 3 proposed a comprehensive methodology addressing essential Big Thick Data generation aspects. The methodology showed a concrete step forward compared to state-of-studies in sociology (e.g., [27]), in ESM studies [53] and in those of HAR (e.g., [55]) not only because it is based on an innovative application for data collection (i.e., iLog [26, 240]), but because it integrates aspects from different disciplines towards the creation of quantitative Big Thick Data. This occurs through conceptualizing the annotations provided by the person not as dissociated from the sensors but integrated into the personal context. Each annotation the person provides regarding activities, social interactions, and places visited corresponds to a set of sensors that enrich the understanding of behaviour in context. In doing so, the methodology is equipped with a process for managing personal data, i.e., preparation and distribution, which is ethics-aware and privacy-compliant. Additionally, the method incorporates services to enhance replicability, ensuring the generated data can be easily reproduced and utilized for various purposes. In this sense, considering the leading platforms for data collection and their distribution, the services offered by the iLog methodology are the only ones that are end-to-end, i.e., they cover all the salient aspects of personal data management.

By introducing this methodology, the thesis aimed to contribute to advancing research and practices in managing Big-Thick Data. It seeks to fill the existing gap in methods for handling complex data that capture the richness of individuals' contextual information while adhering to ethical standards and privacy considerations.

The demonstrated applications of the methodology showcase its versatility in collecting diverse types of data (see Chapters 4, 5, and 6) that align with the various components of the context described in the Chapter 1, and its validity, according to the experts' opinion provided in Chapter 7. These chapters are of fundamental importance to the thesis, highlighting two critical aspects: the validity and reliability of the collected data and the effectiveness of the methodology's services in terms of validity and usability. Therefore, the subsequent sections present additional considerations that contribute to a more comprehensive understanding of the strengths and potential challenges associated with the proposed methodology and its services.

8.1 On validity and reliability

Ultimately, a methodology is evaluated by the reliability of the data collected and, in our case, by its reuse at an interdisciplinary level, which demonstrates the validity of the operationalization of concepts into observed variables. Two case studies were presented to show how the methodology applies to complex studies on the diversity of everyday life and social practices (Case Study I) and to verticalize aspects of the situational context, such as social interactions (Case Study II). Both case studies produced datasets widely reused in various disciplinary fields, thus demonstrating the validity of the approach and the reliability of the results. The main findings are reported below.

Case Study I The notion of Situational Context as a means of quantifying Big Thick Data provides that it is possible to reconstruct the context of the person if, for each situation experienced, there is information regarding the event in which they are participating, where they are, who they are with and how do they feel. This information can come from self-reported annotations and personal device sensors (e.g., smartphones). Thus defined, personal context varies from person to person but also from (cultural) context to (cultural) context.

Chapter 4 showed a comprehensive approach to reconstructing the situational context of 700+ individuals from eight countries. Following the iLog methodology, a set of measurements was designed to collect information on the personal context and frame it at a social and cultural level through the notion of social practices [210]. Therefore, a survey lasting one month was proposed during which participants provided information on their context by answering four questions every half hour (i.e., "*What* are you doing?"; "*Where* are you?"; "*Who* are you with?" ?"; "What's your *mood*?") as well as information from all the smartphone's sensors (e.g., GPS, Bluetooth, application running on the device). Furthermore, they responded to three questionnaires regarding their social practices (e.g., participation in cultural events, sporting activities, shopping and cooking habits) to encourage a better comparison of cultural aspects.

The result of the study was 216k+ annotated situational contexts based on 300+ GB of sensor data. This database has favoured the proliferation of scientific literature concerning aspects of the situational context, showing how the use of annotations favours the prediction of human behaviours [61], such as the patterns of social media usage impact academic activities [223], but also how the prediction of [225] mood and [226] activities through sensors is favoured by the use of Big Thick Data and their cultural framing through the multi-country approach.

In this sense, Case Study I shows how the data collected through the methodology is reliable, enabling accurate analyses of the person's context. Furthermore, it shows how the approach is valid at an interdisciplinary level as different scientific communities adopt it.

Case Study II If Case Study I shows how to obtain notes on the person's context, it does not explain how the main aspects can be explored more deeply towards a *thicker* observation. Furthermore, the study was conducted in a research context in which participants were paid for an activity they would not carry out in their daily lives (i.e., answering questions with high frequency), thus posing a sustainability problem for longer-term studies.

For these reasons, Case Study II (see Chapter 5) delved into the verticalization of individual contextual aspects, focusing specifically on the person's social interactions. In this case, ≈ 200 participants from five countries were invited to install AskForHelp and provide their sensor data through the iLog application (as in Case Study I). AskForHelp was designed to enable interactions between people who do not know each other through a Q&A mechanism, in which a participant could ask anything of interest and receive answers from a subset of people selected based on a pool of characteristics that made them more suitable to respond (e.g., if a participant asked for information on a location, the people who were nearby were involved). During data collection, participants received badges and messages with a gamification approach, encouraging participation to achieve different objectives (e.g., asking a certain amount of questions). This resulted in ≈ 2000 questions with 5000+ answers and 10+ GB of sensor data.

Thus, the application facilitated interactions mapped onto smartphone sensors, providing valuable data for enhancing contextual understanding of the person's social context. Furthermore, the study shows the effectiveness of using non-monetary incentives. The interaction with the participants and the feedback provided through messages and badges have aroused particular enthusiasm and interest on the part of the students involved (see also [228, 241]).

This study has also been the subject of several scientific publications, such as [59, 230, 231], showing both the validity and replicability of the data collected through the iLog methodology.

8.2 About the end-to-end methodology and its services

As mentioned in Chapter 2 and extended throughout the thesis, a fundamental aspect of a methodology is its reproducibility. A study is considered replicable when different researchers obtain the same results by repeating the same procedures. Clearly, in an interdisciplinary methodology, the services and support materials play a fundamental role in facilitating the correct execution of the different phases. Therefore, the validity and usability of the methodology's services are pivotal for the replicability and applicability of a Big Thick Data study. In this context, validity refers to the services' capability to generate meaningful and accurate results. At the same time, usability pertains to the ease with which researchers can employ these services for their studies. Case Studies III and IV provide several insights regarding these two aspects.

Case Studies III Chapter 6 described two case studies showing how the methodology was replicated independently by two researchers with different backgrounds, producing reliable results. In particular, researchers from the University of Thessaly (Greece) with a background in economics were interested in studying students' daily behaviours and leisure activities. On the other hand, researchers at FTP University (Vietnam) focused on studying daily eating patterns to map unhealthy behaviour. Both researchers had access to examples of previous data collections (see LivePeople Catalog [172]). They received the support materials for the design of the study proposed in Appendix A, as well as those relating to ethical and privacy aspects (Appendix B) and to the management of data collection (Appendix C). In this way, they could design and conduct the investigation, assisted by technical support for the study configuration via the iLog app and sensor data preparation.

In both cases, intensive longitudinal investigations were proposed. The first aimed at collecting information on daily activities with questions asked every half hour, while the second aimed at observing eating activities, with questions asked every two hours. The duration of the two studies was one month, reaching, in both cases, more than 100 participants and with response rates in line with previous investigations. Furthermore, executing the entire methodology took approximately six months, significantly reducing the timescales compared to the Diversity 1 survey described in Chapter 4.

Overall, both case studies showed the replicability of the methodology and the usability of the support materials by non-experts, also providing helpful feedback towards its future implementation (see Sections 8.3 and 8.4 on follow).

Case Study IV As mentioned in Section 1.4, the methodology and supporting materials are an integral part of a research infrastructure (RI) whose foundations have been defined in the WeNet Project and whose ultimate goal is to help the entire process of generation and sharing of complex data such as Big Thick Data. With this in mind, the workshop described in Chapter 7 was carried out, which aimed to validate the services already developed in the Project and design new ones to make RI a product usable by researchers.

Therefore, the workshop was focused on the primary services offered by the RI, namely data collection (described in Section 3.3), preparation (Section 3.4) and data sharing (Section 3.5). Seventeen experts from different disciplinary fields were divided into three panels, each focusing on one of the services. The workshop included an initial reflection and evaluation of existing services to convey the ideation process. From this first aspect, the importance of adopting approaches centred on the person and personal data to promote human-aware research and technologies emerged. In particular, the data collection and distribution services appeared to be fundamental, without which the experts would not have been able to replicate the type of studies proposed in this thesis.

Some limitations were then highlighted, including the difficulty in conceiving and designing a study on Big Thick Data, the timing and outputs of data preparation and the types of metadata, as detailed in Section 8.3. In any case, these were supported by various implementation ideas, which contributed to the definition of future work described in Section 8.4.

8.3 Limits

Despite the general enthusiasm reported by the experts who applied the methodology and the extensive reuse of the data collected, some limitations emerged from the various case studies.

First, there is an unequal participation rate in the different pilot studies. In particular, in some pilots reported in Chapters 4 and 5, particularly low participation numbers were recorded in countries such as India, Paraguay and Denmark compared to those of Mongolia and Italy. This aspect is probably due to the complexity of the recruitment strategies (e.g., finding the participants, communicating the contents of the survey and encouraging participation). It highlights the need to adopt different protocols depending on where the study occurs. This hypothesis is corroborated by the case studies described in Chapter 6, where the recruitment procedure took place in Vietnam in several waves. In contrast, in the case of Greece, sending a few emails inviting participation was sufficient.

Still considering participation, another limitation concerns dropout rates. These are partly due to the functioning of the data collection technology, the limits of which are not the subject of this thesis. Instead, the central aspects concern the adaptation of the investigations in a cross-cultural context, the use of multiple platforms for data collection, the monitoring of the studies and the incentive system. As regards the first aspect, it is clear that the formulation of the questions, but also their number and frequency (i.e., the respondent burden) are perceived differently depending on the culture, which is why, in addition to following the standards for translation of surveys, more time needs to be spent on adapting surveys at the local level. Secondly, it emerged that active monitoring and the helpdesk are particularly complex and time-consuming aspects that require more in-depth training for experiment supervisors. Third, the use of different platforms for data collection has disadvantaged mainly participation, where many dropouts have occurred when switching from one platform to another (e.g., from LimeSurvey to iLog or from LimeSurvey to AskForHelp to iLog). Finally, there is evidence that different incentives correspond to different participation rates.

Regarding the use of support materials and the application of the methodology, the design aspects are among the most complex to conduct. In particular, the process that goes from the research idea to the definition of the measurements is complex to articulate if there is no reference to a modus operandi and a set of literature and standards that facilitate the process. Added to this is the problem of approval by ethics committees, which sometimes can take several months.

In addition to the aspects of data collection (described above), the data preparation has attracted some criticism from experts, who generally prefer to access the raw data and manage it independently for research purposes. Furthermore, some issues emerged regarding the quality of the prepared data, which led to the withdrawal of the datasets and the creation of a new version.

Finally, regarding distribution, some minor limits concern the absence of some metadata considered particularly useful by some experts (e.g., the number of rows of the dataset and the amount of missing data). Particularly relevant is the problem of searching for datasets associated with their download. The absence of a more significant number of tags and filters and of an automated procedure for requesting the dataset has, until now, discouraged several practitioners interested in their reuse.

8.4 Future Studies

Looking ahead, an aspect that appears relevant to the reliability of the data collected concerns the incentives for participation, i.e., all those mechanisms that, on the one hand, simplify the activities of the participant and, on the other, amplify their involvement.

To simplify participants' tasks, maintaining the quality of the data and the number of annotations necessary for behavioral analyzes in daily life, one area that seems promising is linked to the Skeptical Learning algorithm [60, 62, 181]. The algorithm is aimed at validating annotations provided by the user and can be applied in real time during an experiment. Presumably, after a short training period, the algorithm could work with an Active Learning approach, predicting the different annotations which should then be validated by the participant, thus reducing the number of questions from 48 per day (as in the case of the Diversity study 1) to just one per day.

Considering greater participant engagement during data collection, studies in the field of gamification are certainly promising, as is the badges and messages approach described in Chapter 5. These should be accompanied by recruitment materials, such as videos and web pages. A further implementation would be to provide a personal dashboard to each participant, through which they can view the amount of their contributions, as well as statistics deriving from them.

From the services and support material point of view, the Citizen Science approach shows how the creation of communities of people interested in research and science is favorable both for citizens who have the opportunity to approach topics of interest to them, and for researchers, who not only have access to a crowd of participants , but also, and above all, to the relationships and exchanges of ideas and opinions that derive from being part of a community. Therefore, the creation of a community based on a Human-Aware AI approach would favor the proliferation of studies and ideas regarding Big Thick Data. This would facilitate the design and execution of studies, which has proven to be a problematic point, as reported in the Chapter 7 workshop.

Still considering the design, an aspect to address is the user-friendly configuration of the data collection service. This is not only to make the service scalable (as happens with platforms such as LimeSurvey and Qualitrics [118, 119]) but also to encourage greater understanding of the measurement tools, facilitating communication and the adaptation of studies.

In this sense, the Catalog, as such, is an essential service, where the search and data visualization functions are a vital help not only for distribution purposes but also for the design of the research itself. Therefore, the implementation of some features of the Catalog, such as the possibility of searching datasets by projects or research approaches, as well as the facilitation of data download for exploration and analysis, are an aspect to be considered in future work.

Appendices

A

Study design support material

Contents

This section presents the support material for the design phase.

A.1 Study design template

This document is a support in the process of designing and managing a research study. In our framework there are four phases of the research and data collection process:

1. Research design and management (of which this document is the template)

2. Ethics and privacy procedures

3. Data collection

4. Data preparation and upload

In particular, the Research design and management phase consists of five different tasks:

1. Research design: that is the definition of the research objectives, the topic to be investigated, and the methods of investigation (sample, investigation tools, incentives).
2. Management and planning: that is the management of times and personnel to conduct the four phases of the research process.
3. Ethics and privacy compliance: i.e., the verification that one's research is conducted according to the ethical and legal dictates of one's country and of the European Union (GDPR) where necessary.
4. Dissemination: the communication strategies of the a priori and a posteriori scientific research product of the data collection campaign.
5. Afterthoughts: considerations on the process.

For a complete example, please have a look at [109]

A.1.1 Research Design

A.1.1.1 Research Goal

This section should describe:

1. Purpose of the research and theoretical justification (max 15 references)
2. General purposes
3. Hypothesis and specific objectives
4. Expected results

A.1.2 Survey methodology

This section should describe the survey measurement and the timing. To facilitate communication with the technical leader, the "iLog: questions and sensors specifications template" (see A.2) should be used to design the stimuli to be administered via the iLog app.

A.1.2.1 Number and possible stratification of the sample

This section should describe the population to be investigated.

A.1.2.2 Recruitment strategy

This section should describe the ways and means to involve the participants.

A.1.2.3 Incentives

This section should describe the incentives (monetary and/or non-monetary) given to the participants.

A.1.3 Management Planning

This section defines the plan for conducting the study. This entails defining the timing and roles for each phase of the Study to ensure its successful completion. Within our paradigm, the Research Leader oversees the life cycle of the project, focusing on the research design and management. This includes defining the suitable schedule (i.e., Gantt), listing the requirements of each role, guiding defining feasible outcomes such as assessing the need for a request for an opinion from the Chairman of the Research Ethics Committee (CESP), deliverables, milestones, and so on. Beyond the Research leader, there are 2 main roles:

1. Study leader: The role of the Study leader is to support the Research leader in overseeing the study, especially for the sample and personnel design. This also includes designing incentives and recruitment strategies, while also ensuring the overall compliance of the study with privacy and ethical regulations. Furthermore, it also manages the actual running of the study.

2. Technology leader: The role of the technology leader is to oversee the technological development of the platform by tightly collaborating with the Study leader to meet the requirements of the Study, especially for the iLog app configuration and the helpdesk.

It can be useful to define a Gantt, specifying the tasks of each role.

A.1.4 Ethics and Privacy Compliance

This section should include the need assessment for a request for an opinion from the Chairman of the Research Ethics Committee (IRB). The following questions are asked to aid in the decision (if the answer to questions 2 to 9 is yes, then an opinion from the IRB is needed).

1. Is the data collection part of a project that has already received an opinion from the CESP?

2. Will the project involve the collection of new information about individuals?

3. Will the project compel individuals to provide information about themselves?

4. Will information about individuals be disclosed to organisations or people who have not previously had routine access to the information?

5. Are you using information about individuals for a purpose it is not currently used for, or in a way it is not currently used?

6. Does the project involve you using new technology that might be perceived as being privacy intrusive? For example, the use of biometrics or facial recognition.

7. Will the project result in you making decisions or acting against individuals in ways that can have a significant impact on them?

8. Is the information about individuals of a kind particularly likely to raise privacy concerns or expectations? For example, health records, criminal records, or other information that people would consider to be private.

9. Will the project require you to contact individuals in ways that they may find intrusive?

If the IRB opinion is needed, this section should also include all relevant aspects regarding the drafting of documents and interactions with the Ethics Committee and the Privacy Office. Support documents for this process can be found in the Appendix B.1.

A.1.5 Dissemination

This section describes the scientific dissemination before and after the data collection. Particularly:

1. **Before**: because of the reproducibility crisis, many journals require the study to be pre-registered. In this case, insert here the documentation or the link to the site where the registration was made (e.g., OSF)

2. **After**: indicate the tentative title of the publication(s) and of the data descriptor. If no pre-registration has been made, add a short description for each paper. Once completed and published, indicate the bibliographic reference here.

A.1.6 Afterthoughts

This section should provide insights and lessons learned after the end of the study to be addressed in future iterations.

A.2 iLog: questions and sensors specifications template

This document is a template for specifying the questions and sensors that should be collected via iLog. According to the experience sampling method framework, questions can be of three types:

1. Signal Contingent: questions are asked randomly throughout the day at random times.

2. Interval Contingent: questions are asked at fixed times. Time diaries are part of this category.
3. Event contingent: the participant reports on the events as they happen.

Depending on the length of the survey and the number of questions, it may be useful to add break options to lighten the participant response burden.

As for the sensors, depending on the type they can be collected on charge (i.e., when the event happens) or with predefined frequencies.

The tables and examples in the document serve as guidelines to specify the elements needed to configure iLog.

A.2.1 Overall characteristics of the study

Title:	
Description:	
Start Date:	
End Date:	
IRB Approval N.:	

Table A.1: Data collection information

A.2.2 Experience sampling questions

A.2.2.1 Signal contingent

Code	**Timing**	**Question**	**Answer options**
A1	Twice a day From 7 AM to 10 PM	What is your mood?	1. Very good 2. Fairly good 3. Fairly bad 4. Very bad

Table A.2: Signal contingent questions features

A.2.2.2 Interval contingent

A.2.2.3 Event contingent

A.2.2.4 Break options

A.2.3 Sensor selection

Code	Timing	Question	Answer options
A1	8:00 AM 12:00 PM 16:00 PM 18:00 PM	What are you eating?	1. Vegetables 2. Fruits 3. Meat 4. . . .

Table A.3: Interval contingent questions features

Code	Event	Question	Answer options
A1	When Activities="Moving"	What sport are you doing?	Free text

Table A.4: Event contingent questions features

Code	Break name	Timing	Additional remarks
A1	I will go to sleep	6 hours	The break option should be available only once a day

Table A.5: Break options features

No.	Sensor	Frequency (up to)	Select
1	Accelerometer	up to 20 times per second	X
2	Linear Acceleration	up to 20 times per second	
3	Gyroscope	up to 20 times per second	
4	Gravity	up to 20 times per second	
5	Rotation Vector	up to 20 times per second	
6	Magnetic Field	up to 20 times per second	
7	Orientation	up to 20 times per second	
8	Ambient Temperature	up to 20 times per second	
9	Pressure	up to 20 times per second	
10	Relative Humidity	up to 20 times per second	
11	Proximity	up to 20 times per second	
12	Location	Once every minute	
13	WIFI Network Connected to	On change	
14	WIFI Networks Available	Once every minute	
15	Bluetooth Devices	Once every minute	
16	Bluetooth LE (Low Energy) Devices	Once every minute	
17	Running Applications	Once every 5 seconds	
18	Screen Status [ON/OFF]	On change	
19	Airplane Mode [ON/OFF]	On change	
20	Battery Charge [ON/OFF]	On change	
21	Battery Level	On change	
22	Doze Mode [ON/OFF]	On change	
23	Headset Status [ON/OFF]	On change	
24	Ring mode [Silent/Normal]	On change	
25	Music Playback (no track information)	On change	
26	Notifications received	On change	
27	Touch event	On change	
28	Cellular network info	Once every minute	
29	Movement Activity	Once every 30 seconds	
30	Step Counter	up to 20 times per second	
31	Step Detection	On change	
32	Light	up to 20 times per second	

Table A.6: Sensor selection

B

Ethics and legal documentation

Contents

This appendix presents the support material for managing the ethical and legal aspects of a data collection. The material was defined starting from various sources, such as the templates and guidelines provided by the University of Trento [1], as well as the proposed documentation from the University of Cambridge [2] and from the UK Data Archive platform [3]. As regards the documentation for the distribution of data, its first draft was based on the procedures defined by EUROSTAT regarding the processing of microdata [4].

[1] https://www.unitn.it/en/ricerca/1755/research-ethics-committee

[2] https://www.research-integrity.admin.cam.ac.uk/research-ethics

[3] https://ukdataservice.ac.uk/learning-hub/research-data-management/ethical-issues/ethical-obligations/

[4] https://ec.europa.eu/eurostat/web/microdata/access

B.1 Request for IRB evaluation

Presentation of the project

Project title

Project manager Please, attach resume

Name and Surname	....
Role/position	....
University/Research center	....
Email address	....
Phone number	....

Other researchers involved Please, attach resume

Name and Surname	....
Role/position	....
University/Research center	....
Email address	....
Phone number	....

Does the project requires permission to other organizations (such as hospitals, schools, prisons) for accessing data or involve participants? If Yes, please attach copy of authorization letter

Do the manager and the members of the research group, as well as their families, have specific interests in relation to the outcome of the study? Please, attach signed statement

Does the research manager have enough time equipment, facilities and staff to conduct the research? Please, attach statement, countersigned by the responsible person of the structure to which it belongs

Are there any procedures that require specific professional skills (e.g. doctor, psychologist, nurse, etc.) in accordance with current legislation? If Yes, specify.

Project details

Funding entities or sponsors Please, specify their contributions

Beginning date of the research

Expected duration

Summary of the research program Please, enclose a schematic representation of the protocol

Project description

- Initial basis and theoretical justification
- Objectives
- Proposed investigation method
- Description of the procedure (attach a copy of the material used and protocol)
- References

Details of the participants

What types of subjects are taking part in the study?

