Department of Information Engineering and Computer Science

Bachelor's Degree in
Computer Science

Final dissertion

A research infrastructure for human behaviour studies

Supervisor

Fausto Giunchiglia

Student

Daniele Marcolan

Academic year 2020/2021

Contents

Introduction **3**

1 Overview of the Research Infrastructure **4**
1.1 The Data Science Challenge . . . 5
1.2 Data Lifecycle . . . 5
1.3 Main Objectives of the Research Infrastructure . . . 6

2 General Flow of Data and Data Management **8**
2.1 Compliance with Ethical and Legal Aspects . . . 8
2.1.1 Cross-Country Cooperation . . . 10
2.1.2 Sharing of Data with Non-EU Countries . . . 10
2.2 Data Collection Services . . . 11
2.2.1 Experiment Input Data . . . 12
2.2.2 Experiment Output Data and Results . . . 13
2.2.3 Research Infrastructure's Service Data . . . 13
2.2.4 Research Data for Later Use . . . 14
2.3 Data Preparation Services . . . 15
2.3.1 Data Minimization . . . 15
2.4 Cataloging and Storage of Data . . . 16
2.4.1 Making Data Findable . . . 17
2.4.2 Making Data Openly Accessable . . . 17
2.4.3 Making Data Interoperable . . . 18
2.4.4 Making Data Reusable . . . 18

3 Experiment Lifecycle **18**
3.1 Data Collection Services . . . 18
3.1.1 Data Collection Design . . . 18
3.1.2 Data Collection Execution . . . 22
3.1.3 Data Collection GDPR Compliance . . . 23
3.2 Data Elaboration Procedures . . . 24
3.3 Datasets Management . . . 24

4 Procedures for GDPR Approval **25**
4.1 Ethics Operation Steps . . . 26
4.1.1 Ethics Committee Approval . . . 26
4.1.2 Data Protection Impact Assessment . . . 27
4.1.3 DPIA for Data Processors . . . 27
4.1.4 Informed Consent . . . 27
4.1.5 Data Minimization . . . 28
4.1.6 Data Preparation . . . 28
4.1.7 Privacy Statement . . . 29
4.1.8 Request for Data Processor . . . 29
4.1.9 Experiment Execution . . . 29
4.1.10 Data Curation and Preservation . . . 29

5 Conclusions **30**

Bibliography **30**

Introduction

This thesis is grounded on a large international and multidisciplinary project called WeNet - The Internet of us, which aims at harnessing the diversity of people, using computer science, sociology and engineering. The WeNet main aim is to bootstrap an online virtual community where the diversity of its members is leveraged and exploited to improve their well-being and quality of interactions. In this framework diversity is assumed to be a key distinguishing feature of life, and it is defined as the variability that exists across humans and social relations, in terms of geographical locations or mobility constraints, personal or interpersonal skills, cultural, religious, economic, or social statuses, beliefs, attitudes, desires, or intentions. WeNet's innovative paradigm impacts human interactions in general, especially those that may benefit from a collaborative approach (creative industries, medical diagnosis, etc...). The WeNet consortium is developing a research infrastructure to exploit the project results and strengthen the european innovation ecosystem with a worldwide perspective, a research infrastructure which aims to empower machine-mediated diversity-aware social interactions. The WeNet platform will be the basis of a series of studies within universities across the world with diverse student populations. The studies will look at how the platform can improve students' quality of life inside and outside the academic environment.

The main problem to which this thesis tries to provide a solution concerns the management of documentation. There are many documents drawn up in recent years by various people within the project, but there is a lack of overall documentation, which gives a global idea of the project, starting from its grounding ideas and ending with the description of the macro processes that govern it. The current documentation results therefore chaotic and dispersive, and there is no conceptual model to refer to.