- students
- adults (older than 18 years and capable of expressing their consent)
- children and teenagers under 18 years
- seniors (over the age of 65 years and able to express their consent)
- european subjects
- people with intellectual disability/mental, not able to give its consent
- other people whose ability to express consent may be compromised (please specify why)
- people with physical disability (specify what type)
- institutionalized subjects (i.e., prisoners, hospital patients, etc.)
- patients and/or customers reported by physicians, psychologists or other categories of professionals
- other people whose ability to express consent may be compromised (please specify why)
- you cannot determine the category of subjects (eg., administration via internet)

Is it possible that some of the subjects are in a position of dependence of the researcher or one of his collaborators, such that it may be assumed that the expression of consent to participate in the study is not completely free or free from all pressure (such as student/professor, doctor/patient, employee/employer)? If Yes, indicate how you intend to provide to minimize the chance that the subject will feel compelled to take part in the search

Participants selection Please, specify any inclusion and exclusion criteria

How will you spread information and invitations to participate in the research? Please, provide copies of flyers and/or letters to be sent

Risk and risk management

The research will involve

- the use of questionnaires (please attach a copy)
- structured or semi-structured interviews (attach a copy of the questions that will be asked; if this is not possible, indicate the topics to be covered)
- in depth interviews
- focus groups
- autobiographical narrations
- diary collection (diary keeping)
- observation of the behaviour of the subjects without their knowledge
- observation of subject behaviour
- audio or video recordings of subjects
- administration of stimuli, tasks or procedures and recording of behavioural responses, opinions or judgements
- administration of stimuli, tasks or procedures that the subject may find troublesome, stressful, physically or psychologically painful, both during and after the conduct of the study
- recording of ocular movements
- use of tms (transcranial magnetic stimulation)
- immersion in virtual reality environments
- recording of evoked potential
- administration of tests, questionnaires or experimental protocols via the internet (web, e-mail)
- use of neuropsychological tests
- neuroimaging techniques (e.g., fmri)
- the implementation of behaviors that could diminish the self-esteem of the subjects, or induce embarrassment, sorrow or depression
- procedures for deceiving the subjects
- administration of substances or agents (e.g., medicines, alcohol)
- collection of tissue samples or human fluids (e.g., blood tests)
- participation in clinical trials
- other (specify): [*e.g., Use of an application installed on their smartphone that can collect sensor data*]

Does the research involve the use of procedures which could be stressful or dangerous for participants? If yes, please describe the nature of the risks and the reasonably expected consequences of the procedures used.

Is there a specific insurance policy for additional liability to the university? If Yes, please attach the insurance contract in full copy.

How do you plan to deal with any complications or adverse reactions?

It is expected that there may be benefits to those taking part in the study? Which ones?

Information and consent

Please, attach a copy of informative and declaration of consent

How the participants will express their consent? If, for the realization of the study, it is not possible to inform participants about the objectives of the study prior to its start, please specify the modality in which this information will be later given

Is there a help desk for participants? If yes, please specify. Otherwise, please specify how participants will be able to obtain the necessary information about the study and the processing of their data

How will the participants be informed about the possibility of receiving, directly or indirectly, information related to their psycho-physical condition that became available during the research?

Anonymity and confidentiality of personal data

Please, add detailed information on how personal data will be processed.

How will participants be guaranteed anonymity E.g., use of identification codes

If is necessary to keep the participant identification data, please specify the reasons and how the participants will be informed

What security measures are in place to ensure that data confidentiality is respected?

Storage and security of collected data and research results

Who will have access to the collected data and the (intermediate or not) results of the research?

For how many years will the collected data be retained after the conclusion of the research?

Please, indicate if and how personal and/or sensitive data will be stored (who is responsible for proper storage and where they will be stored)

B.2 Informative

INFORMATION ABOUT THE STUDY

[*Insert here a short description of the study, considering:*

- *Study aim(s)*
- *Study topic(s)*

]

[*Insert here a short description of the tasks, considering:*

- *Tools used in the study*
- *Activities required of the participant to carry out the study, (e.g., answer questions about the use of time ("Where are you?"; "Who are you with?"; "What are you doing?") asked at regular intervals)*
- *Any additional information on the tasks, (e.g., it will be possible not to respond to notifications immediately, but to accumulate up to a maximum of 24. After that the app will start to delete them, starting from the least recent)*

]

[*Insert here a short description of the incentives, considering:*

- *Type of incentives and amount*
- *Requirements to receive compensation (e.g., The compensation will be given to all participants who have answered at least 85% of the notification and kept active the GPS and Bluetooth for at least 50% of the duration of the study)*

]

We specify that you always have the right not to answer some questions or to interrupt the compilation. The partially collected data will still be useful for the purpose of the investigation. The results of the survey will be disseminated in aggregate form and therefore it will not be possible to trace the subjects to which the data refer.

For information relating to the survey, you can contact the Scientific Director [...], at the following address: [*email*]

INFORMATION ON THE PROCESSING OF PERSONAL DATA

We wish to inform you that the EU Regulation 2016/679 "General Regulation on the protection of personal data" (from now on "GDPR"), the D.lgs. n. 196/2003 "Code regarding the protection of personal data" and the relative Annex A.4 "Ethical rules for the processing of personal data for statistical and scientific purposes (decision of the IDPA n. 515 of 19 December 2018, in the Official Journal n. 11 of 14 January 2019)".

Pursuant to the aforementioned legislation, the processing of your personal data by researchers involved in the research activity will be based on compliance with the principles set forth in art. 5 of the GDPR and, in particular, to those of legality, correctness, transparency, relevance, not excess and in order to guarantee an adequate security of personal data.

The Data Controller is the University of [...], [*Address*], [*email*]

The contact of the Data Protection Officer is: [*email*]

The processing of your personal data is carried out for the realization of the scientific purposes of the research project.

B.3 Privacy statement

INFORMATION ON THE PROCESSING OF PERSONAL DATA FOR SCIENTIFIC RESEARCH AIMS (ART. 13 REG. EU 2016/679)
Title of the research project (hereinafter "Project"): [...]

Dear participant,

We wish to inform you that the EU Regulation 2016/679 "General Regulation on the protection of personal data" (from now on "GDPR"), the D.lgs. n. 196/2003 "Code regarding the protection of personal data" and the relative Annex A.4 "Ethical rules for the processing of personal data for statistical and scientific purposes (decision of the IDPA n. 515 of 19 December 2018, in the Official Journal n. 11 of 14 January 2019)" Pursuant to the aforementioned legislation, the processing of your personal data by researchers involved in the research activity will be based on compliance with the principles set forth in art. 5 of the GDPR and, in particular, to those of legality, correctness, transparency, relevance, not excess and in order to guarantee an adequate security of personal data.

As a data subject, we provide you with the following information regarding the processing of your personal data.

Data Controller and Data Protection Officer

The Data Controller is the University of [...], [*Address*], [*email*]

The contact of the Data Protection Officer is: [*email*]

Purpose of the processing

The processing of your personal data is carried out for the realisation of the scientific purposes of the Project, namely [...] The Project was drafted in accordance with the methodological standards of the disciplinary sector concerned and is deposited with the University of [...], where it will be kept for [... *years*] from the planned conclusion of the research.

Legal basis of the processing

The processing of your personal data is carried out by the Data Controller in the execution of his tasks of public interest pursuant to art. 6, paragraph 1, lett. e) of the GDPR.

If the processing of special categories of personal data (sensitive data) is carried out, it will be done for the purpose of scientific research pursuant to art. 9, paragraph 2, lett. j) of the GDPR and (where appropriate) other legal grounds.

Category and type of personal data processed

The data collected and further processed are:

- Profile data [*Insert list, e.g., gender, age, nationality*]
- Intensive longitudinal survey data [*Insert list, e.g., daily questions on routines*]
- Sensor data [*Insert list, e.g., position, Wi-Fi network connections, running applications, screen status, flight mode, battery status, doze modality, headset, audio mode, music playback (no track info), notifications received, touch event, cellular network info, ...*]

At any time during the data collection, you can stop and start the registering of the sensor data from the sensors, if you wish to do so. Besides the data above, the email address will also be collected for the sole purpose of managing the data collection. The email address will not be stored jointly with the remaining data collected and will only be linked for the purposes.

Processing methods

Your data will be processed using the following software: [*Insert list, e.g., Python, Microsoft Excel, R*]

The following security measure will be followed: [*E.g.: The collected personal data and all information related to the abovementioned event are stored on the servers of the European Commission, or external contractors. The operations of which abide by the Commission's security decisions and provisions established by the Directorate of Security for this kind of servers and services, and for servers of external contractors abiding by the necessary security provisions. Access to the collected personal data and all information related to the above-mentioned event is only possible to the above-described populations through an authentication mechanism, such as the using asymmetric key pairs. Your personal data will be processed exclusively by the Owner and / or authorized parties in the framework of the implementation of the Project.*]

Personal data retention period

Your personal data will be kept until the project's goals are reached or a maximum time of one year after the official end of the project.

Nature of data provision

The provision of your data for the aforementioned research purposes is indispensable for the carrying out of the Project and does not derive from a regulatory and / or contractual obligation. Failure to provide data will only make it impossible to participate in the Project. Data recipients and possible transfer abroad None of your personal or sensor data will be published on the web or anywhere else. The collected data will solely serve the purpose of the Project. None of your personal or sensor-related data will be published on the web or anywhere else. No personal data will be transmitted to subjects unrelated to the recipients and the legal framework mentioned in this document.

[*The data provided, made anonymous, will be used for institutional purposes of teaching, research, and the transfer of scientific and technological knowledge (Third Mission). Furthermore, anonymous data may be transferred solely for the purposes of teaching, research and the transfer of scientific and technological knowledge to the categories of subjects mentioned above, even those based outside the EU, who have signed up, for example through a user licence, codes of conduct and agreements aimed at ensuring an adequate level of protection of personal data, or in any case subject to verification that the recipient guarantees adequate protection measures. Pursuant to art. 28 of the GDPR, the recipients of the data may be subjects or categories of subjects, such as, by way of example but not limited to, researchers and research groups belonging to both public and private national and international universities, bodies or research institutes; students, doctoral students and university professors for teaching, research and technology transfer purposes only within university courses and thesis work; research groups, including private ones, for the development of services and for the improvement of the quality of life; Public or private subjects for scientific research purposes and/or interested in developing services useful for improving the quality of life.*]

To find out at any time the subjects to whom your data will be communicated, simply request an updated list by writing to the email address: [*email*]

Any use for profiling for commercial purposes is prohibited.

Dissemination of research results

The dissemination of statistical and / or scientific results (for example through the publication of scientific articles and / or the creation of databases, even with open access methods, participation in conferences, etc.) may only take place anonymously and / or aggregated and in any case in a manner that does not make it identifiable.

Rights of the data subject

As a data subject you have the right to request at any time from the Data Controller the exercise of rights pursuant to articles 15 and ss. of the GDPR where applicable and, in particular, access to their personal data, rectification, integration, cancellation, limitation of the processing that concerns them or to oppose their treatment. According to the art. 17, paragraph 3, lett. d ("for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes in accordance with Article 89(1) in so far as the right referred to in paragraph 1 is likely to render impossible or seriously impair the achievement of the objectives of that processing") the right to cancellation does not exist for data whose processing is necessary for the purposes of scientific research if it risks making it impossible and / or seriously undermining the objectives of the research itself.

For the exercise of the aforementioned rights, you can contact the Data Controller and / or the Data Protection Officer at the aforementioned addresses.

Each data subject has also the right to an effective judicial remedy where he or she considers that his or her rights under this Regulation have been infringed as a result of the processing of his or her personal data in non-compliance with this Regulation.

For information relating to the Project, please contact the project's Scientific Coordinator at the following address: [*email*]

B.4 Data Protection Impact Assessment (DPIA)

	DPIA [*Insert title*]
Prepared by	
Approved by	
Date of approval	
Review frequency	
Next review date	

Foreword - Article 35 Regulation EU 679/2016 (GDPR) – DPIA

1. Where a type of processing in particular using new technologies, and taking into account the nature, scope, context and purposes of the processing, is likely to result in a high risk to the rights and freedoms of natural persons, the controller shall, prior to the processing, carry out an assessment of the impact of the envisaged processing operations on the protection of personal data. A single assessment may address a set of similar processing operations that present similar high risks.

2. The controller shall seek the advice of the data protection officer, where designated, when carrying out a data protection impact assessment.
3. A data protection impact assessment referred to in paragraph 1 shall in particular be required in the case of:
    (a) a systematic and extensive evaluation of personal aspects relating to natural persons which is based on automated processing, including profiling, and on which decisions are based that produce legal effects concerning the natural person or similarly significantly affect the natural person;
    (b) processing on a large scale of special categories of data referred to in Article 9(1), or of personal data relating to criminal convictions and offences referred to in Article 10; or
    (c) a systematic monitoring of a publicly accessible area on a large scale [5].
4. The supervisory authority shall establish and make public a list of the kind of processing operations which are subject to the requirement for a data protection impact assessment pursuant to paragraph 1. The supervisory authority shall communicate those lists to the Board referred to in Article 68.
5. The supervisory authority may also establish and make public a list of the kind of processing operations for which no data protection impact assessment is required. The supervisory authority shall communicate those lists to the Board.
6. Prior to the adoption of the lists referred to in paragraphs 4 and 5, the competent supervisory authority shall apply the consistency mechanism referred to in Article 63 where such lists involve processing activities which are related to the offering of goods or services to data subjects or to the monitoring of their behaviour in several Member States or may substantially affect the free movement of personal data within the Union.
7. The assessment shall contain (mandatory pursuant to Article :
    (a) a systematic description of the envisaged processing operations and the purposes of the processing, including, where applicable, the legitimate interest pursued by the controller;
    (b) an assessment of the necessity and proportionality of the processing operations in relation to the purposes;
    (c) an assessment of the risks to the rights and freedoms of data subjects referred to in paragraph 1; and
    (d) the measures envisaged to address the risks, including safeguards, security measures and mechanisms to ensure the protection of personal data and to demonstrate compliance with this Regulation taking into account the rights and legitimate interests of data subjects and other persons concerned.
8. Compliance with approved codes of conduct referred to in Article 40 by the relevant controllers or processors shall be taken into due account in

[5] Note that the controller shall be responsible and be able to demonstrate compliance with GDPR so it is advised to do a DPIA and monitor the tasks it refers to monitor its performance and the emergence of issues.

assessing the impact of the processing operations performed by such controllers or processors, in particular for the purposes of a data protection impact assessment.

9. Where appropriate, the controller shall seek the views of data subjects or their representatives on the intended processing, without prejudice to the protection of commercial or public interests or the security of processing operations.
10. Where processing pursuant to point (c) or (e) of Article 6(1) has a legal basis in Union law or in the law of the Member State to which the controller is subject, that law regulates the specific processing operation or set of operations in question, and a data protection impact assessment has already been carried out as part of a general impact assessment in the context of the adoption of that legal basis, paragraphs 1 to 7 shall not apply unless Member States deem it to be necessary to carry out such an assessment prior to processing activities.
11. Where necessary, the controller shall carry out a review to assess if processing is performed in accordance with the data protection impact assessment at least when there is a change of the risk represented by processing operations[6].

Part 1: Data protection impact assessment screening

These questions are intended to help you decide whether a DPIA is necessary. Answering 'yes' to any of these questions is an indication that a DPIA would be a useful exercise. You can expand on your answers as the project develops if you need to. You can adapt these questions to align more closely to project you are assessing [7].

1. Will the project involve the collection of new information about individuals?
2. Will the project compel individuals to provide information about themselves?
3. Will information about individuals be disclosed to organisations or people who have not previously had routine access to the information?
4. Are you using information about individuals for a purpose it is not currently used for, or in a way it is not currently used?
5. Does the project involve you using new technology that might be perceived as being privacy intrusive? For example, the use of biometrics or facial recognition.
6. Will the project result in you making decisions or taking action against individuals in ways that can have a significant impact on them?
7. Is the information about individuals of a kind particularly likely to raise privacy concerns or expectations? For example, health records, criminal

[6] See also WP248 http://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=611236

[7] Additional information can be found on the Italian Data Protection Authority: https://www.garanteprivacy.it/documents/1016/0/ALLEGATO+1+Elenco+delle+tipologie+di+trattamenti+soggetti+al+meccanismo+di+coerenza+da+sottoporre+a+valutazione+di+impatto.pdf/b9ceefa9-dd65-df86-fed4-df3c3570f59d?version=1.11)

The French Data Protection Authority provides a software - available in several languages: https://www.cnil.fr/fr/outil-pia-telechargez-et-installez-le-logiciel-de-la-cnil) records or other information that people would consider to be private.

8. Will the project require you to contact individuals in ways that they may find intrusive?

Part 2: Data protection impact assessment report

Use this report template to record the DPIA process and results. You can start to fill in details after the screening questions have identified the need for a DPIA. The template follows the process that is used in the WeNet project. You can adapt the template to allow you to record additional information relevant to the DPIA you are conducting

1. Step one (mandatory): Describe the project and identify the need for a DPIA Explain what the research project aims to achieve, what the benefits will be to WeNet, to individuals and to other parties. You may find it helpful to link to other relevant documents related to the project, for example a project proposal. It is important to include information about the benefits to be gained from the research project in order to help balance any risk identified in the DPIA. This can help inform decisions on the level of risk to privacy that is acceptable, when balanced against the benefits or other justification for the research project. Also summarize why the need for a DPIA was identified (this can draw on your answers to the screening questions) and identify the legal basis for processing. Include a systematic description of the purposes of the processing - 35.7(A) GDPR

2. Step two (mandatory): Describe the information flows You should describe the collection, use and deletion of personal data here. You should also say how many individuals are likely to be affected by the research project. Describe how the personal data will be processed. Provide information about the design and method. It is often helpful to include a diagram or flowchart that explains the information flows. Include a systematic description of the envisaged processing operations - 35.7(A) GDPR - and an assessment of the necessity and proportionality of the processing operations in relation to the purposes - 35.7(B) GDPR

3. Step three: Consultation requirements Describe the groups you will be consulting with and their interest in the research project. Who should be consulted internally and externally? Explain the method you will use for consultation with any stakeholder groups and how you will communicate the outcomes of the DPIA back to them. How will you carry out the consultation? Explain what you learned from the consultation process and how they shaped your approach to the management of privacy risks. Explain what practical steps you will take to ensure that you identify and address privacy risks. You should link this to the relevant stages of the WeNet operating procedures. You can use consultation at any stage of the DPIA process.

4. Step four (mandatory): Identify the privacy and related risks Identify the key privacy risks and the associated compliance and corporate risks (e.g. Data breach, Intruder access, Researchers are not adequately trained). Larger-scale DPIAs might record this information on a more formal risk register.

See also recital 75 GDPR: The risk to the rights and freedoms of natural persons, of varying likelihood and severity, may result from personal data processing which could lead to physical, material or non-material damage, in particular: where the processing may give rise to discrimination, identity theft or fraud, financial loss, damage to the reputation, loss of confidentiality of personal data protected by professional secrecy, unauthorised reversal of pseudonymisation, or any other significant economic or social disadvantage; where data subjects might be deprived of their rights and freedoms or prevented from exercising control over their personal data; where personal data are processed which reveal racial or ethnic origin, political opinions, religion or philosophical beliefs, trade union membership, and the processing of genetic data, data concerning health or data concerning sex life or criminal convictions and offences or related security measures; where personal aspects are evaluated, in particular analysing or predicting aspects concerning performance at work, economic situation, health, personal preferences or interests, reliability or behaviour, location or movements, in order to create or use personal profiles; where personal data of vulnerable natural persons, in particular of children, are processed; or where processing involves a large amount of personal data and affects a large number of data subjects.

The questions under Part 3 can be used to help you identify the Data Protection Local law and the General Data Protection Regulation (GDPR) related compliance risks.

	Privacy issue	**Risk for individuals**	**Compliance risk**	**Associated organisation risk**
1				
2				
3				

5. Step five(mandatory): Identify privacy solutions Describe the actions you could take to reduce the risks, and any future steps which would be necessary (e.g. the production of new guidance or future security testing for systems). The following are notes on the columns of the table below:

Likelihood and severity of the risk (Recital 76 GDPR): The likelihood and severity of the risk to the rights and freedoms of the data subject should be determined by reference to the nature, scope, context and purposes of the processing. Risk should be evaluated on the basis of an objective assessment, by which it is established whether data processing operations involve a risk or a high risk.

Result: is the risk eliminated, reduced, or accepted?

Evaluation: is the final impact on individuals after implementing each solution a justified, compliant and proportionate response to the aims of the research project?

	Risk	Likelihood	Severity	Solution(s)	Result	Evaluation
1						
2						
3						

6. Step six(mandatory): Sign off and record the DPIA outcomes Who has approved the privacy risks involved in the research project? What solutions need to be implemented?

Risk	Approved solution	Approved by
The key underlying risks are:	Key solutions are:	

7. Step seven(mandatory): Integrate the DPIA outcomes back into the project plan Who is responsible for integrating the DPIA outcomes back into the WeNet project plan and updating any project management paperwork? Who is responsible for implementing the solutions that have been approved? Who is the contact for any privacy concerns that may arise in the future?

Action	Due date	Responsibility for action

Part 3: Linking the DPIA to the data protection principles

Answering these questions during the DPIA process will help you to identify where there is a risk that the research project will fail to comply with the GDPR or other relevant local legislation.

1. GDPR Principle 1 (Article 5(1)(a)) Personal data shall be processed fairly and lawfully and, in particular, shall not be processed unless (i) at least one of the conditions in DPA Schedule 2 and GDPR Article 6 is met, and (ii) in the case of sensitive personal data, at least one of the conditions in DPA Schedule 3 and GDPR Article 9 is also met.

1. Have you identified the purpose of the project?
2. How will you tell individuals about the use of their personal data?
3. Do you need to amend your privacy notices?
4. Have you established which conditions for processing apply?
5. If you are relying on consent to process personal data, how will this be collected and what will you do if it is withheld or withdrawn?

2. GDPR Principle 2 (Article 5(1)(b)) Personal data shall be obtained only for one or more specified and lawful purposes, and shall not be further processed in any manner incompatible with that purpose or those purposes.

1. Does your project plan cover all of the purposes for processing personal data?
2. Have you identified potential new purposes as the scope of the project expands?

3. GDPR Principle 3 (Article 5(1)(c)) Personal data shall be adequate, relevant and not excessive in relation to the purpose or purposes for which they are processed.

1. Is the quality of the information good enough for the purposes it is used?
2. Which personal data could you not use, without compromising the needs of the project?

4. GDPR Principle 4 (Article 5(1)(d))–accurate, kept up to date, deletion Personal data shall be accurate and, where necessary, kept up to date.

1. If you are procuring new software does it allow you to amend data when necessary?
2. How are you ensuring that personal data obtained from individuals or other organisations is accurate?

5. GDPR Principle 5 (Article 5(1)(e)) Personal data processed for any purpose or purposes shall not be kept for longer than necessary for that purpose or those purposes.

1. What retention periods are suitable for the personal data you will be processing?
2. Are you procuring software that will allow you to delete information in line with your retention periods?

6. GDPR Articles 12-22 Personal data shall be processed in accordance with the rights of data subjects.

1. Will the systems you are putting in place allow you to respond to subject access requests more easily?
2. If the project involves marketing, have you got a procedure for individuals to opt out of their information being used for that purpose?

7. GDPR Principle 6 (Article 5 (1)(f)) Appropriate technical and organisational measures shall be taken against unauthorised or unlawful processing of personal data and against accidental loss or destruction of, or damage to, personal data.

1. Do any new systems provide protection against the security risks you have identified?

2. What training and instructions are necessary to ensure that staff know how to operate a new system securely?

8. GDPR Article 24 Personal data shall not be transferred to a country or territory outside the European Economic Area unless that country of territory ensures an adequate level of protection for the rights and freedoms of data subjects in relation to the processing of personal data.

1. Will the project require you to transfer data outside of the European Economic Area (EEA)?
2. If you will be making transfers, how will you ensure that the data is adequately protected?

B.5 Data Processor Agreement

SECTION I

Clause 1

Purpose and scope

1. The purpose of these Standard Contractual Clauses (the Clauses) is to ensure compliance with Article 28(3) and (4) of Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data.
2. The controllers and processors listed in Annex I have agreed to these Clauses in order to ensure compliance with Article 28(3) and (4) of Regulation (EU) 2016/679.
3. These Clauses apply to the processing of personal data as specified in Annex II.
4. Annexes I to IV are an integral part of the Clauses.
5. These Clauses are without prejudice to obligations to which the controller is subject by virtue of Regulation (EU) 2016/679.
6. These Clauses do not by themselves ensure compliance with obligations related to international transfers in accordance with Chapter V of Regulation (EU) 2016/679.