The proposal of this thesis is therefore to be a document that allows a global and high-level vision of the research infrastructure developed by the WeNet consortium - in order to give an organic description of its processes, and of the major challenges to be faced, especially in the management of the personal data collected - as well as being a starting point for anyone who wants to subsequently work on the project documentation. The secondary objective of this thesis is the comparison with the state of the art of european research infrastructures based on data collection, built on the model described above. The work done consists in the analysis and recollection of the aforementioned documents, with in mind the goal of creating a document that can be a starting point for the development of a unique conceptual model, within which the previous work can be relocated and where the subsequent work can be inserted. Starting from the analysis of existing macro-processes, this document can be a basis for works that analyze more in depth the research infrastructure. The objectives of this document, therefore, are:

- To show how this research infrastructure can represent a new paradigm for digital social interaction.
- To be a document that sums up the essential essential points of the project and identifies the macro-processes operating during the development of the research infrastructure.
- To highlight the distinctive features and philosophy, compared to the state of the art of the data-based research infrastructures.

The thesis is divided into four fundamental chapters, organized as follows:

1. The first is an overview of the research infrastructure, which represents the core of the WeNet project. It shows its philosophy, its purpose and its foundations, as well as describing the fundamental points of a standard research infrastructure.

2. The second concerns the process of collecting, managing and storing personal data, as well as the definition of ethical, privacy and security standards. This is probably the aspect of the research infrastructure that requires the most planning effort, as it will be the basis of every operation and experiment within the research infrastructure.

3. The third concerns the lifecycle of the experiments organized by the collaborating partners for the research infrastructure. It describes the basic steps for carrying out experiments that meet the criteria defined in the previous chapter.

4. The fourth concerns the steps necessary to comply with the European specifications of the law on the General Data Protection Regulation (GDPR).

1 Overview of the Research Infrastructure

Diversity permeates our everyday life and covers many dimensions, such as competence, culture, gender or economic across humans and social relations. Technology has evolved to a point where humans from diverse backgrounds, cultures, and experiences have an unprecedented ability to connect with each other. Yet technology does not in-and-by-itself provide support for developing and maintaining the social relationships that transcend geographical and cultural backgrounds. The key objective of the research infrastructure is to build a platform which addresses this gap by providing a diversity-aware, machine-mediated paradigm of social relations. The goal is connecting people that can support each other, and the key is leveraging their diversity. The paradigm includes a family of computational diversity-aware models supporting human interaction. Learning models construct diversity profiles based on people's past behaviour and interactions. A diversity-aware search builds upon these profiles to connect the matching people together. To support people's interactions, a diversity alignment mechanism lifts communication barriers to ensure that messages between humans are interpreted correctly, and a diversity-aware incentive mechanism generates incentives to motivate people to support each other. The entire paradigm is developed taking into consideration ethical guidelines. The platform provides the technological infrastructure to set out a series of studies that will be carried within universities worldwide with diverse student populations, and with the final goal of improving students' quality of life inside and outside the academic environment. Beyond universities, This innovative paradigm impacts human interactions in general, especially those that may benefit from a collaborative approach, such as creative industries or medical diagnosis. The WeNet consortium is developing a research infrastructure that will allow the exploitation of the project results and strengthen the European innovation eco-system in a worldwide perspective.

Diversity-aware platform design is a paradigm that responds to the ethical challenges of existing social media platforms. Available platforms have been criticized for minimizing users' autonomy, marginalizing minorities, and exploiting users' data for profit maximization. The consortium presents a design solution that centers the well-being of users. It presents the theory and practice of designing a diversity-aware platform for social relations. In this approach, the diversity of users is leveraged in a way that allows like-minded individuals to pursue similar interests or diverse individuals to complement each other in a complex activity. The end users of the envisioned platform are students, who participate in the design process. Diversity-aware platform design involves two main steps:

1. Defining a framework and operationalizing the grade of diversity of students.

2. Collecting data to build diversity-aware algorithms, overcoming ethical challenges encountered during the design of a diversity-aware platform.

1.1 The Data Science Challenge

Data Science is becoming a new technology driver and requires re-thinking a number of infrastructure components, solutions and processes to address the following general challenges:

- Exponential growth of data volume produced by different research instruments and/or collected from sensors.
- Need to consolidate digital infrastructure as persistent research platform to ensure research continuity and cross-disciplinary collaboration, deliver/offer persistent services, with adequate governance model.

The recent advancements in data technologies facilitate the paradigm change that is characterized by the following features:

- Automation of all processes regarding data, including data collection, storing, classification, indexing and other components of the general data curation and provenance.
- Transformation all processes, events and products into digital form by means of multi-dimensional multifaceted measurements, monitoring and control; digitising existing artifacts and other content.
- Possibility to re-use the initial and published research data with possible data re-purposing for secondary research.
- Global data availability and access over network for cooperative group of researchers, including wide public access to scientific data.
- Existence of necessary infrastructure components and management tools that allows fast infrastructures and services composition, adaptation and provisioning on demand for specific research projects and tasks.
- Advanced security and access control technologies that ensure secure operation of the complex research infrastructures and scientific instruments and allow creating trusted secure environment for cooperating groups and individual researchers.

1.2 Data Lifecycle

A standard model for the research infrastructures should support the whole data lifecycle and explore the benefit of the data storage/preservation, aggregation and provenance in a large scale and during long/unlimited period of time. Important is that this infrastructure must ensure data security (integrity, confidentiality, availability, and accountability), and data ownership protection. With current needs to process data that require powerful computation, there should be a possibility to enforce data/dataset policy that they can be processed on trusted systems and complying other requirements. Researchers must trust the research infrastructure to process their data on facilities and be ensured that their stored research data are protected from non-authorized access. Privacy issues are also arising from distributed remote character of the infrastructure that can span multiple countries with different local policies. Emergence of computer aided research methods is transforming the way how research are done and scientific data are used. The following types of scientific data are defined:

- Raw data collected from observation and from experiment (according to an initial research model).
- Structured data and datasets that went through data filtering and processing (supporting some particular formal model).
- Published data that supports one or another scientific hypothesis, research result or statement.
- Data linked to publications to support the wide research consolidation, integration, and openness.

Once the data is published, it is essential to allow the other researchers to be able to validate and reproduce the data that they are interested in, and possibly contribute with new results. Capturing information about the processes involved in transformation from raw data up until the generation of published data becomes an important aspect of scientific data management. Scientific data provenance becomes an issue that also needs to be taken into consideration by research infrastructure's providers. Another aspect to take into consideration is to guarantee reusability of published data within the scientific community. Understanding semantic of the published data becomes an important issue to allow for reusability, and this had been traditionally been done manually. Learning from best practices in semantic web community on how to provide a reusable published data, will be one of consideration that is addressed by the research infrastructures. Data are typically distributed both on the collection side and on the processing/access side: data need to be collected (sometimes in a time sensitive way or with other environmental attributes), distributed and replicated. Linking distributed data is one of the problems to be addressed by the research infrastructures.