Clause 2

Invariability of the Clauses

1. The Parties undertake not to modify the Clauses, except for adding information to the Annexes or updating information in them.
2. This does not prevent the Parties from including the standard contractual clauses laid down in these Clauses in a broader contract, or from adding other clauses or additional safeguards provided that they do not directly or

indirectly contradict the Clauses or detract from the fundamental rights or freedoms of data subjects.

Clause 3

Interpretation

1. Where these Clauses use the terms defined in Regulation (EU) 2016/679 respectively, those terms shall have the same meaning as in that Regulation.
2. These Clauses shall be read and interpreted in the light of the provisions of Regulation (EU) 2016/679 respectively.
3. These Clauses shall not be interpreted in a way that runs counter to the rights and obligations provided for in Regulation (EU) 2016/679 or in a way that prejudices the fundamental rights or freedoms of the data subjects.

Clause 4

Hierarchy In the event of a contradiction between these Clauses and the provisions of related agreements between the Parties existing at the time when these Clauses are agreed or entered into thereafter, these Clauses shall prevail.

Clause 5 - Optional

Docking clause

1. Any entity that is not a Party to these Clauses may, with the agreement of all the Parties, accede to these Clauses at any time as a controller or a processor by completing the Annexes and signing Annex I.
2. Once the Annexes in (a) are completed and signed, the acceding entity shall be treated as a Party to these Clauses and have the rights and obligations of a controller or a processor, in accordance with its designation in Annex I.
3. The acceding entity shall have no rights or obligations resulting from these Clauses from the period prior to becoming a Party.

SECTION II – OBLIGATIONS OF THE PARTIES

Clause 5

Description of processing(s) The details of the processing operations, in particular the categories of personal data and the purposes of processing for which the personal data is processed on behalf of the controller, are specified in Annex II.

Clause 6

Obligations of the Parties

Instructions

1. The processor shall process personal data only on documented instructions from the controller, unless required to do so by Union or Member State law to which the processor is subject. In this case, the processor shall inform the controller of that legal requirement before processing, unless the law prohibits this on important grounds of public interest. Subsequent instructions may also be given by the controller throughout the duration of the processing of personal data. These instructions shall always be documented.
2. The processor shall immediately inform the controller if, in the processor's opinion, instructions given by the controller infringe Regulation (EU) 2016/679 or the applicable Union or Member State data protection provisions.

Purpose limitation The processor shall process the personal data only for the specific purpose(s) of the processing, as set out in Annex II, unless it receives further instructions from the controller.

Duration of the processing of personal data Processing by the processor shall only take place for the duration specified in Annex II.

Security of processing

1. The processor shall at least implement the technical and organisational measures specified in Annex III to ensure the security of the personal data. This includes protecting the data against a breach of security leading to accidental or unlawful destruction, loss, alteration, unauthorised disclosure or access to the data (personal data breach). In assessing the appropriate level of security, the Parties shall take due account of the state of the art, the costs of implementation, the nature, scope, context and purposes of processing and the risks involved for the data subjects.
2. The processor shall grant access to the personal data undergoing processing to members of its personnel only to the extent strictly necessary for implementing, managing and monitoring of the contract. The processor shall ensure that persons authorised to process the personal data received have committed themselves to confidentiality or are under an appropriate statutory obligation of confidentiality.

Sensitive data If the processing involves personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, genetic data or biometric data for the purpose of uniquely identifying a natural person, data concerning health or a person's sex life or sexual orientation, or data relating to criminal convictions and offences ("sensitive data"), the processor shall apply specific restrictions and/or additional safeguards.

Documentation and compliance

1. The Parties shall be able to demonstrate compliance with these Clauses.

2. The processor shall deal promptly and adequately with inquiries from the controller about the processing of data in accordance with these Clauses.
3. The processor shall make available to the controller all information necessary to demonstrate compliance with the obligations that are set out in these Clauses and stem directly from Regulation (EU) 2016/679. At the controller's request, the processor shall also permit and contribute to audits of the processing activities covered by these Clauses, at reasonable intervals or if there are indications of non-compliance. In deciding on a review or an audit, the controller may take into account relevant certifications held by the processor.
4. The controller may choose to conduct the audit by itself or mandate an independent auditor. Audits may also include inspections at the premises or physical facilities of the processor and shall, where appropriate, be carried out with reasonable notice.
5. The Parties shall make the information referred to in this Clause, including the results of any audits, available to the competent supervisory authority/ies on request.

Use of sub-processors

1. PRIOR SPECIFIC AUTHORISATION: The processor shall not subcontract any of its processing operations performed on behalf of the controller in accordance with these Clauses to a sub-processor, without the controller's prior specific written authorisation. The processor shall submit the request for specific authorisation at least [SPECIFY TIME PERIOD] prior to the engagement of the sub-processor in question, together with the information necessary to enable the controller to decide on the authorisation. The list of sub-processors authorised by the controller can be found in Annex IV. The Parties shall keep Annex IV up to date.
2. Where the processor engages a sub-processor for carrying out specific processing activities (on behalf of the controller), it shall do so by way of a contract which imposes on the sub-processor, in substance, the same data protection obligations as the ones imposed on the data processor in accordance with these Clauses. The processor shall ensure that the sub-processor complies with the obligations to which the processor is subject pursuant to these Clauses and to Regulation (EU) 2016/679.
3. At the controller's request, the processor shall provide a copy of such a sub-processor agreement and any subsequent amendments to the controller. To the extent necessary to protect business secret or other confidential information, including personal data, the processor may redact the text of the agreement prior to sharing the copy.
4. The processor shall remain fully responsible to the controller for the performance of the sub-processor's obligations in accordance with its contract with the processor. The processor shall notify the controller of any failure by the sub-processor to fulfil its contractual obligations.
5. The processor shall agree a third party beneficiary clause with the sub-processor whereby - in the event the processor has factually disappeared,

ceased to exist in law or has become insolvent- the controller shall have the right to terminate the sub-processor contract and to instruct the sub-processor to erase or return the personal data.

International transfers

1. Any transfer of [personal] data to a third country or an international organisation by the processor shall be done only on the basis of documented instructions from the controller or in order to fulfil a specific requirement under Union or Member State law to which the processor is subject and shall take place in compliance with Chapter V of Regulation (EU) 2016/679.
2. The controller agrees that where the processor engages a sub-processor in accordance with Clause 7.7. for carrying out specific processing activities (on behalf of the controller) and those processing activities involve a transfer of personal data within the meaning of Chapter V of Regulation (EU) 2016/679, the processor and the sub-processor can ensure compliance with Chapter V of Regulation (EU) 2016/679 by using standard contractual clauses adopted by the Commission in accordance with of Article 46(2) of Regulation (EU) 2016/679, provided the conditions for the use of those standard contractual clauses are met.

Clause 7

Assistance to the controller

1. The processor shall promptly notify the controller of any request it has received from the data subject. It shall not respond to the request itself, unless authorised to do so by the controller.
2. The processor shall assist the controller in fulfilling its obligations to respond to data subjects' requests to exercise their rights, taking into account the nature of the processing. In fulfilling its obligations in accordance with (a) and (b), the processor shall comply with the controller's instructions
3. In addition to the processor's obligation to assist the controller pursuant to Clause 8(b), the processor shall furthermore assist the controller in ensuring compliance with the following obligations, taking into account the nature of the data processing and the information available to the processor:
   (a) the obligation to carry out an assessment of the impact of the envisaged processing operations on the protection of personal data (a 'data protection impact assessment') where a type of processing is likely to result in a high risk to the rights and freedoms of natural persons;
   (b) the obligation to consult the competent supervisory authority/ies prior to processing where a data protection impact assessment indicates that the processing would result in a high risk in the absence of measures taken by the controller to mitigate the risk;
   (c) the obligation to ensure that personal data is accurate and up to date, by informing the controller without delay if the processor becomes aware

that the personal data it is processing is inaccurate or has become outdated;
    (d) the obligations in Article 32 Regulation (EU) 2016/679.
4. The Parties shall set out in Annex III the appropriate technical and organisational measures by which the processor is required to assist the controller in the application of this Clause as well as the scope and the extent of the assistance required.

Clause 8

Notification of personal data breach In the event of a personal data breach, the processor shall cooperate with and assist the controller for the controller to comply with its obligations under Articles 33 and 34 Regulation (EU) 2016/679, where applicable, taking into account the nature of processing and the information available to the processor.

Data breach concerning data processed by the controller In the event of a personal data breach concerning data processed by the controller, the processor shall assist the controller:

1. in notifying the personal data breach to the competent supervisory authority/ies, without undue delay after the controller has become aware of it, where relevant/(unless the personal data breach is unlikely to result in a risk to the rights and freedoms of natural persons);
2. in obtaining the following information which, pursuant to Article 33(3) Regulation (EU) 2016/679, shall be stated in the controller's notification, and must at least include:
    (a) the nature of the personal data including where possible, the categories and approximate number of data subjects concerned and the categories and approximate number of personal data records concerned;
    (b) the likely consequences of the personal data breach;
    (c) the measures taken or proposed to be taken by the controller to address the personal data breach, including, where appropriate, measures to mitigate its possible adverse effects.
    (d) Where, and insofar as, it is not possible to provide all this information at the same time, the initial notification shall contain the information then available and further information shall, as it becomes available, subsequently be provided without undue delay.
3. in complying, pursuant to Article 34 Regulation (EU) 2016/679, with the obligation to communicate without undue delay the personal data breach to the data subject, when the personal data breach is likely to result in a high risk to the rights and freedoms of natural persons.

Data breach concerning data processed by the processor In the event of a personal data breach concerning data processed by the processor, the processor shall notify the controller without undue delay after the processor having become aware of the breach. Such notification shall contain, at least:

1. a description of the nature of the breach (including, where possible, the categories and approximate number of data subjects and data records concerned);
2. the details of a contact point where more information concerning the personal data breach can be obtained;
3. its likely consequences and the measures taken or proposed to be taken to address the breach, including to mitigate its possible adverse effects.

Where, and insofar as, it is not possible to provide all this information at the same time, the initial notification shall contain the information then available and further information shall, as it becomes available, subsequently be provided without undue delay. The Parties shall set out in Annex III all other elements to be provided by the processor when assisting the controller in the compliance with the controller's obligations under Articles 33 and 34 of Regulation (EU) 2016/679.

SECTION III – FINAL PROVISIONS

Clause 9

Non-compliance with the Clauses and termination

1. Without prejudice to any provisions of Regulation (EU) 2016/679, in the event that the processor is in breach of its obligations under these Clauses, the controller may instruct the processor to suspend the processing of personal data until the latter complies with these Clauses or the contract is terminated. The processor shall promptly inform the controller in case it is unable to comply with these Clauses, for whatever reason.
2. The controller shall be entitled to terminate the contract insofar as it concerns processing of personal data in accordance with these Clauses if:
   (a) the processing of personal data by the processor has been suspended by the controller pursuant to point (a) and if compliance with these Clauses is not restored within a reasonable time and in any event within one month following suspension;
   (b) the processor is in substantial or persistent breach of these Clauses or its obligations under Regulation (EU) 2016/679;
   (c) the processor fails to comply with a binding decision of a competent court or the competent supervisory authority/ies regarding its obligations pursuant to these Clauses or to Regulation (EU) 2016/679.
3. The processor shall be entitled to terminate the contract insofar as it concerns processing of personal data under these Clauses where, after having informed the controller that its instructions infringe applicable legal requirements in accordance with Clause 7.1 (b), the controller insists on compliance with the instructions.
4. Following termination of the contract, the processor shall, at the choice of the controller, delete all personal data processed on behalf of the controller and

certify to the controller that it has done so, or, return all the personal data to the controller and delete existing copies unless Union or Member State law requires storage of the personal data. Until the data is deleted or returned, the processor shall continue to ensure compliance with these Clauses.

ANNEX I LIST OF PARTIES

Controller(s): [Identity and contact details of the controller(s), and, where applicable, of the controller's data protection officer]

1. Name: ...

Address: ...

Contact person's name, position and contact details: ...

Signature and accession date: ...

2.

...

[Identity and contact details of the controller's data protection officer]

Processor(s): [Identity and contact details of the processor(s) and, where applicable, of the processor's data protection officer]

1. Name: ...

Address: ...

Contact person's name, position and contact details: ...

Signature and accession date: ...

2.

...

[Identity and contact details of the processor's data protection officer]

ANNEX II: DESCRIPTION OF THE PROCESSING

Categories of data subjects whose personal data is processed.............................

Categories of personal data processed...........................

Sensitive data processed (if applicable) and applied restrictions or safeguards that fully take into consideration the nature of the data and the risks involved, such as for instance strict purpose limitation, access restrictions (including access only for staff having followed

specialised training), keeping a record of access to the data, restrictions for onward transfers or additional security measures...........................

Nature of the processing............................

Purpose(s) for which the personal data is processed on behalf of the controller.......................

Duration of the processing............................

....

For processing by (sub-) processors, also specify subject matter, nature and duration of the processing

ANNEX III TECHNICAL AND ORGANISATIONAL MEASURES INCLUDING TECHNICAL AND ORGANISATIONAL MEASURES TO ENSURE THE SECURITY OF THE DATA EXPLANATORY NOTE:

The technical and organisational measures need to be described concretely and not in a generic manner.

Description of the technical and organisational security measures implemented by the processor(s) (including any relevant certifications) to ensure an appropriate level of security, taking into account the nature, scope, context and purpose of the processing, as well as the risks for the rights and freedoms of natural persons. Examples of possible measures:

Measures of pseudonymisation and encryption of personal data

Measures for ensuring ongoing confidentiality, integrity, availability and resilience of processing systems and services

Measures for ensuring the ability to restore the availability and access to personal data in a timely manner in the event of a physical or technical incident

Processes for regularly testing, assessing and evaluating the effectiveness of technical and organisational measures in order to ensure the security of the processing

Measures for user identification and authorisation

Measures for the protection of data during transmission

Measures for the protection of data during storage

Measures for ensuring physical security of locations at which personal data are processed

Measures for ensuring events logging

Measures for ensuring system configuration, including default configuration

Measures for internal IT and IT security governance and management

Measures for certification/assurance of processes and products

Measures for ensuring data minimisation

Measures for ensuring data quality

Measures for ensuring limited data retention

Measures for ensuring accountability

Measures for allowing data portability and ensuring erasure]

For transfers to (sub-) processors, also describe the specific technical and organisational measures to be taken by the (sub-) processor to be able to provide assistance to the controller Description of the specific technical and organisational measures to be taken by the processor to be able to provide assistance to the controller.

ANNEX IV: LIST OF SUB-PROCESSORS

EXPLANATORY NOTE: This Annex needs to be completed [and kept up to date] in case of specific authorisation of sub-processors (Clause 7.7(a), Option 1). The controller has authorised the use of the following sub-processors:

1. Name: ...

Address: ...

Contact person's name, position and contact details: ...

Description of the processing (including a clear delimitation of responsibilities in case several sub-processors are authorised): ...

2. ...

B.6 Data download

Data distribution process instructions

This document concerns the data owned by KnowDive or KnowDive members (from here on KnowDive Data). Specifically, the KnowDive Data are of two types:

1. data collected by asking the user, e.g., via questionnaires, as it is usually the case in, e.g., market studies, social studies
2. data collected from sensors of devices related to the user, most typically, smart phones and smart watches

KnowDive Data can be data fully anonymous and therefore out of General Data Protection Regulation (GDPR). The KnowDive Data will be shared according to the process described below. For what concerns anonymous data, the recital 26 of the GDPR claims that:

"[...] The principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable."

Since there is a possibility, at least in principle, to trace back the person generating the data, two main guidelines are followed:

1. The anonymization process we follow is very strict and precise.
2. The access to KnowDive data is restricted and limited by licensing and a successful submission of a proposal.

Data Requester's Roles

Duly designated representative of the entity: She is someone with the authority to make commitments on behalf of the organization. She:

1. Signs the application form for the research entity.
2. Signs a confidentiality undertaking and initials the terms of use.

Contact person in the research entity:

1. Coordinates submission of research proposals at the level of the entity.
2. Countersigns each research proposal submitted by researchers linked to the entity; the contact person confirms by his/her signature that all persons named in the research proposal are employed by, or are formally related to (e.g., PhD students), the research entity.
3. Informs researchers named in the research proposal about the obligations laid down in the terms of use of data.
4. In a network project, confirms participation of individual researchers from the entity, if another research entity is coordinator.
5. Is identified in the application form for the research entity and confidentiality undertaking.

Principal researcher:

1. submits and signs the research proposal and the individual confidentiality declaration.
2. identifies individual researchers participating in the research project.
3. is granted access to the secure platform with KnowDive Data.
4. is responsible for the lawful access to KnowDive Data for all researchers named in the research proposal.
5. protects KnowDive Data in accordance with the conditions specified in the relevant documents (confidentiality undertaking and terms of use, and individual confidentiality declaration).
6. informs KnowDive of any changes to the research proposal.
7. follows the guidelines for publication.
8. at the end of the project:
    - provides KnowDive with a copy of all reports, which have been produced using the data.
    - destroys received KnowDive Data and derived files after expiration/completion of the research project.

Data manager indicated in the research proposal (if different from principal researcher):

1. is granted access to the secure platform with KnowDive Data.
2. is responsible for the practical access to KnowDive Data for all researchers named in the research proposal.
3. protects KnowDive Data in accordance with the conditions specified in the relevant documents (confidentiality undertaking and terms of use and individual confidentiality declaration).
4. destroys received KnowDive Data and derived files after expiration/completion of the research project.

Individual researcher(s) named in the research proposal:

1. signs individual confidentiality declarations (each separately).
2. protects KnowDive Data in accordance with the conditions specified in the relevant documents (confidentiality undertaking and terms of use and individual confidentiality declaration); follows the guidelines for publication attached to the data.

All persons must immediately inform KnowDive about any breach of the confidentiality rules laid down in the confidentiality undertaking, terms of use and individual confidentiality declaration.

The data distribution process

The data distribution process is divided into 5 parts, namely:

1. Requesting access to KnowDive Data (submitting a research proposal)
2. Validation of research proposal
3. Granting access to KnowDive Data
4. Changes to research proposal
5. Closing of project

In a nutshell, the output of the distribution process, depending on the access type(s), will be:

1. sending to the principal researcher files containing anonymized data.
2. in the case of non-anonymous data, KnowDive acts as a contact point between the requester of the data and the research institution that generated it.

In both cases, in this document we refer to KnowDive as the owner and distributor of the dataset(s).

Access to KnowDive Data is only valid for the period specified in the research proposal. At the end of that period, the researcher must: (a) destroy any original scientific-use files sent by KnowDive and any confidential data derived from the files, and (b) send KnowDive the research results.

Requesting access to KnowDive Data (submitting a research proposal)

The organization can request access to KnowDive Data by submitting one or more research proposals. The researchers named in the proposal should be:

1. an employee of the research entity (or be working for them as a contractor, only in justified cases), or
2. senior (Ph.D.) students under guidance of a supervisor employed by the research entity; supervisor must be identified in the research proposal as a principal researcher and a senior student as an individual researcher.

Once the organization has been recognized as a research entity, its name will be included in the list of recognized entities on the KnowDive website.

Drafting of research project description

In case of anonymized data, in addition to removing direct identifiers from the records, some variables are further anonymized, i.e., grouped together, aggregated etc. This sometimes limits the usage of KnowDive Data. The individual KnowDive Data set documentation on KnowDive website provide for this crucial information about data preparation. Please consider it while drafting the research project description.

Check before you apply for KnowDive Data

Before applying for KnowDive Data, the organization should check if fulfils all pre-conditions.

- □ It has the name of the contact person in the research entity at hand
- □ They have checked that all researchers who will have access to the data are employed by or linked to their research entity
- □ They have read the description of anonymization methods applied on KnowDive Data and the KnowDive data documentation
- □ They have the credential to submit the proposal to KnowDive

Eligibility

KnowDive grants access to KnowDive Data only to recognized research entities. To qualify for recognition, an organization must:

1. Have research as one of its main activities (e.g., universities, research institutions) or be a research department within other organization (e.g., within bank, statistical institute, etc.).
2. Provide evidence of publication of research results.
3. Be independent and autonomous in formulating scientific conclusions.
4. Have adequate data security safeguards.

Responsibilities

Before being recognized as a research entity, an organization must fill in the research proposal form and sign a confidentiality undertaking with the terms of use. The confidentiality undertaking commits the signatory and all researchers having access to confidential data to:

1. accessing confidential data only for the agreed purposes.
2. guaranteeing the physical and logical security of the data, including prevention, and acting in case of violation of confidentiality.
3. respect copyright by citing datasets with appropriate references and following the guidelines for publication.

Validation of research proposal

The standard consultation period is 4 weeks.

Granting access to KnowDive Data

Once the research proposal is approved, depending on the access type(s) of your choice, we will grant you access to anonymized data (see Instructions for data download).

Changes to research proposal

Once a research proposal has been accepted, new researchers can be added (provided they sign a confidentiality declaration), and the duration of the project can be extended.

1. Form to ask staff changes
2. Form to ask project extension

The adding of a new partner organizations or a new dataset always requires a consultation with KnowDive.

1. Form to add partner organization
2. General form for changes (for changes not covered by dedicated forms – e.g., adding datasets, changing the scope of the research)

Closing of project

Access to KnowDive Data is only valid for the period specified in the research proposal. At the end of that period, you must:

1. destroy any original files sent by KnowDive and any confidential data derived from the files (and sign the relevant form: Declaration of data destruction), and
2. send KnowDive the references to your research results.

In deciding on the project end-date consider the length of the process of publication of research results which might require access to data.

Research proposal template

This application form is intended for entities wishing to be recognized as research entities. As a first step, please complete and send this form electronically (in PDF) to datadistribution@knowdive.it. Please do not sign the form at this stage. The information provided in the application form will be examined by KnowDive, which will take the decision on whether to grant 'research entity' status. The following criteria will be considered when deciding on the status of the entity:

1. the purpose of the entity.
2. the established record or reputation of the entity as a body producing quality research and making it publicly available.
3. the internal organizational arrangements for research, including, where relevant, the fact that the research entity is independent, autonomous in formulating scientific conclusions and separated from policy areas of the body to which it belongs.
4. The technical and scientific soundness of the proposal.
5. the safeguards in place to ensure security of the data.

Applicants will be notified by email of the outcome of the assessment.

General information

Official full name of the entity: ....

Short name or acronym: ....

Postal address: ....

Website: ....

Country: ....

Legal status: ....

University or higher education establishment

- □ Research organization
- □ Governmental organization
- □ International organization
- □ Public commercial organization
- □ Private commercial organization, including consultancy. Please indicate the type of organization (e.g., limited company, partnership, private enterprise): ....
- □ European Economic Interest Grouping
- □ Private organization, nonprofit
- □ Other, please specify: ....

Duly designated representative of the research entity

Name: ....

Position: ....

Email: ....

Country: ....

Identification of the researchers (and data manager) who will have access to the data

Principal researcher

Name: ....

Position: ....

E-mail: ....

Official full name of the research entity: ....

Website: ....

Data manager - the person to whom confidential data will be sent - if different from principal researcher

In case of network project: data managers in the other organizations must be marked in the field “Position” under heading “Individual researchers”.

Name: ....

Position: ....

E-mail: ....

Official full name of the research entity: ....

Website: ....

Individual researchers

Individual researcher (1)

Name: ....

Position: ....

E-mail: ....

Official full name of the research entity: ....

Website: ....

Individual researcher (2)

Name: ....

Position: ....