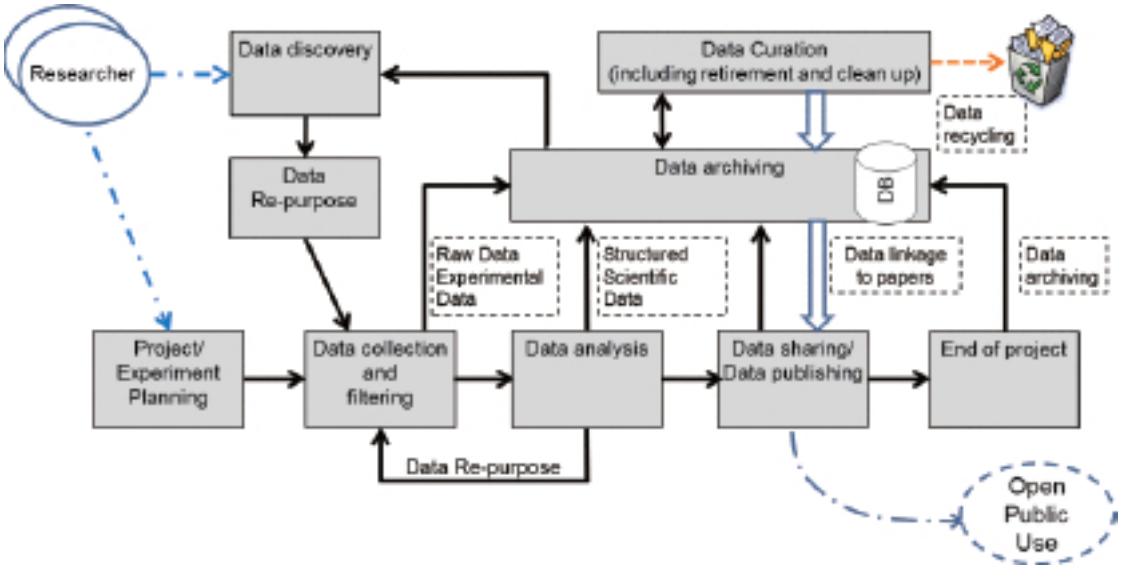

Figure 1.1: A general model of data flowing in a data-based research infrastructure

The generic scientific data lifecycle includes a number of consequent stages: research project or experiment planning; data collection; data processing; publishing research results; discussion, feedback; archiving (or discarding). A general lifecyle of gathered data is shown in Figure 1.1. New scientific data lifecycle management requires data storage and preservation at all stages what should allow data re-use/re-purposing and secondary research on the processed data and published results. However, this is possible only if the full data identification, cross-reference and linkage are implemented in the infrastructure. Data integrity, access control and accountability must be supported during the whole data during lifecycle. Data curation is an important component of the scientific data lifecycle management and must also be done in a secure and trustworthy way. Support data security and access control to scientific data during their lifecycle: data acquisition (experimental data), initial data filtering, specialist processing; research data storage and secondary data mining, data and research information archiving.

1.3 Main Objectives of the Research Infrastructure

The main goal of the research infrastructure developed by the WeNet consortium is to develop the culture, science and engineering methodologies, algorithms, social interaction protocols for an online platform that will empower machine mediated diversity-aware interactions between people. The project aims to verify the diversity among students based on social practices - intended as the set of skills, knowledge and motivations, both between people within the same university and between universities in different countries - through the development of new techniques of measurement. The research infrastructure ultimate goal is to encourage diversity between people by promoting social

relations and to develop of a platform that favors social relationships through artificial intelligence systems. In other words, through technology, it wants to create a medium that favors interaction between people, encouraging mutual support practices. The basic assumption is that people with more skills in one field are more likely to help others with needs in that specific field. The double objective in this case is to build, on the one hand, an optimized relationship model and, on the other, to provide the best relationship solution.

The platform is a software framework that allows application developers, innovators and Web entrepreneurs to quickly and easily develop and deploy diversity-aware applications. The first application developed allows you to organize a meeting to have lunch together, sharing meals around the world, and to discover potentially interesting available social meal events to apply to. The app uses the resources and functionality exposed by the platform to provide users with a personalised experience based on diversity dimensions. The app was developed using Telegram as an enabling platform. It was therefore designed and implemented as a chat application, powered by a chatbot engine implemented as a deterministic automaton developed in Python, allowing the users to create and manage shared meals and to apply to become a participant of an existing shared meal. The core of the application is the Research Infrastructure, which is designed to managing the full lifecycle of data generated in the WeNet pilot experiments (collection, processing, storage and access). The Research Infrastructure directly hosts data and metadata about experiments in one cloud location, after having anonymized them.

The project will run experiments with students in 18 pilot sites worldwide, involving 10,000 participants throughout the whole duration of the project. Generally, the purpose of the research infrastructure is the development of the scientific foundations, methodologies and algorithms empowering machine mediated diversity-aware people interactions. This can be further characterized as:

- Development of a computational sociological theory of diversity.
- Development of diversity aware interaction protocols. These protocols will establish:
  - The engagement and productivity of the participants, based on a diversity-aware theory of incentives.
  - A common taxonomy of interactions, where the machine will always try to minimize the human cognitive load while still minimizing the probability of a mistake in the interaction.
  - A common framework for sharing the local context, which will allow to properly enact the interactions.
- Development of diversity-aware algorithms and tools for learning the user embodiment in the world.
- Development of diversity-aware algorithms and tools for learning how to recognize and evolve the users' social interactions.