E-mail: ....

Official full name of the research entity: ....

Website: ....

Please copy the rows if more researchers involved in the project.

Purpose of the research proposal

Title of the research proposal

Please describe the research project(s) for which data access is requested, objectives of the research project(s) and provide details on the underlying contract if the research project is commissioned by another body; maximum 2 pages and 15 references.

Please state the duration for which data access is requested (maximum five years), please respect the format: dd/mm/yyyy. Start date must be at least in 6 weeks

Datasets to be used

Please indicate the KnowDive dataset(s) to be used

Please state how the above-mentioned dataset will be used. In case of access to several datasets, please state which data will be used for which part of the research project.

Please state the methods of analysis to be used

Results of the analysis

1. Please describe the expected outcomes of the analysis of the data.
2. Please describe how the results of the research will be published or otherwise disseminated: through which channels (printed publications, online publications, conferences, web, etc.)

Safekeeping of KnowDive Data

1. Please describe how data and intermediate results will be securely stored in the premises of the research entity (see: Safekeeping of the data part of the Terms of use). In case of a network project, all entities listed under item 1 need to provide this information.
2. Please describe how the anonymity of the statistical units will be ensured in results of your research. Please describe how you will ensure the anonymity of the statistical units in published results of your research. The Guidelines for publication set minimum thresholds for cell size and other applicable rules to be respected. To know more about anonymization methods, see the Catalog documentation.

I hereby certify that the information contained in this questionnaire is complete, accurate and correct and that any future change will be reported immediately to KnowDive. I understand that KnowDive is authorized to check at any time the accuracy of the information given in this questionnaire. I understand that KnowDive may also request more information, if necessary. I confirm that I submit this request to be granted access to KnowDive data. The decision of KnowDive may or may not authorize me to be granted access to KnowDive Data. In addition, I commit myself to take and maintain all necessary measures in compliance with the requirements stated in the confidentiality declaration and according to the "Regole deontologiche per trattamenti a fini statistici o di ricerca scientifica" (https://www.garanteprivacy.it/web/guest/home/docweb/ /docweb-display/docweb/9069637) where applicable. Furthermore, I commit myself to comply with the European Charter and Code of Research""

Principal researcher

Place and Date: ....

Signature: ....

Contact person in the research entity

Place and Date: ....

Signature: ....

Confidentiality undertaking and terms of use

TERMS OF USE OF THE REQUESTED DATASET FOR SCIENTIFIC-USE

This Agreement (hereinafter referred to as the "Agreement") is executed by and between:

1. On the one part:

KnowDive [8], hereby duly represented by [. . .] (hereinafter referred to as "KnowDive").

2. On the other part:

FULL COMPANY NAME ...., [VAT NUMBER.], a legal entity under the laws of [COUNTRY............], having its registered office at [FULL ADDRESS............] (hereinafter referred to as the "Research entity").

Hereinafter individually referred to as the "Party" and jointly as the "Parties".

Whereas:

- The Research proposal submitted by the Research entity was deemed worthy of approval by KnowDive;
- The research proposal concerns the following dataset(s): .... with regard to the proposal n. ....;
- The dataset requested by the Research entity is protected by copyright, pursuant to the relevant legislation in force, and in particular to the Italian Law 633/1941 and Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019, transposed and implemented by the Italian Legislative Decree no. 177 of 8 November 2021 (G.U. No. 283 of 27.11.2021);
- The copyright holders are respectively KnowDive, which has prepared the requested dataset in order to make it suitable for the specific research activities indicated by the Research entity in the Section B.6 and .... which has carried out the data collection, with regard to the collected data;
- KnowDive may also act as a licensee, with wide powers to license;
- The requested and prepared dataset is considered anonymous, also according to the Opinion 05/2014 on Anonymisation Techniques - WP216 adopted on 10 April 2014 by Article 29 Data Protection Working Party, and therefore outside the scope of EU Regulation 2016/679 (GDPR);
- Indeed, as indicated in recital 26 of the above-mentioned Regulation: "The principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable. This Regulation does not therefore concern the processing of such anonymous information, including for statistical or research purposes";

[8] The institution responsible for the procedure and the delegated person will be defined during the incubation period.

- The Parties are willing to define the terms and conditions of the use of the requested dataset.

NOW, THEREFORE, the Parties agree as follows:

1. These Terms of Use concern the use of the requested dataset by the Research entity.
2. KnowDive shall provide the Research entity free-charge access to the requested dataset.
3. The Research entity undertakes to respect the approved research project, the following provisions, as well as all its own tasks arising from the summary of the document regarding the "Data Distribution Process" (Section B.6), which are considered an integral part hereof.
4. The Research entity may not use the requested dataset for any unlawful purpose or for purposes not set in the approved research project or otherwise not authorized by KnowDive.
5. The Research entity can't, directly or indirectly, sell, license or sub-license, rent or otherwise transfer to third parties the dataset provided by KnowDive, nor permit any third party to do so.
6. The Research entity undertakes not to use any deanonimisation technique and not to link the KnowDive data to other datasets, including public ones, and declares that it is aware that any attempt at deanonimisation could constitute an unlawful processing of personal data, entail the applicability of EU Regulation 2016/679 and expose also, but non only, to administrative fines.
7. KnowDive, in relation only to the dataset provided to the Research entity, grants a license to use it and in particular the following permissions:
   (a) KnowDive hereby grants to the above-mentioned Research entity a temporary, non-exclusive, royalty-free, non-transferable, not sub-licensable license for the (temporary) storage and to use the provided dataset;
   (b) the license granted under this Agreement relates to the use of the provided dataset exclusively for non-commercial purposes;
   (c) the creation of derivative works by the Research entity is permitted within the limits indicated in Section B.6 and only if necessary and functional to the publications related to the research activity approved by KnowDive;
   (d) the license only allows to use the provided data to conduct statistical analysis for scientific purposes
   (e) the data provided may be disclosed, only in aggregate form, as a result of the approved research activities, in compliance with the provisions of Section B.6, without any information that may permit the identification of individual statistical units;
8. This license shall be valid only until the end of the period specified in the research proposal or the different period expressly authorised and specified or otherwise communicated by KnowDive;
9. KnowDive may, at any time and at its sole discretion, especially in case of

violation of the terms of use, revoke the above-mentioned license, without prior notice and without owing anything to the Research entity.

10. At the end of that period, the Research entity must destroy any original scientific-use files sent by KnowDive, any copy, derivative works and any confidential data derived from the files, and sign the Declaration of data destruction, which will be sent to KnowDive.
11. Any publication related to the approved research project and to the provided dataset made by the Research entity shall be made available to KnowDive and shall contain the following references: . . .
12. All rights concerning the publications of the Research entity generated during the research project approved by KnowDive will remain with the Research entity itself.
13. The Research entity undertakes to respect the technical and organizational measures set in Section B.6 to ensure the security of the data provided by KnowDive.
14. The Research entity ensure that none of the data provided by KnowDive will be accessed by non-authorised persons or parties; these data will be accessed only by the individual researches listed in Section B.6.
15. The Research entity undertakes to inform KnowDive without undue delay about any information reasonably required and any breach of the security and/or confidentiality of the provided data.
16. The Research entity shall ensure that the Principal Researcher, the Data Manager and each Individual researcher undertake to comply, during every day scientific and professional activities within the approved project, with the ethical rules and codes of ethics applicable to them, as set out in Section B.6, and, where applicable, with the so-called European Charter for Researchers.
17. The Research entity is responsible for all and any loss or damages incurred by KnowDive, that is a result of any conduct of the Research entity itself.
18. The Research entity is responsible and liable for any and all actions performed by using the dataset provided.
19. The Research entity is responsible for restoring all damages caused to KnowDive by anyone acting on its behalf.
20. The Research entity shall indemnify and hold KnowDive harmless from and against any claims of third parties and against any fines arising from any violation of any third party right or any other unlawful act committed during the execution of the project approved by KnowDive, including damages or fines related to data processing issues.
21. Nothing in this Agreement shall be construed as conferring rights to use in advertising or otherwise the name “KnowDive” or his logos or trademarks without his prior written approval.
22. This Agreement shall commence on the date of the last signature to this Agreement and shall continue until the end date of the Project as detailed in the research proposal, when (unless agreed otherwise between the parties in writing) this Agreement shall automatically terminate.
23. This Agreement may be terminated by KnowDive by written notice having

immediate effect if the Research entity is in material breach of any of its obligations, representations or warranties hereunder and have failed to effect any remedy in due time after a written notice requiring such remedy has been given by KnowDive specifying a time of not more than thirty (30) days within which the remedy is to be effected;

24. The parties acknowledge and agree that the provisions of articles [16, 17, 18, 19, 20, 25, 26, 27, 28, 29] of this Agreement are intended to survive, and continue in effect after the termination or expiry of this Agreement.
25. Should any provision of this Agreement be or become invalid, illegal or unenforceable, it shall not affect the validity of the remaining provisions of this Agreement.
26. The Research entity shall not be entitled to act or to make legally binding declarations on behalf of KnowDive. Nothing in this Agreement shall be deemed to constitute any kind of formal grouping or entity between the Parties.
27. No rights or obligations of the Research entity arising from this Agreement may be assigned or transferred, in whole or in part, and no obligations of the Research entity may be subcontracted.
28. Any further modification to the text of this Agreement (including Section B.6) require a written agreement to be signed between the Parties.
29. This Agreement is drawn up in English, which language shall govern all documents and notices between the Parties.
30. The Agreement is governed by the laws of Italy without reference to its conflict of law principles. Any dispute arising out of the Agreement shall be settled by the competent court located in [...].

For KnowDive,

Name; ....

Place and Date: ....

Signature: ....

For the Research entity,

Name; ....

Place and Date: ....

Signature: ....

Articles [2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15 16, 22, 23, 24, 30] of the Agreement are expressly approved by the Parties.

KnowDive,

Signature: ....

Research entity,

Signature: ....

Guidelines for publication

Introduction

These guidelines explain how KnowDive data are to be used by researchers. The researchers must read the guidelines before they use KnowDive data and respect the rules for publication laid down below, in accordance with Section B.6 signed by the duly designated representative of the research entity, the researchers are bound for publication by the following guidelines.

Researchers must ensure that all research published or otherwise disseminated does not contain information that allows individual statistical units (persons, households, enterprises, etc.) To be identified. In all reports, including both published and unpublished papers, researchers must ensure they have abided by the strict application of the guidelines for publication from Section B.6.

What is confidential data?

Data are confidential if the respondents can be identified. Simply removing name and address details from the KnowDive data files does not prevent the identification of the survey respondents. Specific or unique characteristics of the survey respondent may lead to their recognition. The scientific-use files delivered to researchers are especially prepared to make the identification of survey respondents more difficult. This is done by:

- Reducing the level of detail of the data.
- Modifying certain values.
- And/or suppressing risky records or variables.
- Why it is important to protect confidentiality (identity) of the respondents

We collect data from individual respondents to produce statistics. Researchers may be granted access to files containing information on individual respondents (access to KnowDive data) to conduct statistical analysis for scientific purposes. KnowDive data contain information provided by individuals or organizations. Each record in the KnowDive data files represents information provided by the respondents. Researchers granted access to confidential data are only permitted to use the data to conduct statistical analysis for scientific purposes.

Researchers are prohibited from identifying individuals or organizations represented in the files.

Disclosure of individual information constitutes a breach of the law, but also a breach of the trust the respondents place in the statistical system. Such a breach could harm the reputation of that statistical system and lead to a reduction in the quality of official statistics.

Conditions of access

Only researchers belonging to a recognized research entity may request access to KnowDive data. The research entity's duly designated representative is obligated to sign a confidentiality undertaking. Access to KnowDive data may be granted if:

1. The research proposal submitted by the researcher(s) has been approved. Each research proposal must be countersigned by the contact person identified in the confidentiality undertaking.
2. All researchers requesting access to KnowDive data have signed a confidentiality declaration.

Researchers must keep KnowDive data files secure to ensure it is not accessible by anyone who is not authorized to access the data. Results of the statistical analysis that may contain information on individual respondents should also be kept secure. The KnowDive data must be stored on a password-protected computer. Access to the data must be restricted to authorized researchers explicitly named in the research proposal. The intermediate results of analysis containing confidential data must be stored in a protected environment. After expiry or completion of the project indicated in the research proposal (or in the event of termination of access by KnowDive), the principal researcher must destroy the dataset and any data or variables derived from it and sign a declaration to the effect that it has been ensured that all data have been destroyed. This obligation applies to the original data sent by KnowDive and to all derived data, except for the aggregated and/or analyzed data as presented in the research results/reports.

Specific rules for publication

When publishing the results of the statistical analysis for scientific purposes, researchers must comply with the specific rules laid down below.

1. Below 20 observations (unweighted sample), results must not be published.
2. From 20 to 49 observations (unweighted sample), results may be published but are to be individually identified (e.g., shown in brackets).
3. For confidentiality reasons, reports that include sample sizes must only mention 'less than 20 observations' and '20 to 49 observations' (i.e., not the actual number) for these two thresholds respectively. For unweighted sample sizes below 20 observations, the actual number of observations must not be derived from (or combined with) other information available in the reports, e.g., column or row totals.

Data matching

Researchers are prohibited from attempting to link the KnowDive data to other (including public) datasets, unless explicitly agreed by KnowDive. Matching two datasets increases the likelihood of the identification of statistical respondents represented in both datasets. What to do in case of a change in the set-up of a research project The research proposal is valid for the specified purpose (research project), period, datasets, and research entity (or entities). A new research proposal must be submitted to KnowDive if any of the following situations arises:

1. The data are to be used for new research project.

2. A different set of data is needed.
3. A new research entity joins the project.

If a more recent release of the data is needed for an ongoing research proposal and/or a researcher is replaced or added to the team in the same research entity taking part in the project, a principal researcher or contact person in the research entity should inform KnowDive of these changes in writing (see forms below). An individual confidentiality declaration must be signed by each researcher taking part in the project. Form to ask staff changes form to ask project extension

Your responsibility to protect confidential data

Both the confidentiality undertaking signed by the duly designated representative of the research entity and the individual confidentiality declaration signed by the individual researcher provide the basis for legal action in cases where conditions of these documents have been neglected. The commission can take action in the event of a breach of confidentiality as follows:

1. By revoking the offending researcher's (and if necessary, his/her research entity's) access to KnowDive data.
2. By suggesting the research entity takes disciplinary action against the researcher.
3. By claiming civil law compensatory damages from the research entity. The confidentiality undertaking includes a reference to the applicable law and competent court.
4. And/or by filing a complaint or by reporting the breach to the police based on national legislation. The commission may participate in national proceedings as plaintiff.

Depending on the situation, sanctions may be applied on researchers or their research entities.

Legal basis

Legal basis for granting access to confidential data can be found in the regulation GDPR.

Any questions?

If any issue in this manual is unclear, or in case of further questions on the use of confidential data, please contact us: datadistribution@knowdive.it.

General form for changes

This form is to be used for changes of the project's core information, such as adding datasets or amending the scope. Please do not use this form for administrative changes (project extension, staff changes, and research entity changes). The dedicated forms for administrative changes: project extension, staff changes, and research entity changes.

How to use this form?

1. Fill in the form with the corresponding information.
2. Print the form, without this cover page, on the letterhead paper of your organization.
3. Have the form initialed, dated, and signed by the principal researcher and the contact person of the leading entity.
4. Have the individual confidentiality declaration dated and signed by all new researchers to be working with KnowDive Data.
5. Send all forms to datadistribution@knowdive.it

We would like to .... in the research project entitled .... reference number ....

Please explain very briefly what and why should be changed in the current research project proposal. In case of dataset to be added, at least the following sections must be included: (i) description of the research proposal, (ii) objectives, (iii) need for KnowDive Data, (iv) datasets selection and type, (v) variables description, and (vi) data use and methods.

All other information contained in the research proposal referred to above remain unchanged. I hereby certify that the information contained in this questionnaire is complete, accurate and correct and that any future change will be reported immediately to KnowDive. I understand that KnowDive is authorized to check at any time the accuracy of the information given in this questionnaire. I understand that KnowDive may also request more information, if necessary. Furthermore, I commit myself to take and maintain all necessary measures in compliance with the requirements stated in the confidentiality declaration and according to the "Regole deontologiche per trattamenti a fini statistici o di ricerca scientifica" (https://www.garanteprivacy.it/web/guest/home/docweb//docweb-display/docweb/9069637).

Principal researcher

Place and Date: ....

Signature: ....

Declaration of data destruction

How to use this form?

1. Fill in the form with the corresponding information.
2. Print the form, without this cover page, on the letterhead paper of your organization.
3. Have the form initialed, dated, and signed by the principal researcher and the contact person of the leading entity.
4. Have the individual confidentiality declaration dated and signed by all new researchers to be working with KnowDive Data.
5. Send all forms to datadistribution@knowdive.it

I .... the principal researcher of the research project proposal number: .... proposal entitled: ....

□ I declare that the data have been used for the project described above and that the references to the published results will be (have been) sent to KnowDive.
  - Please use this link: https://xxx to send the references to publications (once available).
  - If no publications have been (will be) produced, please explain why: ....

□ I declare the data have not been used for the project described above for the following reason(s):

Furthermore,

□ I declare that all original files sent by KnowDive, and any data or variables derived from these files have been destroyed. I have ensured that there is no remaining copy of the dataset(s). The obligation to destroy derived data does not concern non-confidential: aggregated and/or analyzed data as presented in the published results/reports.

□ I will use the data for the project number of the research project proposal entitled: .... validated by KnowDive on ....

Principal researcher

Place and Date: ....

Signature: ....

C

iLog and data collection support material

C.1 iLog sensor list

iLog collects data in the background from a pre-selected list of sensors, with no user intervention. The data are generated as a time series, consisting of tuples composed of a timestamp and one or more values. The collected sensors are reported in the Tables C.1, C.2, C.3. In these tables, the value *Small/Big* in the last column (column *Category*) intuitively means that the size of the dataset generated by these sensors is comparatively small (or big). The sensor data collected by iLog are organized into three categories as follows:

- *Hardware (HW) sensors* are sensors that one can find in a phone, e.g., accelerometer, gyroscope, GPS. The complete list of HW sensors used in this survey is reported in Table C.1;
- *Software (SW) sensors*, by which we mean all the SW events that can be collected from the Operating system and SW, for instance, the Wifi the HW is connected to and so on. The complete list of SW sensors is reported in Table C.2;
- *QU sensors* (where QU stands for Questionnaire), by which we mean events connected with the compilation of the Time Use Diary, mainly related to the various execution times, e.g., when a question arrived or was answered. Table C.3 reports the complete list of QU sensors.

In these three sensor tables, the frequency by which the sensors are captured is reported according to the following conventions: *on change* means that the value of the sensor is recorded only when the current value is changed (along with a timestamp of when it happened), *up to X samples per second* means that for each

No	HW Sensor	Estimated Frequency	Category
1	Accelerometer	up to 10 samples per second	Big
2	Gyroscope	up to 10 samples per second	Big
3	Light	up to 10 samples per second	Big
4	Location	Once every minute	Small
5	Magnetic Field	up to 10 samples per second	Big
6	Pressure	up to 10 samples per second	Big

Table C.1: List of Hardware sensors

second the value of the sensor will be stored up to a maximum of X times (these values are estimated), and *once every Y* means that the values of a sensor are recorded once the time Y has passed (these values are estimated). The data collected from the *HW sensors* (Table C.1) are as follows:

- Sensor 1 (Accelerometer) measures the acceleration to which the phone is subjected, and it captures it as a 3D vector;
- Sensor 2 (Gyroscope) measures the rotational forces to which the phone is subjected, and it captures it as a 3D vector;
- Sensor 3 (Light) measures the ambient illumination around the phone, measured in illuminance (lux);
- Sensor 4 (Location) returns the geocoordinates of where the phone is located; for more accuracy, this sensor combines GPS and WIFI/cellular connections;
- Sensor 5 (Magnetic Field) measures the magnetic field to which the phone is subjected, and it captures it as a 3D vector;
- Sensor 6 (Pressure) measures the ambient air pressure to which the phone is subjected.

The data collected from the *SW sensors* (Table C.2) are as follows:

- Sensor 7 (Airplane Mode) returns whether the phone's Airplane mode is on or off; Airplane mode turns off all the connectivity features of the phone;
- Sensor 8 (Battery Charge) returns whether the phone is currently charging its battery;
- Sensor 9 (Battery Level) returns the phone's battery level;
- Sensor 10 (Bluetooth Devices) returns all Bluetooth devices detected by the phone;
- Sensor 11 (Bluetooth Low Energy) returns all the low-energy Bluetooth devices detected by the phone;
- Sensor 12 (Cellular Network info) returns information related to the cellular network (cellid, dbm, type) to which the phone is connected;

No	SW Sensor	Estimated Frequency	Category
7	Airplane Mode [ON/OFF]	On change	Small
8	Battery Charge [ON/OFF]	On change	Small
9	Battery Level	On change	Small
10	Bluetooth Devices	Once every minute	Small
11	Bluetooth LE (Low Energy) Devices	Once every minute	Small
12	Cellular network info	Once every minute	Small
13	Doze Mode [ON/OFF]	On change	Small
14	Headset Status [ON/OFF]	On change	Small
15	Movement Activity Label	Once every 30 seconds	Small
16	Movement Activity per Time	Once every 30 seconds	Small
17	Music Playback (no track information)	On change	Small
18	Notifications received	On change	Small
19	Proximity	up to 10 samples per second	Small
20	Ring mode [Silent/Normal]	On change	Small
21	Running Applications	Once every 5 seconds	Small
22	Screen Status [ON/OFF]	On change	Small
23	Step Counter	up to 10 samples per second	Small
24	Step Detection	On change	Small
25	Touch event	On change	Small
26	User Presence	On change	Small
27	WIFI Network Connected to	On change	Small
28	WIFI Networks Available	Once every minute	Small

Table C.2: List of Software sensors

- Sensor 13 (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;

- Sensor 14 (Headset status) returns whether the headphones of the phone were connected;

- Sensor 15 (Movement activity label) returns a label identifying the activity performed by the user. Android uses Google's Activity Recognition API and low-power signals from multiple sensors in the device to compute this value. Possible activities are: *still, in_vehicle, on_bicycle, on_foot, running, tilting, walking*;

- Sensor 16 (Movement activity per Time) similar to the previous sensor, again computed via the Google API, but data are presented grouped by time instead of being grouped by labels;

- Sensor 17 (Music Playback) returns whether music is being played on the phone (yes or no) using the default music player from the operating system;

- Sensor 18 (Notifications received) measures when the phone receives a notification and when the user dismisses it;
- Sensor 19 (Proximity) measures the distance between the user's head and the phone. Depending on the phone it may be measured in centimeters (i.e., the absolute distance) or as labels (e.g, 'near', 'far');
- Sensor 20 (Ring Mode) returns the current ring status of the phone (normal/silent/vibrate);
- Sensor 21 (Running Applications) returns the name of the application (or application package) that is currently running in the foreground of the phone;
- Sensor 22 (Screen status) returns whether the phone's screen is on or off;
- Sensor 23 (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;
- Sensor 24 (Step Detection), similar to the previous, uses the Android API to generate a step value each time the user takes a step;
- Sensor 25 (Touch event) generates a touch value each time the user touches the screen;
- Sensor 26 (User Presence) sensor that detects when the user is present near the phone, for example, when the user unlocks the screen;
- Sensor 27 (WIFI Network connected to) returns information related to the WiFi network to which the phone is connected to, if connected, will also report the WiFi network ID;
- Sensor 28 (WIFI Networks available) returns all WiFi networks detected by the smartphone.

In this table we use two different concepts: *Time diary* and *Task*, the difference being that *Time diary questions /answers* are administered at fixed time intervals (as it is the case, e.g., with the questions reported in the Figures 4.1, 4.2, and 4.3), while *Task questions /answers* can be administered any time, depending on an event triggering them (as it is the case, e.g., with the questions reported in Figure 4.5). They are as follows:

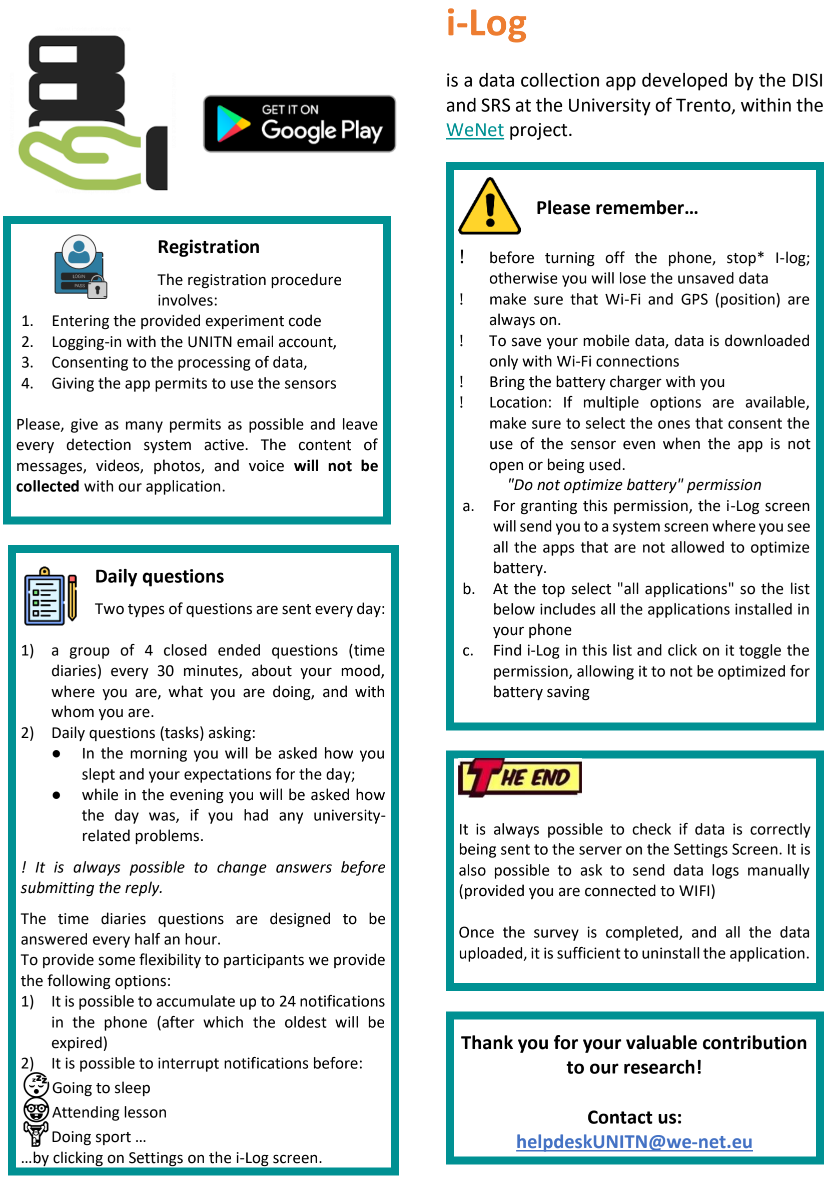
- Sensor 29 (Time Diary questions) contains the question, the question ID (for traceability purposes), and the timestamp when it was generated on the server and sent to the cloud provider for delivery;
- Sensor 30 (Time Diary confirmation) contains the timestamp at which each question, identified by its unique ID, has been delivered to the device of the participant (which may coincide or not with the time the participant sees it);

- Sensor 31 (Time Diary answers) contains the answer, the timestamp when it was saved in the server, and the difference between answer and notification times in milliseconds;
- Sensor 32 (Task questions) contains the question and the timestamp when it was sent from the server;
- Sensor 33 (Task confirmation) contains the timestamp when each specific question was notified to the participant;
- Sensor 34 (Task answers) contains the answer, the timestamp when it was saved in the server, and the difference between answer and notification times in milliseconds.

No	QU Sensor	Estimated Frequency	Category
29	Time Diary questions	On change	Small
30	Time Diary confirmation	On change	Small
31	Time Diary answers	On change	Small
32	Task questions	On change	Small
33	Task confirmation	On change	Small
34	Task answers	On change	Small

Table C.3: List of Questionnaire sensors

C.2 iLog participants instructions



i-Log

is a data collection app developed by the DISI and SRS at the University of Trento, within the WeNet project.

GET IT ON
Google Play

Registration

The registration procedure involves:

1. Entering the provided experiment code
2. Logging-in with the UNITN email account,
3. Consenting to the processing of data,
4. Giving the app permits to use the sensors

Please, give as many permits as possible and leave every detection system active. The content of messages, videos, photos, and voice **will not be collected** with our application.

Daily questions

Two types of questions are sent every day:

1) a group of 4 closed ended questions (time diaries) every 30 minutes, about your mood, where you are, what you are doing, and with whom you are.
2) Daily questions (tasks) asking:
   - In the morning you will be asked how you slept and your expectations for the day;
   - while in the evening you will be asked how the day was, if you had any university-related problems.

! It is always possible to change answers before submitting the reply.

The time diaries questions are designed to be answered every half an hour.
To provide some flexibility to participants we provide the following options:

1) It is possible to accumulate up to 24 notifications in the phone (after which the oldest will be expired)
2) It is possible to interrupt notifications before:

Going to sleep
Attending lesson
Doing sport ...
...by clicking on Settings on the i-Log screen.

Please remember...

! before turning off the phone, stop* I-log; otherwise you will lose the unsaved data
! make sure that Wi-Fi and GPS (position) are always on.
! To save your mobile data, data is downloaded only with Wi-Fi connections
! Bring the battery charger with you
! Location: If multiple options are available, make sure to select the ones that consent the use of the sensor even when the app is not open or being used.

"Do not optimize battery" permission

a. For granting this permission, the i-Log screen will send you to a system screen where you see all the apps that are not allowed to optimize battery.
b. At the top select "all applications" so the list below includes all the applications installed in your phone
c. Find i-Log in this list and click on it toggle the permission, allowing it to not be optimized for battery saving

It is always possible to check if data is correctly being sent to the server on the Settings Screen. It is also possible to ask to send data logs manually (provided you are connected to WIFI)

Once the survey is completed, and all the data uploaded, it is sufficient to uninstall the application.

Thank you for your valuable contribution to our research!

Contact us:
helpdeskUNITN@we-net.eu

Figure C.1: Example of iLog flyer from WeNet Diversity 1

C.3 iLog field supervisor handbook

C.3.1 What is iLog?

iLog[1] is a unique and innovative app for data collection. It was designed with privacy and ethics in mind, addressing issues of:

- Transparency: making people aware that the app is collecting data
- Accountability: any unexpected result can be traced
- Data protection: enabling the best safety measures and compliance with GDPR
- Lack of bias: avoiding misunderstanding by matching self-reported data with data collected by the machines

iLog is developed to collect information from users and their smartphone usage. The first contains data from the internal sensors in the phone in a completely non-intrusive manner, without the need for any interaction from the user, but keeping him informed of the current collection through an always visible notification. Each phone is equipped with many sensors, the use of which is monitored by the app, in full compliance with the rules agreed for privacy. Therefore, no sensitive content such as text messages, internet searches, or call content will be recorded. iLog can capture a varied and growing number of mobile phone sensors and events (Currently up to 41).

The second component requires, instead, the active collaboration of the user, as it involves the administration of a questionnaire. This will appear directly on the smartphone and will be comprised of two groups of questions: the first, concerning the expectations on the day and whether these will be expected or not, will appear once a day, in the morning and the evening; the second, which consists of 4 short queries, will appear at intervals of 30 minutes in 24 hours. The four questions will concern, in order, what you are doing, where you are, who you are, and finally, how you judge your mood.

C.3.2 How does it work?

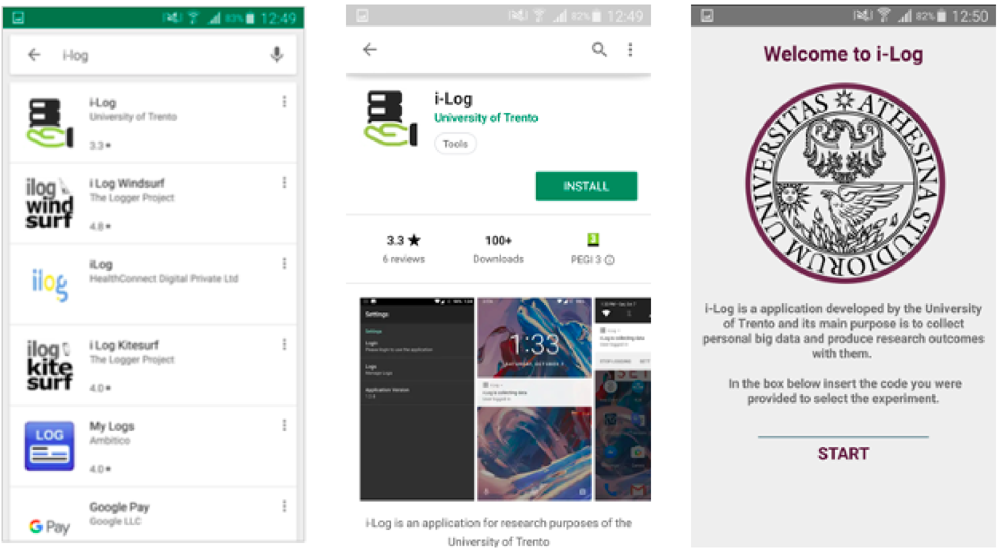
Before tackling the various parts of the app's functioning, it is useful to dwell on some general characteristics:

1. iLog does not compromise the phone's usability in any way.
2. From tests carried out, the battery consumption is at most 7% per hour, not far from regular consumption. When the battery reaches 5%, iLog turns off automatically.
3. iLog is multilingual and chooses the language set in the phone.

[1] http://datascientia.disi.unitn.it/ilog/

4. iLog starts automatically when you turn on the phone. You can stop the experiment at any time by pressing the stop button.
5. When the experiment is in progress, the user is always informed via a notification. If the notification is not active, it means that data collection is NOT in progress.

C.3.2.1 Registration procedure



i-Log is an application for research purposes of the University of Trento

i-Log is a application developed by the University of Trento and its main purpose is to collect personal big data and produce research outcomes with them.

In the box below insert the code you were provided to select the experiment.

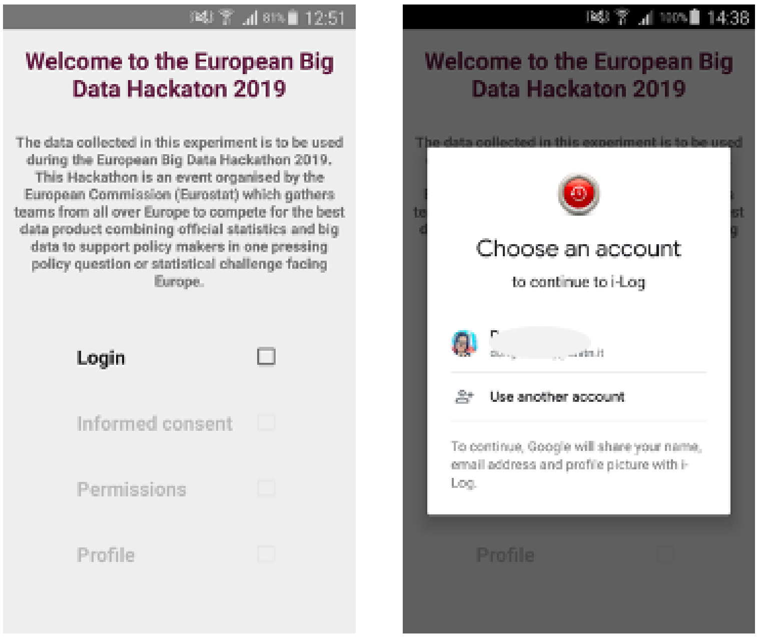
Figure C.2: Download iLog on App Store

At the first start, you will have to complete a procedure in which you will be asked to log in with your account, consent to the processing of data, permission to use the sensors and the battery, and a short profiling procedure. Give as many permits as possible and leave every detection system active. The content of messages, videos, photos, and voice cannot be detected in any way.

You can always manage the permissions and other configuration options from the Settings menu on the app icon. However, leaving every detection system active will ensure that data collection takes place at its best and in the most complete way.

What to do if...

- If you had to turn off the phone, before doing so, stop * even I-log; otherwise, you will lose data collected in the last 30 minutes.
- The sensors are critical: make sure that Wi-Fi and GPS (localization) are always on.
- Remember to bring the battery charger while the experiment is running.



12:51

Welcome to the European Big Data Hackaton 2019

The data collected in this experiment is to be used during the European Big Data Hackathon 2019. This Hackathon is an event organised by the European Commission (Eurostat) which gathers teams from all over Europe to compete for the best data product combining official statistics and big data to support policy makers in one pressing policy question or statistical challenge facing Europe.

Login □

Informed consent □

Permissions □

Profile □

14:38

Welcome to the European Big Data Hackaton 2019

Choose an account

to continue to i-Log

[redacted]

Use another account

To continue, Google will share your name, email address and profile picture with i-Log.

Profile □

Figure C.3: Access iLog with your email

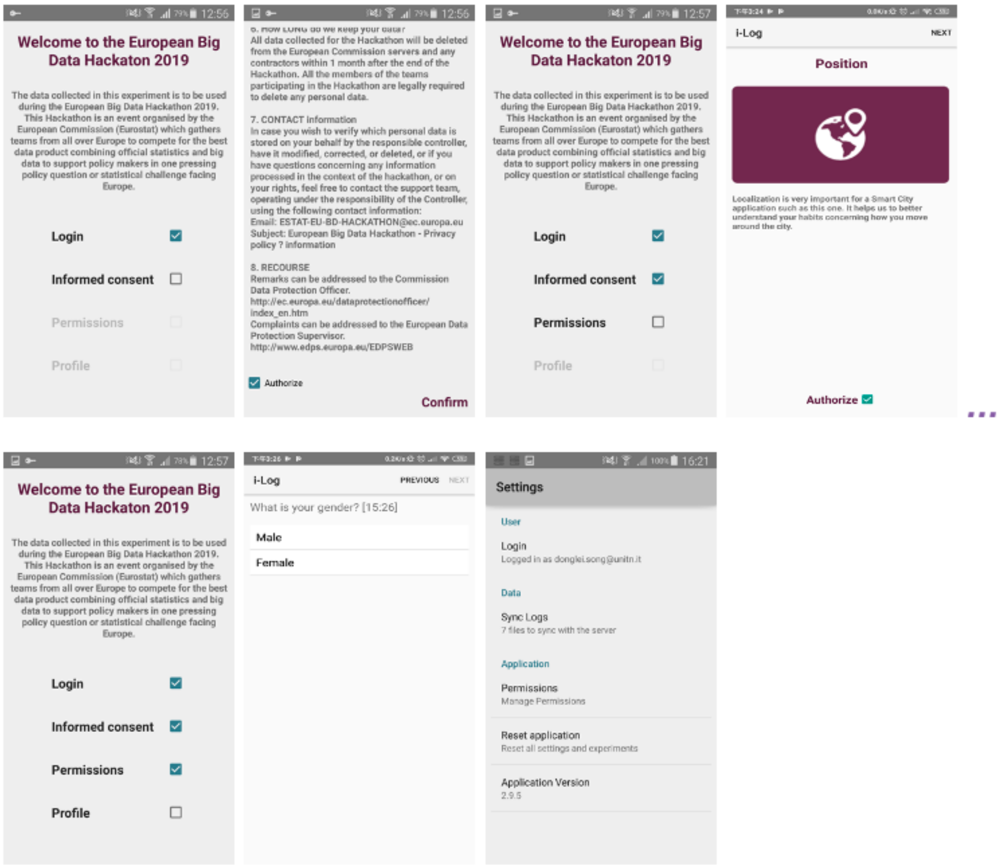
C.3.2.2 Daily questions

Answer times As specified in the paragraph iLog app, this data collection method should, among other things, reduce the problems previously encountered in similar investigations related to memory (forgetfulness, different memories, ...) and the lack of time. For this reason and following the principle with which the app was structured, we would ask you to fill in the questionnaires as they appear on your phone. This will not always be possible.

What to do if you cannot answer If you cannot answer the questions immediately, you can accumulate up to 5 questions. In this case, answer by starting from the oldest if possible. Before going to sleep, remember to communicate it to the iLog app by clicking on the icon in the Settings menu and then on the "Sleep" option.

C.3.3 Privacy

All the iLog projects are evaluated by the Ethics Committee for Experimentation with Human Beings and are held by the EU Regulation 2016/679 "General Regulation on the protection of personal data" (GDPR), the D.lgs. n. 196/2003 "Code regarding the protection of personal data" and the relative Annex A.4 "Code of ethics and good conduct for the processing of personal data for statistical and scientific purposes" (Provision of the Guarantor No. 2 of the June 16, 2004, Official Journal August 14, 2004, No. 190) sanction the right of every person to the protection of personal data.



Welcome to the European Big Data Hackaton 2019

The data collected in this experiment is to be used during the European Big Data Hackathon 2019. This Hackathon is an event organised by the European Commission (Eurostat) which gathers teams from all over Europe to compete for the best data product combining official statistics and big data to support policy makers in one pressing policy question or statistical challenge facing Europe.

Login ☑

Informed consent ☐

Permissions ☐

Profile ☐

6. HOW LONG do we keep your data?
All data collected for the Hackathon will be deleted from the European Commission servers and any contractors within 1 month after the end of the Hackathon. All the members of the teams participating in the Hackathon are legally required to delete any personal data.

7. CONTACT Information
In case you wish to verify which personal data is stored on your behalf by the responsible controller, have it modified, corrected, or deleted, or if you have questions concerning any information processed in the context of the hackathon, or on your rights, feel free to contact the support team, operating under the responsibility of the Controller, using the following contact information:
Email: ESTAT-EU-BD-HACKATHON@ec.europa.eu
Subject: European Big Data Hackathon - Privacy policy ? information

8. RECOURSE
Remarks can be addressed to the Commission Data Protection Officer.
http://ec.europa.eu/dataprotectionofficer/index_en.htm
Complaints can be addressed to the European Data Protection Supervisor.
http://www.edps.europa.eu/EDPSWEB

☑ Authorize

Confirm

Welcome to the European Big Data Hackaton 2019

The data collected in this experiment is to be used during the European Big Data Hackathon 2019. This Hackathon is an event organised by the European Commission (Eurostat) which gathers teams from all over Europe to compete for the best data product combining official statistics and big data to support policy makers in one pressing policy question or statistical challenge facing Europe.

Login ☑

Informed consent ☑

Permissions ☐

Profile ☐

i-Log NEXT

Position

Localization is very important for a Smart City application such as this one. It helps us to better understand your habits concerning how you move around the city.

Authorize ☑

Welcome to the European Big Data Hackaton 2019

The data collected in this experiment is to be used during the European Big Data Hackathon 2019. This Hackathon is an event organised by the European Commission (Eurostat) which gathers teams from all over Europe to compete for the best data product combining official statistics and big data to support policy makers in one pressing policy question or statistical challenge facing Europe.

Login ☑

Informed consent ☑

Permissions ☑

Profile ☐

i-Log PREVIOUS NEXT

What is your gender? [15:26]

Male

Female

Settings

User

Login
Logged in as donglei.song@unitn.it

Data

Sync Logs
7 files to sync with the server

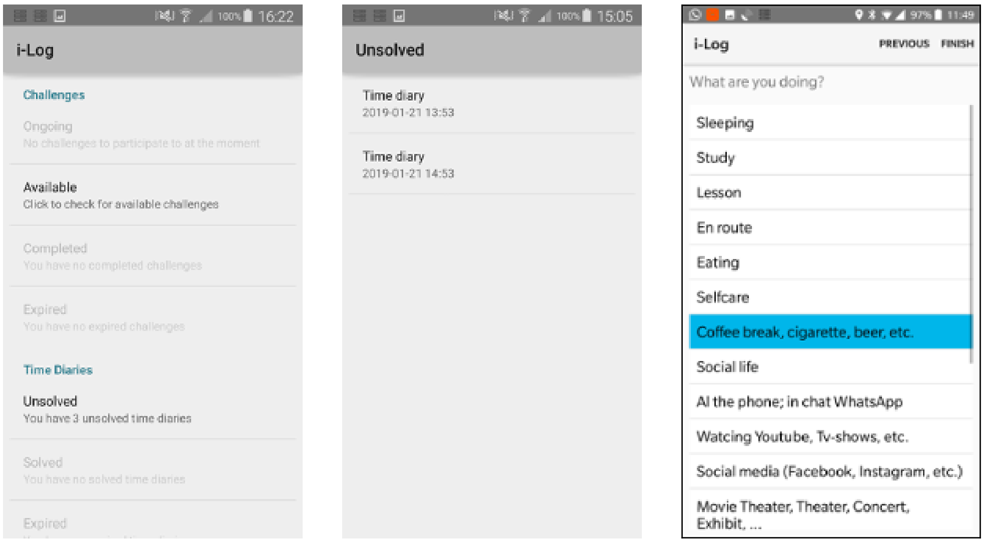
Application

Permissions
Manage Permissions

Reset application
Reset all settings and experiments

Application Version
2.9.5

Figure C.4: iLog study registration procedure



i-Log

Challenges

Ongoing
No challenges to participate to at the moment

Available
Click to check for available challenges

Completed
You have no completed challenges

Expired
You have no expired challenges

Time Diaries

Unsolved
You have 3 unsolved time diaries

Solved
You have no solved time diaries

Expired

Unsolved

Time diary
2019-01-21 13:53

Time diary
2019-01-21 14:53

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, ...

Figure C.5: Example of iLog daily questions

Under the aforementioned legislation, the processing of personal data by researchers involved in the research activity of the WeNet Project will be based on compliance with the principles outlined in Art. 5 of the GDPR and, in particular, to those of legality, correctness, transparency, relevance, not excess and to guarantee adequate security of personal data.

The data are collected, stored, and processed anonymously. Each participant will be assigned an ID so as not to be able to relate it to his true identity. Furthermore, personal data may be communicated anonymously to other universities, institutes, research institutions, and researchers for similar research purposes. Participants can request their data, withdraw from the survey anytime, and ask for data and ID.

C.3.4 FAQs

I can't find the app on the Google Play store Please check that you have an Android 6.0 or higher operating system installed on your smartphone. Typically, the information is found in Settings » System Info. If so, please check punctuation in your Google Play Store (iLog) search.

Is iLog compatible with my smartphone A: iLog works on every phone that runs Android version 6 or up, you can check your operating system version in the Settings of your device. On some smartphones, there are some aggressive battery-saving techniques, i.e., Xiaomi, Huawei, latest Samsung, among others. On these phones, it is usually possible to exclude some applications from this, and iLog should be one of them. The way to do it depends on different factors, and you should probably search on Google how to do it for your phone/operating system. To be clear, it happens also with WhatsApp and other prevalent applications, however, they are so popular that they are excluded by the battery-saver policies by default. On devices that use stock Android operating systems, it is enough to grant iLog the battery permission during installation, but unfortunately, this is not enough on many devices.