In other words, the idea is to test the concept and technology extensively and at scale in a rich and relevant application domain (University and adult schools' student life). This will provide empirical evidence of how to use the platform in general and provide feedback on the acceptance and perceived usefulness of the used technology. The intended outcomes for individuals regards some of the known issues that students deal with, that include: health (stress, sleep, poor eating, homesickness), time management (how to find a balance between studying, socializing, resting), socialization and integration, pressure to perform academically and space issues (limited physical space to study/work). The two themes that will be the focus of the WeNet Smart University pilots will be:

- The need for students to develop and maintain healthy habits (health is used here in a large sense, including mental and physical)
- The need for students to play multi-faceted social roles and to negotiate affiliations to multiple groups.

2 General Flow of Data and Data Management

The project will run several pilot trials with users that will serve multiple purposes. From a data perspective, the proposed approach develops and follows a methodology that involves five main steps:

- Collection of data.
- Preparation of data.
- Data analysis in the local project experiments.
- Sharing and treating data in the project experiments and research through the Research Infrastructure.
- Reusing and transferring data toward the scientific community through the Research Infrastructure.

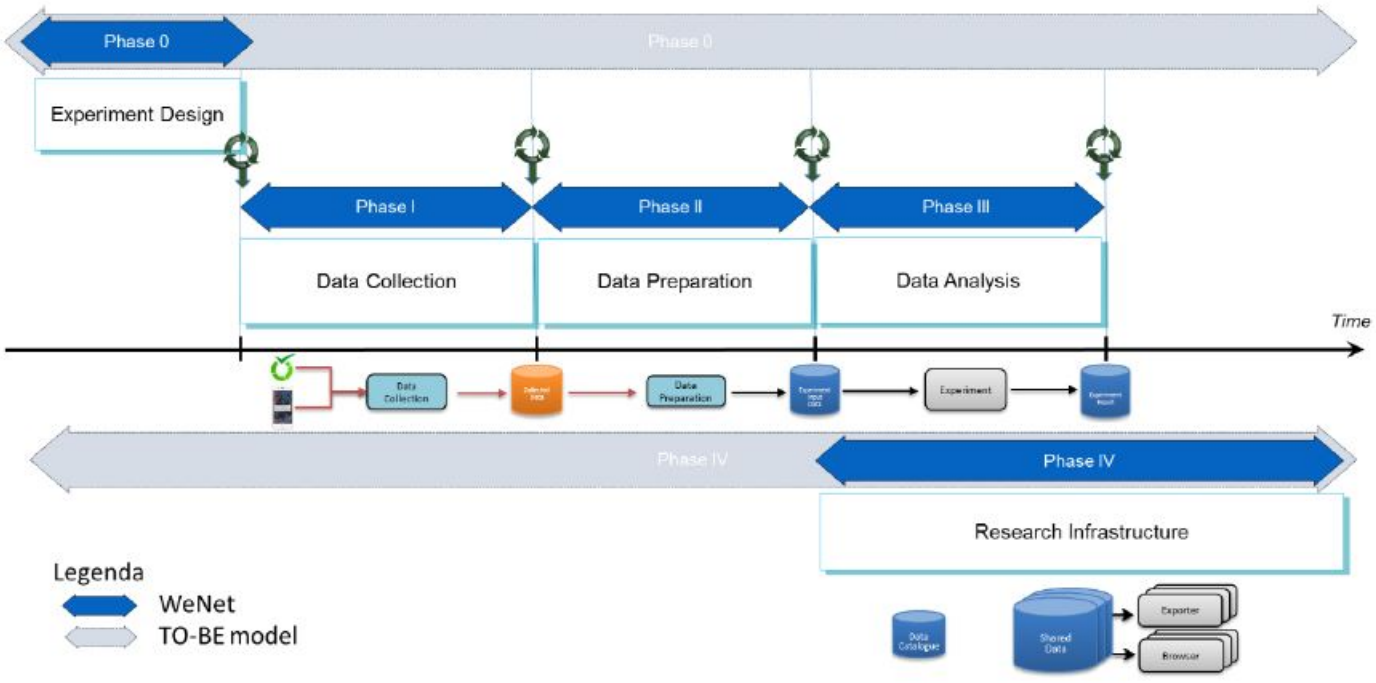

Figure 2.1: The data management process

2.1 Compliance with Ethical and Legal Aspects

The research project will comply not only with the General Data Protection Regulation (GDPR) from the European Union, but also will adhere to high ethical standards for data management. Part of the data collected within the project will be personal data and, as such, subject to GDPR. The processing of data will follow the following six principles mentioned in paragraph 1 of GDPR Article 5:

1. Lawfulness, fairness and transparency.

2. Purpose limitations.
3. Data minimization.
4. Accuracy.
5. Storage limitations.
6. Integrity and confidentiality.

Data and privacy protection are a central issue when speaking about the ethics of research activities in Europe. In research settings, data protection imposes obligations on researchers to provide research subjects with detailed information about what will happen to the personal data that they collect. It also requires the partners processing the data to ensure that the data are properly protected, minimized, and destroyed when no longer needed. One of the best ways to mitigate the ethical concerns arising from the use of personal data is to anonymize it so that it no longer relates to identifiable persons. Data that no longer relates to identifiable persons, such as aggregated and statistical data, or data that has been otherwise rendered anonymous (the data subject cannot be re-identified), are not personal data and are therefore outside the scope of data protection law (GDPR).

The project should collect only the data needed to meet its research objectives. Collecting unneeded personal data may be deemed unethical and unlawful. Before start any trial, pilot, experiment or study within the project the partners should conduct a data minimization review to ensure that data are collected on a 'need to know' basis and report it in order to document that the principle was properly considered. Personal data processing requires free and fully informed consent from the involved persons. In the project, all participation must be voluntary. As such, the data controller and related partners must (in advance) obtain and clearly document the participants' informed consent. The collected data must be kept in a form that enables the data subjects to be identified for a period not exceeding what is necessary for the purposes for which they are processed. Data retention occurs for a specified time period, based on the legal agreements in force and the business needs. Furthermore, as an European Project, WeNet embraces the "right to be forgotten" that is also part of GDPR. The possibility of stopping participation in the experiment and requesting the deletion of their personal data will remain present for all individuals and for the entire duration of the data collection and data maintenance procedures. The data subject may also request their data to be deleted after the end of the project. These objectives will be validated and authorized both by the ethics committee and by the privacy office present in each university. In addition, the project has equipped itself with documentation representing the minimum requirements for participating in an investigation. This consists of:

- Ethics committee approval: about the objectives, methods and resources used in the current phase of research.
- Legal office approval: about the documentation, the participant's information and the privacy measures. This contains also:
  - Informed Consent.
  - Privacy statement.
  - Information about the experiment execution.
  - Data preparation.
  - Request for Data Processor.
  - Data curation and preservation.
- Data Protection Impact Assessment: to assess, identify and minimise risks that may result from data processing.
- Data Protection Impact Assessment (DPIA) for Data Processors.
- Data Minimization: a definition of all the steps done by the data controller concerning the minimization process.

2.1.1 Cross-Country Cooperation

Because of the international nature of the project, data are transferred within the research infrastructure partners to recipients in other countries both European and non-European, which represent a particular ethical challenge. The type of data transferred is anonymized, therefore the GDPR does not impose any restrictions on the transfer of such data within the European Union, either to third countries or from third countries. This means that the data collected by the extra-EU partner can be transferred to the EU partner and vice versa, provided it is anonymous. Anyway, understanding 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. Looking at different cultures, we can conclude that there is much diversity with regard to privacy norms and behavior in the world. Being located in the European context, the project certainly adopts a "western" perspective on privacy and has to comply with the GDPR standards. The platform created within the project will then build on principles such as privacy by design, data minimization, and the protection of individual user's data. Anyhow, diversity exists also with regard to privacy norms. Such diversity will not be reflected in the WeNet technology. Therefore, the consortium should also be aware that the use of the platform in cultures outside of Europe or the West might lead to unexpected consequences, precisely because norms and behaviors with regard to privacy differ.