Can I install the app on tablets? We prefer using personal phones for the experiment since they are easier to carry everywhere. If you can only use a tablet, you could try but remember to always take the tablet with you to participate in the experiment in the best possible way. Nevertheless, there are some technical issues with tablets not having all the sensors the phones have, so they should only be used as a last resort.

How do I know if the application is working? The application is running if at least one notification saying "iLog, Tracking is activated" appears in the smartphone's notification bar. You should see one (or multiple) small icons on the top left of the screen and one (or multiple) notifications when you swipe down from the top (NB. this is smartphone-dependent; some brands do not show the icons, on some of them you have to swipe with one finger, two fingers, etc).

Which email account should I use to log in? You should use your personal email, which is the main Gmail account that you use to access the services on your smartphone.

Do I have to agree to all the required permissions? Yes, otherwise you will not be able to complete the installation procedure. You can manage the permissions by accessing the app menu, but to make the experiment successful, I would ask you to always keep the GPS on.

Does iLog acquire all kinds of permission for my phone? No, iLog only accesses some of the permissions according to the purpose of the study. For more information please refer to the instruction 1.5 Permissions.

Why my steps of permission are different from the instructions? Depending on your Android version some permission granting may be slightly different

A friend of mine installed the app and I would like to participate in the experiment too Right now, we are not receiving additional participants because they would need to fill out the first survey (which is unavailable anymore) to participate.

Does iLog consume giga/data connection? No, iLog was designed not to consume data on your smartphone. The data is collected automatically only when you are connected to a Wi-Fi. You can change this option by accessing the Settings on the iLog screen.

Is the application using a lot of disk space? It should not. The application stores the logs momentarily on the device before synchronizing them over Wi-Fi. If you are always connected to Wi-Fi (like most people do), it does not occupy any space. If you experience high storage occupation rates, try to see if you have unsynchronized logs.

Is the application going to affect the battery life of the smartphone? To a certain extent, yes, but it depends. Depending on the experiment you will be participating in (and consequently on the sensors used), the application can consume no to a considerable amount of battery. But keep in mind that iLog runs continuously and is not like other applications that you open, use for 5 minutes, and then close. Compared to it, Facebook, Pinterest, and others consume way more battery! In the end, if you use Facebook, how long will your battery last? I bet less than some hours... We put a lot of effort into minimizing the impact on the battery life of iLog. We can give you some pieces of advice, and some lessons learned so that you can use them in the most efficient way possible:

1. Keep the smartphone's Wi-Fi on and, if possible, let it connect to available networks. If the phone is connected to a Wi-Fi, iLog detects the user's location (when applicable) from there and not from the GPS sensor (that is the most energy-depending component of the device, apart from the screen (that we do not use));
2. Inside or close to buildings, there is no GPS signal. If you are in the office, or at your desk, you should be connected to a Wi-Fi network, otherwise, iLog keeps trying to detect the GPS signal and this is an energy-demanding task.

Is iLog compatible with external Battery Manager applications On the Google Play store nowadays it is possible to find many applications that promise to save the battery life of your device. Some devices come with this kind of application pre-installed. As a general rule, iLog is not compatible with them because they simply kill all the applications that run in the background. Smarter applications allow white-listing and exclude some applications from this behavior: iLog should be one of them. To be clear, it happens also with Whatsapp and other very popular applications, however, they are so popular that they are excluded by the battery-saver policies by default.

Is iLog compatible with my smartphone's (temporary) Battery Saver settings? On many Android phones there is the possibility to temporarily enable a specific setting that limits battery consumption, it is usually called Battery Saver (the name can be different, it depends on your phone brand/model) and is accessible from the top menu (scrolling down). This functionality should be used rarely by the user in those situations when the battery level is low and you need it to last some more hours. When this functionality is enabled, iLog (as well as any other application running in the background and/or fetching data in the background) is blocked by the operating system and cannot collect data. To overcome this limitation, most operating systems allow the exclusion of applications from this behavior: iLog should be one of them. The way to do it depends on different factors and you should probably search on Google how to do it for your phone/operating system. To be clear, it happens also with WhatsApp and other very popular applications, however, they are so popular that they are excluded by the battery-saver policies by default. If you are using Xiaomi phones, please turn off "Battery Saver" manually after finishing the installation. Refer to i-log instruction 1.5. Permission-Battery-Xiaomi.

If you are using *Samsung* phones, please turn off the "Battery Saver" and apps similar to the "Smart Manager app" manually after finishing the installation. Refer to i-log instruction 1.5. Permission-Battery-Samsung.

If you are using a *Sony* phone, you can manually put iLog into "Power-saving exceptions" after finishing the installation. Refer to i-log instruction 1.5. Permission-Battery-Sony.

If you are using *Huawei* phones, please turn off "Battery optimization" manually after finishing the installation. Refer to i-log instruction 1.5. Permission-Battery-Huawei.

How do I access notifications? Notifications should appear directly on your smartphone screen. If not, try to check from the (Settings) menu of the app, which is located on the smartphone screen (not the iLog icon). If you don't see the settings button, pull the screen down with two fingers.

I can't find the app menu The (Settings) menu of the app is located on the smartphone screen (not the iLog icon). If you don't see the settings button, pull the screen down with two fingers.

Do I always have to answer every half hour? No, although, for the sake of a successful experiment, it is advisable to try to answer every half hour. Applications accumulate up to a maximum of 12 notifications (i.e. 6 hours of detection). After that, iLog will erase the older one to make space for new questions that appear every half hour. Therefore, to avoid losing too many notifications (and the possibility of being paid) and forgetting what you were doing or having to fill in too many time diaries in one go, I recommend checking the app at least every 2 hours. Additionally, you have the option to stop notifications in three moments:

1. when you go to sleep
2. when you are in class
3. when you play sports

You can find this option in the app settings.

Can I stop notifications? Yes, you have the option to stop notifications in three moments:

1. when you go to sleep
2. when you are in class
3. when you play sports

You can find this option in the app settings.

Is it normal that I don't receive notifications on iLog (when applicable) No it is not normal, if you had internet connection (which is required to receive questions) then please send a request for help to the experiment helpdesk (more information below).

Can I turn off the smartphone? Yes, you can. Remember before stopping the app (you can find the option on the iLog drop-down menu). In this way all your data that you have not yet synchronized will be kept on your smartphone. If you do not stop iLog, all unsaved data (and therefore not sent to the server) will be lost. To avoid this, I recommend checking now and then the number of notifications you haven't sent yet. You can see them through: Smartphone screen (Do not click on the iLog icon - if you don't see the settings button, pull the screen down with two fingers) » Settings » Data » Manage Log » "n. files to sync".

What happens if my smartphone shuts down? If you do not stop iLog, all unsaved data (and therefore not sent to the server) will be lost. To avoid this, I recommend checking now and then the number of notifications you haven't sent yet. You can see them through: Smartphone screen (Do not click on the iLog icon - if you don't see the settings button, pull the screen down with two fingers) » Settings » Data » Manage Log » "n. files to sync".

Furthermore, even if iLog has been designed to consume a maximum of 7% of daily battery, I recommend that you always carry the charger with you for the duration of the experiment.

My phone broke, and I just got a new one. Can I continue the experiment? Yes of course. Once you download the app you will only have to be careful to log in with the same credentials that you used when you installed it the first time. In this way, we can also recover all your data.

Please, remember that for the experiment's success (and therefore to be paid), you will have to fill in at least 85% (i.e., 11 days) of the notifications.

Time diaries, what does it mean ...?

- *Personal care* includes activities such as brushing, combing, make-up, having a shower, etc.
- *Social Life / Entertainment*: hang out with friends or colleagues that involve relaxing activities such as going to the pub, walking, or going to dance / karaoke, etc.
- *Social media* refers to the use of Facebook, Instagram, Twitter, etc.
- *Internet (for leisure)* refers to the use of Internet not for study/work activities, which involves watching TV series, movies, YouTube, or commercial sites such as Amazon and eBay, look at the personal Mail, ...
- *Cultural activity (Cinema, Theater, ...)* includes all cultural activities carried out outside the home. Also included are Concerts, participation in Conferences, ...

- *Arts / Hobbies* are all kinds of entertainment, even occasional, not included in the previous wording. If you are an artist or a professional musician and you are preparing for a performance, you will prefer to indicate the hours spent for this purpose as Study or Work
- *Rest* is considered as the afternoon rest, different from the option e. To sleep
- *Home* is the DOMICILE you have during the period in which you attend classes. If you live with your parents, select Parents Home, which concerns the place where there is your legal RESIDENCE.
- *Library* is intended as a university library, different from the Municipal or Foundations, for which the option *Others Library* should be selected.
- *Friends home*, acquaintances or distant relatives who are not your family members
- *Work* includes both the activities you do for your (paid) job and those you do for an internship (at university or another institution/company). For activities other than those mentioned, you can select the Study or Volunteer options.

D

Data preparation support materials

Contents

This section presents the templates useful for describing the data collected and the project that carried out the data collection. Together with the datasets, this documentation is the main output of the data preparation and management phase and is useful both for the internal description of the datasets (to facilitate their management) and for external distribution (to facilitate their understanding and reuse). To this end, the documentation is made up of two main templates, namely the template for the codebook (which describes the datasets collected) and the one for the technical report (which describes the investigation and data collection process).

D.1 Codebook template

D.1.1 Data report overview

The dataset examined has the following dimensions:

Feature	**Result**
Number of observations	.
Number of variables	.

Table D.1: Codebook overview

D.1.2 Summary table

Variable	**Description**	**Class**	**Unique values**	**Missing**
Acronym of the variable present in the dataset	Description of the variable	Variable class (numeric, string, datetime, boolean, ...)	Unique values or number of users	Missing values

Table D.2: Overall description of each collected variable

D.1.3 Descriptive statistics

D.1.3.1 Questionnaire

Feature	**Result**
Variable type	.
Number of missing obs.	.
Number of unique values	.
Mode	.
Reference category	.

Table D.3: Descriptive statistics for categorical variables

List of labels:...

D.1.3.2 Active data

D.1.3.3 Passive data

D.2 Technical report template

The technical report describes:

Feature	Result
Variable type	.
Number of missing obs.	.
Number of unique values	.
Median	.
Min and Max	.

Table D.4: Descriptive statistics for numeric variables

Userid	Day 1	Day 2	Day n
1	.	.	.
2	.	.	.
3	.	.	.

Table D.5: Enter in the table the number of daily contributions (e.g., answers to questions, photos) sent by each participant.

	Sensor name		
Userid	Day 1	Day 2	Day n
1	%	%	%
2	%	%	%
3	%	%	%

Table D.6: Descriptive statistics for passive data with predefined frequencies (e.g., accelerometer). Enter the percentage of observations obtained for each participant on each survey day in the table.

	Sensor name		
Userid	Day 1	Day 2	Day n
1	.	.	.
2	.	.	.
3	.	.	.

Table D.7: Descriptive statistics for passive data without predefined frequencies (e.g., running app). Enter the number of observations obtained for each participant on each survey day in the table.

1. The purpose of the data collection (what: defines the boundaries in which data are applicable)

2. The SOA on methodology and datasets (why: absence of this type of data or helps to find similar data available)

3. Explanation of how the data were collected (data collection process)

4. Outline of type of data collected, potential reuse (e.g., example of bivariate statistics)

D.2.1 TITLE: LivePeople Data Descriptor (Template)

AUTHOR: Author 1, Author 2

INSTITUTE: University of Trento, University of ...

Abstract

The abstract should describe the data collection process, the analysis performed, the data, and their reuse potential. It should not provide conclusions or interpretive insights. Minimum length 100 words / maximum length 500 words.

D.2.2 Background & Summary

An overview of the study design, the assay(s) performed, and the created data, including any background information needed to put this study in the context of previous work and the literature. The section should also briefly outline the broader goals that motivated the creation of this dataset and the potential reuse value. A figure that provides a schematic overview of the study and assay(s) design may be included (see e.g., Figure D.1). The Background & Summary should not include subheadings and should be max 700 words.

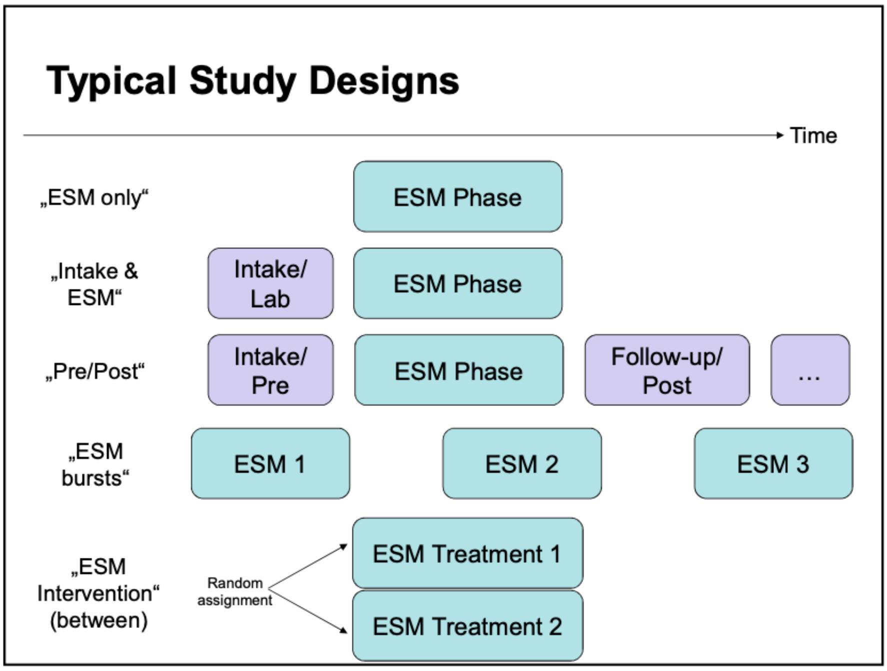


Figure D.1: Typical study design - schematic representation of the protocol

The required information to complete this section should be found in the **Research Design** document and the pre-registration document (optional).

D.2.3 Methods

The Methods should include detailed text describing any steps or procedures used in producing the data, including full descriptions of the experimental design, data acquisition assays, and any computational processing (e.g. normalization, image feature extraction). Related methods should be grouped under corresponding subheadings where possible, and methods should be described in enough detail to allow other researchers to interpret and repeat if required, the full study. Authors should cite previous descriptions of the methods under use, but ideally, the method descriptions should be complete enough for others to understand and reproduce the methods and processing steps without referring to associated publications. There is no limit to the length of the Methods section.

Below are the sub-paragraphs that are typically used to describe the methods.

1. Data collection tools
2. Sample design
3. Incentives strategy

The required information to complete this section should be found in the **Research Design** document and the pre-registration document (optional).

D.2.4 Data Records

The Data Records section should be used to explain each data record associated with this work, including the repository where this information is stored, and to provide an overview of the data files and their formats. Each external data record should be cited numerically in the text of this section, for example, and included in the main reference list as described below. A data citation should also be placed in the subsection of the Methods containing the data-collection or analytical procedure(s) used to derive the corresponding record. Providing a direct link to the dataset may also be helpful to readers (e.g., `https://doi.org/10.6084/m9.figshare.853801`). Tables should be used to support the data records, and should indicate the samples and subjects (study inputs), their provenance, and the experimental manipulations performed on each (please see 'Tables' below). They should also specify the data output resulting from each data-collection or analytical step, should these form part of the archived record.

D.2.5 Technical Validation

This section presents any experiments or analyses that are needed to support the technical quality of the dataset. This section may be supported by figures and tables, as needed. This is a required section; authors must present information justifying the reliability of their data.

This section should contain information about the following:

1. Sample dropouts (and/or attrition effects)
2. Data preparation and anonymization strategies
3. Questions and scale validation

D.2.6 Usage Notes

The Usage Notes should contain brief instructions to assist other researchers with reusing the data. This may include a discussion of software packages that are suitable for analyzing the assay data files, suggested downstream processing steps (e.g., normalization, etc.), or tips for integrating or comparing the data records with other datasets. Authors are encouraged to provide code, programs, or data-processing workflows if they may help others understand or use the data. Please see our code availability policy for advice on supplying custom code alongside Data Descriptor manuscripts. For studies involving privacy or safety controls on public access to the data, this section should describe in detail these controls, including how authors can apply them to access the data, what criteria will be used to determine who may access the data, and any limitations on data use.

D.2.7 Code availability

For all studies using custom code in the generation or processing of datasets, a statement must be included under the heading "Code availability", indicating whether and how the code can be accessed, including any restrictions to access. This section should also include information on the versions of any software used, if relevant, and any specific variables or parameters used to generate, test, or process the current dataset.

References

D.2.8 Author contributions statement

Must include all authors, identified by initials, for example: A.A. conceived the study, A.A. and B.A. conducted the study, C.A. and D.A. analysed the results. All authors reviewed the manuscript.

D.2.9 Acknowledgements

Acknowledgments should be brief and should not include thanks to anonymous referees and editors or effusive comments. Grant or contribution numbers may be acknowledged.

D.2.10 Competing Interest

The corresponding author is responsible for providing a competing interests statement on behalf of all authors of the paper.

Example: All authors declare no competing interests during the data collection, preparation, and analysis of this dataset.

E

Catalog user journey

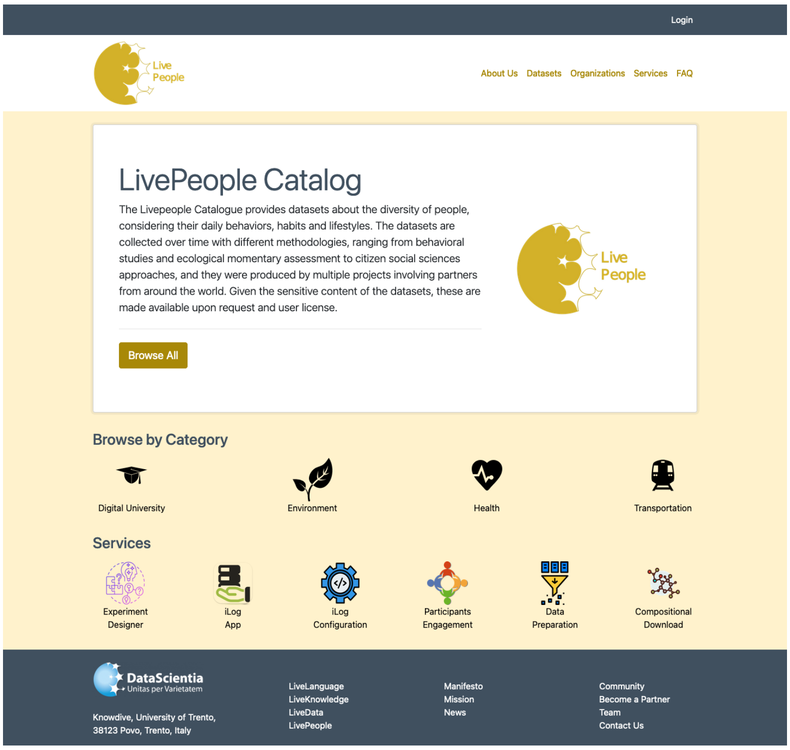
Contents

To demonstrate the use of the Catalog, in this section, we present a three stage user journey. We will also introduce connected services for full use of the Catalog. Therefore, this section is divided into 3+1 sections: (1) Accessing the catalog and (1a) exploring services; (2) Searching for datasets; (3) Exploring data sets and download requests.

E.1 Accessing the catalog and exploring services

The first step concerns the landing page in Figure E.1, where the user will find various information regarding the Dataset Catalogue. Firstly, a description of the type of datasets available, the methodology with which they are collected and a summary of the access methods. By clicking on the Browse button, she will be able to access the Catalogue. Alternatively, if she already know the type of data she is interested in, she will have the possibility to select the Category using the below banner.

An essential component of the page, which allows the user to explore the full potential of the Catalog, concerns the aspect of the services, described in the previous section. The services currently identified are:



LivePeople Catalog

The Livepeople Catalogue provides datasets about the diversity of people, considering their daily behaviors, habits and lifestyles. The datasets are collected over time with different methodologies, ranging from behavioral studies and ecological momentary assessment to citizen social sciences approaches, and they were produced by multiple projects involving partners from around the world. Given the sensitive content of the datasets, these are made available upon request and user license.

Browse by Category

Services

Knowdive, University of Trento,
38123 Povo, Trento, Italy

Figure E.1: Catalog Landing Page

1. *Experiment Designer*: support in the design and management of a data collection
2. *iLog App*: which refers to the description of the data collection app
3. *iLog Configuration*: iLog app configuration service
4. *Participants Engagement*: ability to select an on-demand panel of participants
5. *Data Preparation*: management of collected datasets
6. *Compositional Download*: ability to compose the datasets published in the catalog according to the user's needs

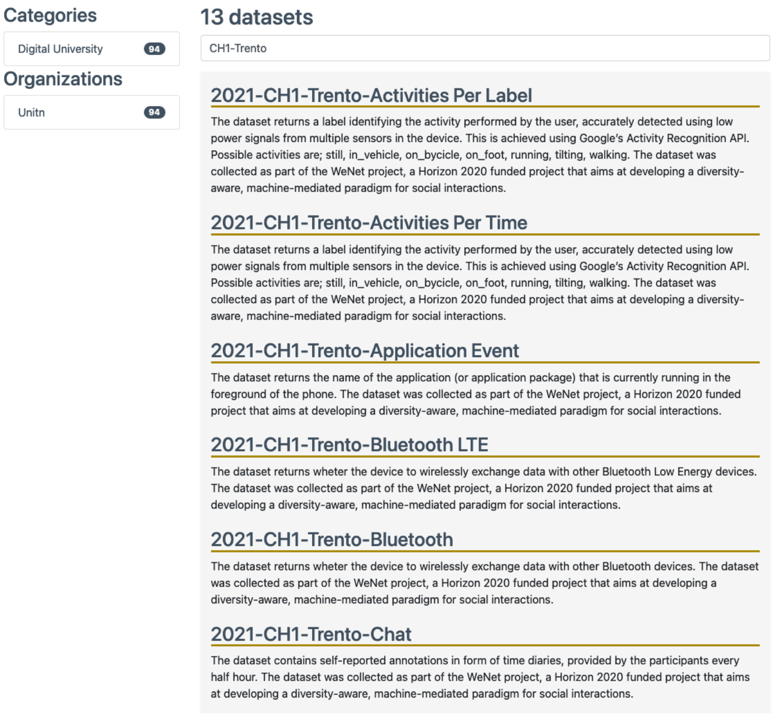
At the time of writing this deliverable the web pages for each service (except for the iLog App) are under development.

In the footer of the page, the user will find other information regarding complementary Catalogs to Catalog within the DataScientia project, a soon-to-be foundation that will act as a citizen science community in the field of generation, management, and sharing of person-centric data and their related resources (see also [242]).

E.2 Dataset search

By clicking on Browse, the user will have the possibility to see the list of datasets available in the Catalog (see Figure E.2). To ensure a detailed description of each resource and to facilitate compositionality, the datasets are presented individually and with a brief description. Through the title, the user will be able to recognize:

the acronym of the data collection, the place where it took place, and the name of the dataset.



Categories

Digital University 94

Organizations

Unitn 94

13 datasets

CH1-Trento

2021-CH1-Trento-Activities Per Label

The dataset returns a label identifying the activity performed by the user, accurately detected using low power signals from multiple sensors in the device. This is achieved using Google's Activity Recognition API. Possible activities are; still, in_vehicle, on_bycicle, on_foot, running, tilting, walking. The dataset was collected as part of the WeNet project, a Horizon 2020 funded project that aims at developing a diversity-aware, machine-mediated paradigm for social interactions.

2021-CH1-Trento-Activities Per Time

The dataset returns a label identifying the activity performed by the user, accurately detected using low power signals from multiple sensors in the device. This is achieved using Google's Activity Recognition API. Possible activities are; still, in_vehicle, on_bycicle, on_foot, running, tilting, walking. The dataset was collected as part of the WeNet project, a Horizon 2020 funded project that aims at developing a diversity-aware, machine-mediated paradigm for social interactions.

2021-CH1-Trento-Application Event

The dataset returns the name of the application (or application package) that is currently running in the foreground of the phone. The dataset was collected as part of the WeNet project, a Horizon 2020 funded project that aims at developing a diversity-aware, machine-mediated paradigm for social interactions.

2021-CH1-Trento-Bluetooth LTE

The dataset returns wheter the device to wirelessly exchange data with other Bluetooth Low Energy devices. The dataset was collected as part of the WeNet project, a Horizon 2020 funded project that aims at developing a diversity-aware, machine-mediated paradigm for social interactions.

2021-CH1-Trento-Bluetooth

The dataset returns wheter the device to wirelessly exchange data with other Bluetooth devices. The dataset was collected as part of the WeNet project, a Horizon 2020 funded project that aims at developing a diversity-aware, machine-mediated paradigm for social interactions.

2021-CH1-Trento-Chat

The dataset contains self-reported annotations in form of time diaries, provided by the participants every half hour. The dataset was collected as part of the WeNet project, a Horizon 2020 funded project that aims at developing a diversity-aware, machine-mediated paradigm for social interactions.

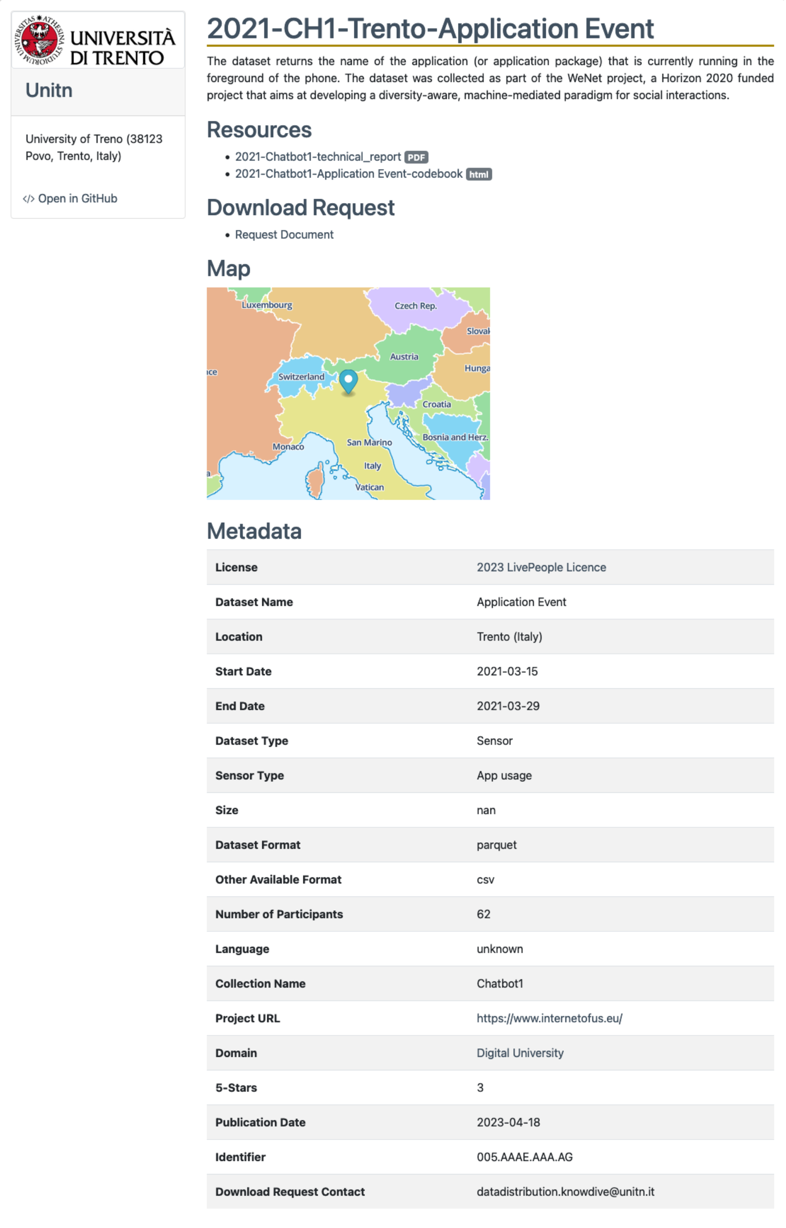
Figure E.2: Catalog Dataset Search

Through the Search function, enabled by the search bar and by the tags available on the left side of the web page, the user will be able to identify the datasets of interest. Some of the possible searches concern the acronym of the data collection and the place of collection (as in the example in Figure E.2), but also the name of the dataset and its type.

E.3 Dataset exploration and download requests

Finally, the user will be able to select the dataset of her interest by accessing the dataset page (see Figure E.3). On the dataset page, the user will find a brief description of the dataset, which includes the data collection project of which it is a part. Furthermore, the user will find a list of resources associated with the dataset, such as the technical report describing the data collection, the codebook containing a set of descriptive statistics, and the labels associated with each variable contained in the dataset and optionally additional materials (e.g., the questionnaire used or the notes of the researchers who carried out the data collection).

If interested in one or more datasets, she can make a download request by filling in the appropriate form to be sent to the email linked in the metadata.



UNIVERSITÀ DI TRENTO

Unitn

University of Treno (38123 Povo, Trento, Italy)

</> Open in GitHub

2021-CH1-Trento-Application Event

The dataset returns the name of the application (or application package) that is currently running in the foreground of the phone. The dataset was collected as part of the WeNet project, a Horizon 2020 funded project that aims at developing a diversity-aware, machine-mediated paradigm for social interactions.

Resources

- 2021-Chatbot1-technical_report PDF
- 2021-Chatbot1-Application Event-codebook html

Download Request

- Request Document

Map

Metadata

License	2023 LivePeople Licence
Dataset Name	Application Event
Location	Trento (Italy)
Start Date	2021-03-15
End Date	2021-03-29
Dataset Type	Sensor
Sensor Type	App usage
Size	nan
Dataset Format	parquet
Other Available Format	csv
Number of Participants	62
Language	unknown
Collection Name	Chatbot1
Project URL	https://www.internetofus.eu/
Domain	Digital University
5-Stars	3
Publication Date	2023-04-18
Identifier	005.AAAE.AAA.AG
Download Request Contact	datadistribution.knowdive@unitn.it

Figure E.3: Catalog Dataset Resources and Metadata

References

[1] Alexander Grosskopf, Gero Decker, and Mathias Weske. *The process: business process modeling using BPMN*. Meghan-Kiffer Press, 2009.

[2] Rob Kitchin. *The data revolution: Big data, open data, data infrastructures and their consequences*. Sage, 2014.

[3] Amon Rapp and Maurizio Tirassa. “Know thyself: a theory of the self for personal informatics”. In: *Human–Computer Interaction* 32.5-6 (2017), pp. 335–380.

[4] Nigel G Fielding, Raymond M Lee, and Grant Blank. *The SAGE handbook of online research methods*. Sage, 2008.

[5] Danah Boyd and Kate Crawford. “Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon”. In: *Information, communication & society* 15.5 (2012), pp. 662–679.

[6] Yusra Asim et al. “Context-aware human activity recognition (CAHAR) in-the-Wild using smartphone accelerometer”. In: *IEEE Sensors Journal* 20.8 (2020), pp. 4361–4371.

[7] Yonatan Vaizman, Katherine Ellis, and Gert Lanckriet. “Recognizing detailed human context in the wild from smartphones and smartwatches”. In: *IEEE pervasive computing* 16.4 (2017).

[8] Madeline Lee Pe, Peter Koval, and Peter Kuppens. “Executive well-being: Updating of positive stimuli in working memory is associated with subjective well-being”. In: *Cognition* 126.2 (2013), pp. 335–340.

[9] Alan Irwin. “Citizen science: a study of people, expertise, and sustainable development”. In: (1995).

[10] CornellLab. 2023. URL: https://www.birds.cornell.edu/home/ (visited on 12/14/2023).

[11] iNaturalist. 2023. URL: https://www.inaturalist.org/ (visited on 12/14/2023).

[12] Biying Fu et al. “Sensing technology for human activity recognition: A comprehensive survey”. In: *IEEE Access* 8 (2020), pp. 83791–83820.

[13] Xingjiao Wu et al. “A survey of human-in-the-loop for machine learning”. In: *Future Generation Computer Systems* (2022).

[14] Andrea Bontempelli et al. “Lifelong Personal Context Recognition”. In: *arXiv preprint arXiv:2205.10123* (2022).

[15] Tobias Bornakke and Brian L Due. “Big–Thick Blending: A method for mixing analytical insights from big and thick data sources”. In: *Big Data & Society* 5.1 (2018), p. 2053951718765026.

[16] Alejandra Gomez Ortega et al. “SIG on Data as Human-Centered Design Material”. In: *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 2022, pp. 1–4.

[17] Yi-Fu Tuan. *Space and place: The perspective of experience*. U of Minnesota Press, 1977.

[18] Clifford Geertz. “Thick description: Toward an interpretive theory of culture”. In: *The cultural geography reader*. Routledge, 2008, pp. 41–51.

[19] Fausto Giunchiglia. “Contextual reasoning”. In: *Epistemologia, special issue on I Linguaggi e le Macchine* 16 (1993), pp. 345–364.

[20] Fausto Giunchiglia, Enrico Bignotti, and Mattia Zeni. “Personal context modelling and annotation”. In: *2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*. IEEE. 2017, pp. 117–122.

[21] Fausto Giunchiglia et al. “Streaming and Learning the Personal Context”. In: *Twelfth International Workshop Modelling and Reasoning in Context*. Also: arXiv preprint arXiv:2108.08234. 2021, p. 19.

[22] Xiaoyue Li et al. “Representing habits as streams of situational contexts”. In: *International Conference on Advanced Information Systems Engineering*. Springer. 2022, pp. 86–92.

[23] Katherine R. Arlinghaus and Craig A. Johnston. “The Importance of Creating Habits and Routine”. In: *American Journal of Lifestyle Medicine* 13.2 (2019), pp. 142–144. eprint: https://doi.org/10.1177/1559827618818044. URL: https://doi.org/10.1177/1559827618818044.

[24] Claudio Bettini et al. “A survey of context modelling and reasoning techniques”. In: *Pervasive and mobile computing* 6.2 (2010), pp. 161–180.

[25] Daniele Riboni and Claudio Bettini. “OWL 2 modeling and reasoning with complex human activities”. In: *Pervasive and Mobile Computing* 7.3 (2011), pp. 379–395.

[26] Fausto Giunchiglia, Enrico Bignotti, and Mattia Zeni. “Personal context modelling and annotation”. In: *2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*. IEEE. 2017, pp. 117–122.

[27] Piergiorgio Corbetta. *Social research: Theory, methods and techniques*. Sage, 2003.

[28] Herbert F Weisberg. *The total survey error approach*. University of Chicago Press, 2009.

[29] WeNet - The Internet of Us. 2023. URL: https://www.internetofus.eu/ (visited on 12/14/2023).

[30] Bison Ivano et al. *D1.4 Final Model of Diversity: The Research Protocol of The Diversity Pilot Study*. Tech. rep. Written by the "WeNet - The Internet of US" (WeNet) project’s consortium under EC grant agreement 823783. 2021.

[31] de Götzen Amalia. *D7.2-User Recruitment Procedure*. Tech. rep. Written by the "WeNet - The Internet of US" (WeNet) project’s consortium under EC grant agreement 823783. 2020.

[32] de Götzen Amalia et al. *D7.3-User Recruitment Procedure*. Tech. rep. Written by the "WeNet - The Internet of US" (WeNet) project’s consortium under EC grant agreement 823783. 2022.

[33] Busso Matteo, Rodas Britez Marcelo, and Giunchiglia Fausto. *D6.6-Research Infrastructure v4.0*. Tech. rep. Written by the "WeNet - The Internet of US" (WeNet) project’s consortium under EC grant agreement 823783. 2023.

[34] Chenu-Abente Ronald et al. *D11.1-H-Requirement n.2*. Tech. rep. Written by the "WeNet - The Internet of US" (WeNet) project’s consortium under EC grant agreement 823783. 2023.

[35] Bona Roberto et al. *D11.2-Processing of Personal Data (POPD) Requirement no. 6*. Tech. rep. Written by the "WeNet - The Internet of US" (WeNet) project's consortium under EC grant agreement 823783. 2019.

[36] Fausto Giunchiglia et al. “Mobile social media and academic performance”. In: *International conference on social informatics*. Springer, Cham. Springer, Cham, 2017, pp. 3–13.

[37] Patrick E Shrout and Joseph L Rodgers. “Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis”. In: *Annual review of psychology* 69 (2018), pp. 487–510.

[38] Sayash Kapoor and Arvind Narayanan. “Leakage and the reproducibility crisis in machine-learning-based science”. In: *Patterns* 4.9 (2023).

[39] P. Ball. *Is AI leading to a reproducibility crisis in science?* 2023. URL: https://www.nature.com/articles/d41586-023-03817-6#ref-CR4 (visited on 12/14/2023).

[40] Monya Baker. “1,500 scientists lift the lid on reproducibility”. In: *Nature* 533.7604 (2016).

[41] Colin F Camerer et al. “Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015”. In: *Nature human behaviour* 2.9 (2018), pp. 637–644.

[42] Louise Corti et al. *Managing and sharing research data: A guide to good practice*. Sage, 2019.

[43] Tim May and Beth Perry. *Social research: Issues, methods and process*. McGraw-Hill Education (UK), 2022.

[44] Jan Jonker and Bartjan Pennink. *The essence of research methodology: A concise guide for master and PhD students in management science*. Springer Science & Business Media, 2010.

[45] Delbert C Miller and Neil J Salkind. *Handbook of research design and social measurement*. Sage, 2002.

[46] Aviva de Groot and Bart van der Sloot. *The handbook of privacy studies: An interdisciplinary introduction*. Amsterdam University Press, 2019.

[47] Norman M Bradburn, Seymour Sudman, and Brian Wansink. *Asking questions: the definitive guide to questionnaire design–for market research, political polls, and social and health questionnaires*. John Wiley & Sons, 2004.

[48] Paul P Biemer et al. *Total survey error in practice*. John Wiley & Sons, 2017.

[49] M Brent Donnellan et al. “The mini-IPIP scales: tiny-yet-effective measures of the Big Five factors of personality.” In: *Psychological assessment* 18.2 (2006), p. 192.

[50] Emorie D Beck and Joshua J Jackson. “Consistency and change in idiographic personality: A longitudinal ESM network study.” In: *Journal of Personality and Social Psychology* 118.5 (2020), p. 1080.

[51] Mihaly Csikszentmihalyi and Reed Larson. “Validity and reliability of the experience-sampling method”. In: *The Journal of nervous and mental disease* 175.9 (1987), pp. 526–536.

[52] Alexander Hart et al. “Using smartphone sensor paradata and personalized machine learning models to infer participants’ well-being: ecological momentary assessment”. In: *Journal of Medical Internet Research* 24.4 (2022), e34015.

[53] I. Myin-Germeys and P. Kuppens. *The open handbook of experience sampling methodology: A step-by-step guide to designing, conducting, and analyzing ESM*

studies (2nd ed.) Center for Research on Experience Sampling and Ambulatory Methods Leuven, 2022.

[54] Fons JR Van de Vijver and Kwok Leung. *Methods and data analysis for cross-cultural research*. Vol. 116. Cambridge University Press, 2021.

[55] Emiro De-La-Hoz-Franco et al. “Sensor-based datasets for human activity recognition–a systematic review of literature”. In: *IEEE Access* 6 (2018), pp. 59192–59210.

[56] Mihaly Csikszentmihalyi, Mihaly Csikszentmihalyi, and Reed Larson. “Validity and reliability of the experience-sampling method”. In: *Flow and the foundations of positive psychology: The collected works of Mihaly Csikszentmihalyi* (2014), pp. 35–54.

[57] Niels Van Berkel, Denzil Ferreira, and Vassilis Kostakos. “The experience sampling method on mobile devices”. In: *ACM Computing Surveys (CSUR)* 50.6 (2017), pp. 1–40.

[58] Apoorv Agarwal et al. “Sentiment analysis of twitter data”. In: *Proceedings of the workshop on language in social media (LSM 2011)*. 2011, pp. 30–38.

[59] Peter Kun et al. “Exploring diversity perceptions in a community through a Q&A chatbot: Design Research Society Conference 2022”. In: (2022).

[60] Wanyi Zhang. “Personal Context Recognition via Skeptical Learning.” In: *IJCAI*. 2019, pp. 6482–6483.

[61] Wanyi Zhang et al. “Putting human behavior predictability in context”. In: *EPJ Data Science* 10.1 (2021), p. 42.

[62] Andrea Bontempelli et al. “Learning in the wild with incremental skeptical gaussian processes”. In: *arXiv preprint arXiv:2011.00928* (2020).

[63] ISTAT. *Multipurpose survey on households: aspects of daily life - microdata for research purposes*. URL: https://www.istat.it/en/archivio/129934 (visited on 12/14/2023).

[64] Pitirim Aleksandrovich Sorokin and Clarence Quinn Berger. *Time-budgets of human behavior*. Vol. 2. Harvard University Press, 1939.

[65] American Time Use Surveys. URL: https://www.bls.gov/tus/ (visited on 12/14/2023).

[66] EUROSTAT. *Harmonised European Time Use Surveys (HETUS)*. URL: https://ec.europa.eu/eurostat/web/time-use-surveys (visited on 12/14/2023).

[67] Jeffrey Goldman et al. “Participatory Sensing: A citizen-powered approach to illuminating the patterns that shape our world”. In: *Foresight & Governance Project, White Paper* (2009), pp. 1–15.

[68] Victor Jupp. “The Sage dictionary of social research methods”. In: *The SAGE Dictionary of Social Research Methods* (2006), pp. 1–352.

[69] CWM Hart. “The Hawthorne Experiments1”. In: *Canadian Journal of Economics and Political Science/Revue canadienne de economiques et science politique* 9.2 (1943), pp. 150–163.

[70] Sindre Rolstad, John Adler, and Anna Rydén. “Response Burden and Questionnaire Length: Is Shorter Better? A Review and Meta-analysis”. In: *Value in Health* 14.8 (2011), pp. 1101–1108. URL: https://www.sciencedirect.com/science/article/pii/S1098301511015245.

[71] Gudrun Eisele et al. "The effects of sampling frequency and questionnaire length on perceived burden, compliance, and careless responding in experience sampling data in a student population". In: *Assessment* 29.2 (2022), pp. 136–151.

[72] GESIS. 2023. URL: https://www.gesis.org/en/home (visited on 12/14/2023).

[73] KU Leuven. *ESM Item Repository.* URL: https://esmitemrepository.com/ (visited on 12/14/2023).

[74] Mila Hall et al. "A systematic review of momentary assessment designs for mood and anxiety symptoms". In: *Frontiers in Psychology* 12 (2021), p. 642044.

[75] Tamlin S Conner and Barbara J Lehman. "Getting started: Launching a study in daily life." In: (2012).

[76] Andreas M Brandmaier et al. "LIFESPAN: A tool for the computer-aided design of longitudinal studies". In: *Frontiers in Psychology* 6 (2015), p. 272.

[77] Ginette Lafit et al. "Selection of the number of participants in intensive longitudinal studies: A user-friendly shiny app and tutorial for performing power analysis in multilevel regression models that account for temporal dependencies". In: *Advances in methods and practices in psychological science* 4.1 (2021), p. 2515245920978738.

[78] Eleanor Singer and Cong Ye. "The use and effects of incentives in surveys". In: *The ANNALS of the American Academy of Political and Social Science* 645.1 (2013), pp. 112–141.

[79] Shalom H Schwartz et al. "Extending the cross-cultural validity of the theory of basic human values with a different method of measurement". In: *Journal of cross-cultural psychology* 32.5 (2001), pp. 519–542.

[80] UkDataArchive. 2023. URL: https://www.data-archive.ac.uk/ (visited on 12/14/2023).

[81] Open Science Framework (OSF). 2023. URL: https://osf.io/ (visited on 12/14/2023).

[82] Matteo Busso, Ronal Chenu Abente Acosta, and Amalia de Götzen. "A research infrastructure for generating and sharing diversity-aware data". In: *arXiv preprint arXiv:2306.09759* (2023).

[83] Feiyu Xu et al. "Explainable AI: A brief survey on history, research areas, approaches and challenges". In: *Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II 8*. Springer. 2019, pp. 563–574.

[84] EUCommission. *Ethics guidelines for Trustworthy AI.* 2023. URL: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (visited on 12/14/2023).

[85] Jahna Otterbacher. "Crowdsourcing stereotypes: Linguistic bias in metadata generated via gwap". In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 2015, pp. 1955–1964.

[86] Jahna Otterbacher et al. "Investigating user perception of gender bias in image search: the role of sexism". In: *The 41st International ACM SIGIR conference on research & development in information retrieval*. 2018, pp. 933–936.

[87] Kalia Orphanou et al. "Mitigating Bias in Algorithmic Systems—A Fish-eye View". In: *ACM Computing Surveys* 55.5 (2022), pp. 1–37.

[88] Asunción Esteve. "The business of personal data: Google, Facebook, and privacy issues in the EU and the USA". In: *International Data Privacy Law* 7.1 (2017), pp. 36–47.

[89] Joy Buolamwini and Timnit Gebru. “Gender shades: Intersectional accuracy disparities in commercial gender classification”. In: *Conference on fairness, accountability and transparency*. PMLR. 2018, pp. 77–91.

[90] Mark Israel and Iain Hay. *Research ethics for social scientists*. Sage, 2006.

[91] Peter Singer. *A companion to ethics*. John Wiley & Sons, 1993.

[92] American Sociological Association (ASA). *Ethical guidelines for statistical practice*. URL: https://www.amstat.org/your-career/ethical-guidelines-for-%20statistical-practice (visited on 12/14/2023).

[93] ESRC. *Framework for research ethics*. 2015. URL: https://www.ukri.org/councils/esrc/guidance-for-applicants/research-ethics-guidance/framework-for-research-ethics/.

[94] European Parliament and Council of the European Union. “General Data Protection Regulation (2016/679, "GDPR")”. In: (2016). URL: https://data.europa.eu/eli/reg/2016/679/oj (visited on 12/14/2023).

[95] Article 29 Data Protection Working Party. “Opinion 05/2014 on Anonymisation Techniques”. In: (2014).

[96] Khaled El Emam and Cecilia Alvarez. “A critical appraisal of the Article 29 Working Party Opinion 05/2014 on data anonymization techniques”. In: *International Data Privacy Law* 5.1 (2015), pp. 73–87.

[97] ISO/IEC JTC 1/SC 27. *Information security, cybersecurity and privacy protection ‘ISO/IEC 27559:2022’*. Tech. rep. International Standards Organization, 2022. URL: https://www.iso.org/standard/71677.html.

[98] Creative Commons. URL: https://creativecommons.org/ (visited on 12/14/2023).

[99] EUROSTAT. URL: https://ec.europa.eu/eurostat (visited on 12/14/2023).

[100] EUROSTAT. *Microdata access*. URL: https://ec.europa.eu/eurostat/web/microdata/access (visited on 12/14/2023).

[101] Alessandro Acquisti, Laura Brandimarte, and George Loewenstein. “Privacy and human behavior in the age of information”. In: *Science* 347.6221 (2015), pp. 509–514.

[102] Rafael Capurro. “Privacy. An intercultural perspective”. In: *Ethics and information technology* 7 (2005), pp. 37–47.

[103] Charles Ess. “" Lost in translation"?: Intercultural dialogues on privacy and information ethics (Introduction to special issue on privacy and data privacy protection in Asia)”. In: *Ethics and Information Technology* 7.1 (2005), p. 1.

[104] Rui Wang et al. “StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones”. In: *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 2014, pp. 3–14.

[105] Daniel Garcia-Gonzalez et al. “A Public Domain Dataset for Real-Life Human Activity Recognition Using Smartphone Sensors”. In: *SENSORS* 20.8 (2020).

[106] George Vavoulas et al. “The mobiact dataset: Recognition of activities of daily living using smartphones”. In: *International Conference on Information and Communication Technologies for Ageing Well and e-Health*. Vol. 2. SciTePress. 2016, pp. 143–151.

[107] Mattia Giovanni Campana and Franca Delmastro. "ContextLabeler dataset: Physical and virtual sensors data collected from smartphone usage in-the-wild". In: *Data in brief* 37 (2021), p. 107164.

[108] Seungeun Chung et al. "Real-world multimodal lifelog dataset for human behavior study". In: *ETRI Journal* 44.3 (2022), pp. 426–437.

[109] Fausto Giunchiglia et al. "A survey on students' daily routines and academic performance at the University of Trento". In: (2022).

[110] Soujanya Chatterjee et al. "SmokingOpp: Detecting the Smoking'Opportunity'Context Using Mobile Sensors". In: *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 4.1 (2020), pp. 1–26.

[111] Yash Jain et al. "ColloSSL: Collaborative Self-Supervised Learning for Human Activity Recognition". In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6.1 (2022), pp. 1–28.

[112] George Vavoulas et al. "The mobifall dataset: Fall detection and classification with a smartphone". In: *IJMSTR* 2.1 (2014), pp. 44–56.

[113] Attila Reiss and Didier Stricker. "Introducing a new benchmarked dataset for activity monitoring". In: *2012 16th international symposium on wearable computers*. IEEE. 2012, pp. 108–109.

[114] Davide Anguita et al. "A public domain dataset for human activity recognition using smartphones". In: *Proceedings of the 21th international European symposium on artificial neural networks, computational intelligence and machine learning*. 2013, pp. 437–442.

[115] Denzil Ferreira, Vassilis Kostakos, and Anind K Dey. "AWARE: mobile context instrumentation framework". In: *Frontiers in ICT* 2 (2015), p. 6.

[116] John Krumm and Dany Rouhana. "Placer: semantic place labels from diary data". In: *Proceedings of ACM - UbiComp*. 2013, pp. 163–172.

[117] Mattia Zeni et al. "Fixing mislabeling by human annotators leveraging conflict resolution and prior knowledge". In: *Proceedings of ACM - IMWUT* 3.1 (2019), pp. 1–23.

[118] LimeSurvey Development Team. *LimeSurvey - The free and open source survey software tool!* 2012. URL: limesurvey.org (visited on 12/14/2023).

[119] Qualtrics Development Team. *Qualtrics*. 2005. URL: https://www.qualtrics.com (visited on 12/14/2023).

[120] Wikipedia. *Comparison of survey software*. URL: https://en.wikipedia.org/wiki/Comparison_of_survey_software (visited on 12/14/2023).

[121] Wikipedia. *Mobile phone based sensing software*. URL: https://en.wikipedia.org/wiki/Mobile%5C_phone%5C_based%5C_sensing%5C_software (visited on 12/14/2023).

[122] Merijn Mestdagh et al. "m-Path: An easy-to-use and flexible platform for ecological momentary assessment and intervention in behavioral research and clinical practice". In: (Jan. 2022).

[123] Mattia Zeni et al. "Improving time use measurement with personal big collection - the experience of the European Big Data Hackathon 2019." In: *Journal of Official Statistics* (2020).

[124] Mattia Zeni, Ilya Zaihrayeu, and Fausto Giunchiglia. "Multi-device activity logging". In: *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*. 2014, pp. 299–302.

[125] Fausto Giunchiglia, Enrico Bignotti, and Mattia Zeni. "Human-Like Context Sensing for Robot Surveillance". In: *International Journal of Semantic Computing* 12.01 (2017), pp. 129–148.

[126] Fausto Giunchiglia, Mattia Zeni, and Enrico Big. "Personal Context Recognition via Reliable Human-Machine Collaboration". In: *2018 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*. IEEE. IEEE, 2018, pp. 379–384.

[127] Jason D Runyan et al. "A smartphone ecological momentary assessment/intervention "app" for collecting real-time data and promoting self-awareness". In: *PloS one* 8.8 (2013), e71325.

[128] Frauke Kreuter et al. "Collecting survey and smartphone sensor data with an app: Opportunities and challenges around privacy and informed consent". In: *Social Science Computer Review* 38.5 (2020), pp. 533–549.

[129] Fausto Giunchiglia et al. "Assessing Annotation Consistency in the Wild". In: *2018 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*. IEEE. IEEE, 2018, pp. 561–566.

[130] Mattias Hellgren. "Extracting more knowledge from time diaries?" In: *Social Indicators Research* 119.3 (2014), pp. 1517–1534.

[131] Dina Najeeb et al. "MindLogger: a brain-computer interface for word building using brainwaves". In: *Proceedings of the 1st Workshop on Mobile Medical Applications*. 2014, pp. 6–11.

[132] Mahmood Hosseini et al. "Crowdsourcing: A taxonomy and systematic mapping study". In: *Computer Science Review* 17 (2015), pp. 43–69.

[133] Huadong Ma, Dong Zhao, and Peiyan Yuan. "Opportunities in mobile crowd sensing". In: *IEEE Communications Magazine* 52.8 (2014), pp. 29–35.

[134] Jennifer L Shirk et al. "Public participation in scientific research: a framework for deliberate design". In: *Ecology and society* 17.2 (2012).

[135] Zooniverse. 2023. URL: https://www.zooniverse.org/ (visited on 12/14/2023).

[136] SciStarter. 2023. URL: https://scistarter.org/ (visited on 12/14/2023).

[137] EU-citizen.science. 2023. URL: https://eu-citizen.science/ (visited on 12/14/2023).

[138] Amazon Mechanical Turk. 2023. URL: https://www.mturk.com/ (visited on 12/14/2023).

[139] Prolific. 2023. URL: https://www.prolific.com/ (visited on 12/14/2023).

[140] Eyal Peer et al. "Data quality of platforms and panels for online behavioral research". In: *Behavior Research Methods* (2022), p. 1.

[141] Benjamin D Douglas, Patrick J Ewell, and Markus Brauer. "Data quality in online human-subjects research: Comparisons between MTurk, Prolific, CloudResearch, Qualtrics, and SONA". In: *Plos one* 18.3 (2023), e0279720.

[142] I Sondaggi Retribuiti. URL: https://www.isondaggiretribuiti.it/sondaggi-retribuiti/ (visited on 12/14/2023).

[143] Fatimah Sidi et al. "Data quality: A survey of data quality dimensions". In: *2012 International Conference on Information Retrieval & Knowledge Management*. IEEE. 2012, pp. 300–304.

[144] Fakhitah Ridzuan and Wan Mohd Nazmee Wan Zainon. “A review on data cleansing methods for big data”. In: *Procedia Computer Science* 161 (2019), pp. 731–738.

[145] Edwin De Jonge and Mark Van Der Loo. *An introduction to data cleaning with R.* Statistics Netherlands Heerlen, 2013.

[146] Mohamed Yakout, Laure Berti-Équille, and Ahmed K Elmagarmid. “Don’t be scared: Use scalable automatic repairing with maximal likelihood and bounded changes”. In: *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data.* 2013, pp. 553–564.

[147] Xu Chu et al. “KATARA: Reliable data cleaning with knowledge bases and crowdsourcing”. In: *Proceedings of the VLDB Endowment* 8.12 (2015), pp. 1952–1955.

[148] Venkat Gudivada, Amy Apon, and Junhua Ding. “Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations”. In: *International Journal on Advances in Software* 10.1 (2017), pp. 1–20.

[149] David Camilo Corrales, Juan Carlos Corrales, and Agapito Ledezma. “How to address the data quality issues in regression models: A guided process for data cleaning”. In: *Symmetry* 10.4 (2018), p. 99.

[150] Timnit Gebru et al. “Datasheets for datasets”. In: *Communications of the ACM* 64.12 (2021), pp. 86–92.

[151] Wikepedia. *Data Cleasning.* URL: https://en.wikipedia.org/wiki/Data_cleansing (visited on 12/14/2023).

[152] Google. *OpenRefine.* URL: https://openrefine.org/ (visited on 12/14/2023).

[153] IBM. *IBM Infosphere Qualitystage.* URL: https://www.ibm.com/products/infosphere-qualitystage (visited on 12/14/2023).

[154] Python. *Pandas profiling.* URL: https://pandas-profiling.github.io/pandas-profiling/docs/master/rtd/index.html (visited on 12/14/2023).

[155] R. *dataMaid.* URL: https://cran.r-project.org/web/packages/dataMaid/index.html (visited on 12/14/2023).

[156] R. *Dealing with missing values R packages.* URL: https://www.analyticsvidhya.com/blog/2016/03/tutorial-powerful-packages-imputing-missing-values (visited on 12/14/2023).

[157] AWS. *Deequ Library.* URL: https://aws.amazon.com/tr/blogs/big-data/test-data-quality-at-scale-with-deequ/ (visited on 12/14/2023).

[158] Abe Gong, James Campbell, and Great Expectations. *Great Expectations.* URL: https://github.com/great-expectations/great_expectations.

[159] Apache Griffin. *Open Source Data Quality Solution for Big Data.* URL: https://griffin.apache.org/ (visited on 12/14/2023).

[160] UNESCO. *Recommendation on Open Science.* URL: https://www.unesco.org/en/open-science/about?hub=686 (visited on 12/14/2023).

[161] Mark D Wilkinson et al. “The FAIR Guiding Principles for scientific data management and stewardship”. In: *Scientific data* 3.1 (2016), pp. 1–9.

[162] F.A.I.R.Principles. 2023. URL: https://www.go-fair.org/fair-principles/ (visited on 12/14/2023).

[163] Michael Hausenblas. *5 star data info*. URL: https://5stardata.info/en/ (visited on 12/14/2023).

[164] ComputerOntario. 2023. URL: https://www.computeontario.ca/what-is-dri (visited on 12/14/2023).

[165] BioBank. 2023. URL: https://www.biobank.it/ (visited on 12/14/2023).

[166] Google. *Dataset Search*. URL: https://datasetsearch.research.google.com (visited on 12/14/2023).

[167] re3data. URL: https://www.re3data.org (visited on 12/14/2023).

[168] Github. *Awesome public datasets*. URL: https://github.com/awesomedata/awesome-public-datasets (visited on 12/14/2023).

[169] Nature. *Data Repository Guidance*. URL: https://www.nature.com/sdata/policies/repositories%5C#general (visited on 12/14/2023).

[170] Nature. *Scientific Data*. URL: https://www.nature.com/sdata/ (visited on 12/14/2023).

[171] Elsevier. *data in Brief*. URL: https://www.journals.elsevier.com/data-in-brief (visited on 12/14/2023).

[172] LivePeople. 2023. URL: https://datascientiafoundation.github.io/LivePeople/ (visited on 12/14/2023).

[173] Gordian Dziwis et al. "Bikerack datasets of the city of Trento". In: (2019).

[174] Fausto Giunchiglia et al. "A worldwide diversity pilot on daily routines and social practices (2020)". In: (2021).

[175] Fausto Giunchiglia et al. "A worldwide diversity chat application pilot on interactions and social practices (2021-1st Wave)". In: (2022).

[176] Fausto Giunchiglia et al. "A worldwide diversity chat application pilot on interactions and social practices (2021-2nd Wave)". In: (2022).

[177] Djamila Romaissa Beddiar et al. "Vision-based human activity recognition: a survey". In: *Multimedia Tools and Applications* 79.41 (2020), pp. 30509–30555.

[178] Jiang Wang et al. "Mining actionlet ensemble for action recognition with depth cameras". In: *2012 IEEE conference on computer vision and pattern recognition*. IEEE. 2012, pp. 1290–1297.

[179] L Minh Dang et al. "Sensor-based and vision-based human activity recognition: A comprehensive survey". In: *Pattern Recognition* 108 (2020), p. 107561.

[180] Inez Myin-Germeys et al. "Emotional reactivity to daily life stress in psychosis". In: *Archives of general psychiatry* 58.12 (2001), pp. 1137–1144.

[181] Andrea Bontempelli. "Human-Machine Alignment for Context Recognition in the Wild". PhD thesis. University of Trento, 2024.

[182] Shem Unger et al. "iNaturalist as an engaging tool for identifying organisms in outdoor activities". In: *Journal of Biological Education* 55.5 (2021), pp. 537–547.

[183] Luis-Daniel Ibáñez et al. "QROWD—A Platform for Integrating Citizens in Smart City Data Analytics". In: *Sustainable Smart Cities: Theoretical Foundations and Practical Considerations*. Springer, 2022, pp. 285–321.

[184] Kristina Host and Marina Ivašić-Kos. "An overview of Human Action Recognition in sports based on Computer Vision". In: *Heliyon* 8.6 (2022).

[185] Marieke Wichers et al. “Critical slowing down as a personalized early warning signal for depression”. In: *Psychotherapy and psychosomatics* 85.2 (2016), pp. 114–116.

[186] Kristin Gustavson et al. “Attrition and generalizability in longitudinal studies: findings from a 15-year population-based study and a Monte Carlo simulation study”. In: *BMC public health* 12 (2012), pp. 1–11.

[187] LSE. *Privacy impact assessment template*. Tech. rep. London School of Economics. URL: https://info.lse.ac.uk/staff/divisions/dts/assets/documents/policies/Privacy-Impact-Assessment-template-v2.docx (visited on 12/14/2023).

[188] STATISTA. URL: https://www.statista.com (visited on 12/14/2023).

[189] OSF. *IRB and Consent Form Examples*. URL: https://osf.io/g4jfv/ (visited on 12/14/2023).

[190] *Android Developers*. URL: https://developer.android.com/guide/topics/sensors/sensors_overview (visited on 12/14/2023).

[191] Anne Cummings, Jing Zhou, and Greg R Oldham. “Demographic differences and employee work outcomes: Effects on multiple comparison groups”. In: *annual meeting of the Academy of Management, Atlanta, GA*. 1993.

[192] S. E. JACKSON. “Consequences of group composition for the interpersonal dynamics of strategic issue processing”. In: *Advances in Strategic Management* 8 (1992), pp. 345–382. URL: https://ci.nii.ac.jp/naid/20001538270/en/.

[193] Susan E Jackson, Karen E May, and Kristina Whitney. “Understanding the dynamics of diversity in decision-making teams”. In: (1995).

[194] Martha L Maznevski. “Understanding our differences: Performance in decision-making groups with diverse members”. In: *Human relations* 47.5 (1994), pp. 531–552.

[195] Anne S Tsui, Terri D Egan, and Charles A O’Reilly III. “Being different: Relational demography and organizational attachment”. In: *Administrative science quarterly* (1992), pp. 549–579.

[196] Lisa Hope Pelled. “Demographic diversity, conflict, and work group outcomes: An intervening process theory”. In: *Organization science* 7.6 (1996), pp. 615–631.

[197] David A Harrison, Kenneth H Price, and Myrtle P Bell. “Beyond relational demography: Time and the effects of surface-and deep-level diversity on work group cohesion”. In: *Academy of Management journal* 41.1 (1998), pp. 96–107.

[198] David A Harrison et al. “Time, teams, and task performance: Changing effects of surface-and deep-level diversity on group functioning”. In: *Academy of Management journal* 45.5 (2002), pp. 1029–1045.

[199] Theodore Schatzki and Karin Knorr Cetina. “The Practice Turn in Contemporary Theory”. In: (2001).

[200] Ludwig Wittgenstein and GEM Anscombe. “Philosophical investigations”. In: (1953).

[201] Hubert Dreyfus. “Being-in-the-World: A Commentary on Heidegger’s Being and Time”. In: (1991).

[202] Anthony Giddens. *Central problems in social theory: Action, structure, and contradiction in social analysis*. Vol. 241. Univ of California Press, 1979.

[203] Anthony Giddens. *The Constitution of Society*. Cambridge: Polity Press, 1984.

[204] Pierre Bourdieu. *Outline of a Theory of Practice*. 16. Cambridge University Press, 1977.

[205] Pierre Bourdieu. *Distinction: A social critique of the judgment of taste*. Harvard University Press, 1984.

[206] Pierre Bourdieu. *The logic of practice*. Stanford University Press, 1990.

[207] Erving Goffman. *Asylums: Essays on the social situation of mental patients and other inmates*. Harmondsworth: Penguin, 1975.

[208] Bruno Latour and Steve Woolgar. *Laboratory Life: The Construction of Scientific Facts*. Vol. 80. Princeton University Press, 1986.

[209] Andreas Reckwitz. “Toward a theory of social practices: A development in culturalist theorizing”. In: *European Journal of Social Theory* 5.2 (2002), pp. 243–263.

[210] Elizabeth Shove, Mika Pantzar, and Matt Watson. *The dynamics of social practice: Everyday life and how it changes*. Sage, 2012.

[211] Elizabeth Shove and Mika Pantzar. “Consumers, producers and practices: Understanding the invention and reinvention of Nordic walking”. In: *Journal of consumer culture* 5.1 (2005), pp. 43–64.

[212] Inge Røpke. “Theories of practice—New inspiration for ecological economic studies on consumption”. In: *Ecological economics* 68.10 (2009), pp. 2490–2497.

[213] Bison Ivano, Veltri Giuseppe Alessandro, and Gaskell George. *D1.1: Early taxonomy of diversity*. Tech. rep. Written by the "WeNet - The Internet of US" (WeNet) project’s consortium under EC grant agreement 823783. 2020.

[214] Carl G Jung. “Psychological types. Collected works of CG Jung”. In: *Tr. HG Baynes. Rev. RFC Hull. Princeton: Princeton UP* (1971).

[215] Isabel Briggs-Myers and Peter B Myers. “Gifts differing: Understanding personality type”. In: (1995).

[216] David Dennis Lee Mascarenas. “A Jungian based framework for Artificial Personality Synthesis.” In: *EMPIRE@ RecSys*. 2016, pp. 48–54.

[217] Douglass J Wilde. *Teamology: the construction and organization of effective teams*. Springer Science & Business Media, 2008.

[218] Douglass J Wilde. *Jung’s personality theory quantified*. Springer Science & Business Media, 2011.

[219] Valdiney V Gouveia, Taciano L Milfont, and Valeschka M Guerra. “Functional theory of human values: Testing its content and structure hypotheses”. In: *Personality and Individual Differences* 60 (2014), pp. 41–47.

[220] Shalom H Schwartz. “Are there universal aspects in the structure and contents of human values?” In: *Journal of social issues* 50.4 (1994), pp. 19–45.

[221] Kirsi Tirri and Petri Nokelainen. “Identification of multiple intelligences with the Multiple Intelligence Profiling Questionnaire III”. In: *Psychology Science* 50.2 (2008), p. 206.

[222] Daniel Gatica-Perez et al. “Discovering Eating Routines in Context with a Smartphone App”. In: *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*. UbiComp/ISWC’19 Adjunct. London, United Kingdom: Association for Computing Machinery, 2019, pp. 422–429. URL: https://doi.org/10.1145/3341162.3349297.

[223] Fausto Giunchiglia et al. “Mobile social media usage and academic performance”. In: *Computers in Human Behavior* 82 (2018), pp. 177–185.

[224] Clemens Stachl et al. “Predicting personality from patterns of behavior collected with smartphones”. In: *Proceedings of the National Academy of Sciences* 117.30 (2020), pp. 17680–17687.

[225] Lakmal Meegahapola et al. “Generalization and Personalization of Mobile Sensing-Based Mood Inference Models: An Analysis of College Students in Eight Countries”. In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6.4 (2023), pp. 1–32.

[226] Karim Assi et al. “Complex daily activities, country-level diversity, and smartphone sensing: A study in denmark, italy, mongolia, paraguay, and uk”. In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI ’23. Hamburg, Germany: Association for Computing Machinery, 2023, pp. 1–23. URL: https://doi.org/10.1145/3544548.3581190.

[227] Paula Helm et al. “Diversity and neocolonialism in Big Data research: Avoiding extractivism while struggling with paternalism”. In: *Big Data & Society* 10.2 (2023), p. 20539517231206802.

[228] Gal Kobi and Segal Avi. *D4.2-Algorithms and implementation of diversity aware single user incentive design*. Tech. rep. Written by the "WeNet - The Internet of US" (WeNet) project’s consortium under EC grant agreement 823783. 2022.

[229] Nicholas Hoernle et al. “The phantom steering effect in Q&A websites”. In: *Knowledge and Information Systems* 64.2 (2022), pp. 475–506.

[230] Noemí Pavón et al. “Analyzing User Experience of the Chatbot SOS TUTORÍA UC”. In: *2023 XLIX Latin American Computer Conference (CLEI)*. IEEE. 2023, pp. 1–9.

[231] Laura Achón et al. “’SOS TUTORÍA UC’: A Diversity-Aware Application for Tutor Recommendation Based on Competence and Personality”. In: *2023 XLIX Latin American Computer Conference (CLEI)*. IEEE. 2023, pp. 1–10.

[232] William G Smith. “Does gender influence online survey participation? A record-linkage analysis of university faculty online survey response behavior.” In: *Online submission* (2008).

[233] Claudia Loebbecke and Arnold Picot. “Reflections on societal and business model transformation arising from digitization and big data analytics: A research agenda”. In: *The Journal of Strategic Information Systems* 24.3 (2015), pp. 149–157.

[234] Viktor Mayer-Schönberger and Kenneth Cukier. *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt, 2013.

[235] David M Berry. “The computational turn: Thinking about the digital humanities”. In: *Culture machine* 12 (2011).

[236] Laura Schelenz et al. “The Theory, Practice, and Ethical Challenges of Designing a Diversity-Aware Platform for Social Relations”. In: *AIES’21*. Association for Computing Machinery, 2021.

[237] EUCommission. *Research and innovation*. 2023. URL: https://research-and-innovation.ec.europa.eu/funding/funding-opportunities/funding-programmes-and-open-calls/horizon-europe/research-infrastructures_en (visited on 12/14/2023).

[238] Lakmal Meegahapola, Salvador Ruiz-Correa, and Daniel Gatica-Perez. “Protecting mobile food diaries from getting too personal”. In: *Proceedings of the 19th International Conference on Mobile and Ubiquitous Multimedia*. 2020, pp. 212–222.

[239] Fausto Giunchiglia et al. “A context model for personal data streams”. In: *Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data*. Springer. 2022, pp. 37–44.

[240] Mattia Zeni, Ilya Zaihrayeu, and Fausto Giunchiglia. “Multi-device activity logging”. In: *Proceedings of ACM - UBICOMP: Adjunct Publication*. 2014, pp. 299–302.

[241] de Götzen Amalia and et al. *D7.4-Formative Evaluation*. Tech. rep. Written by the "WeNet - The Internet of US" (WeNet) project’s consortium under EC grant agreement 823783. 2023.

[242] Miorandi Daniele et al. *D8.3-Exploitation and sustainability action plan*. Tech. rep. Written by the "WeNet - The Internet of US" (WeNet) project’s consortium under EC grant agreement 823783. 2023.