2.1.2 Sharing of Data with Non-EU Countries

The project involves, especially in the first phase, the sharing of data from third countries. The research includes middle-income countries, which will mutually receive various benefits from participation in the project based on sharing of the collection platform and data collection; know-how sharing, and collaboration in the production of scientific research. The involvement of Non-EU countries concerns both legal aspects and process aspects:

- Legal aspects: the first issue is related to the non-applicability of the GDPR outside Europe. This is especially true for non-EU Project partners, who do not receive direct European funding and therefore are not required to comply with the regulations in force in Europe. For this reason, the project provides that non-EU members will be compliant with European regulations as legally binding in the signed Grant Agreement. The enforcement on these countries will be based on the procedure and documents produced in the context of European legislation, with measures that safeguard local law. Therefore, each non-EU member must, initially, indicate a local ethics committee that supports and approves the actions carried out within the project, like the other EU members. If at some universities and non-EU research centers the figure of the ethics committee should be absent, they will take charge of nominating one. It will then be a request for a confirmation that the activity could have been performed legally in an EU country. Here, it should be emphasised that the experiments have already obtained the approval of the ethics committee of the University of Trento. Regarding the scope of data collection in pilots and subsequent experiments, each of the members in charge of performing them will be designated as Data Controller. All data controllers are also required to make a DPIA; unlike the EU members who will present the DPIA to their DPO (Data Protection Officer), non-EU members, if they do not have a similar office, will insert the DPIA among the project documents. The regulations in force in each country, where they do not conflict with the GDPR, will be considered as supplementary to the GDPR or to the regulations deductible from the Grant Agreement. This will happen in any action spheres of the project, including data protection.

- Process aspects: in regards to the initial phase of data collection, complying with the transparency criterion, privacy information forms are created and distributed. Information from European citizens in non-EU countries will not be collected, due to the limit of the Grant Agreement and the impossibility of imposing European law in non-European countries. Once the data collection procedures have been approved and the data collection has been completed, the data will be processed for the first time exclusively by the Data Controller who will be - directly or indirectly - responsible. In this phase, data will be manipulated with personal contents, which will be separated from the main database and carefully kept. After this first filtering phase, the correctly anonymized data can be shared. The develop of the research infrastructure is

experiment-based and it is becoming increasingly data-intensive over its life. The local partners produce a significant amount of data from data collection campaigns with students. The project allows the utilization of data in a comparative analytical way and enables the analytic outputs to be shared across different partners. The project involves low and middle-income countries, and it is concerned with guaranteeing fair benefit-sharing arrangements with all stakeholders. Furthermore, the consortium needs to allow all research-partners' data based experiments to use large scale data from different countries. To address these challenges, the data management policy is based on the following two key points:

- the local PIs retain absolute control and ownership of their collected data.
- no personal data is moved between the partners.

The local PIs ensure secure personal data collection and protection for the personal collected data; the local PIs also maintain and control the datasets with personal information. Local partners will remain the controllers of their collected data and be bound by local statutes appropriate anonymized datasets can then be brought outside the local partner jurisdiction for experiment analysis and research. Throughout this process, the experiment specification acts as an overall plan for the experiments, taking the form of an agreement between the partners involved and providing transparency of process. All the available anonymized data sets are listed in the research infrastructure catalog and they can be requested by the different partners. The local PIs have the responsibilities necessary to deliver a successful Data Governance for their collected data. They have the decision making authority for this specific subject matter of data, their responsibilities include:

- Data definer (define the data, how they will be used and how they will be managed).
- Data producer (producing, updating, deleting and archiving the data that will be managed).
- Data user (using data to perform processes and experiments)
- Maintain quality and data integrity.
- Define data access levels.

The partners will utilize analytic techniques, and research processes, by which anonymized data can be utilized effectively in the production of analytic outputs; which in turn can be shared across partners for research purposes. The accountability for the governance of personal data will be retained within the local PIs. During the initial phase of the Project, non-EU countries will share only anonymized data to Europe. Given than this will become much less sensitive non-personal data, the following of the respective local laws will be deemed sufficient for their handling.

2.2 Data Collection Services

The main source of the data for the project will essentially be students. Data will be generated by students through sensors active on their smartphones and through questionnaires filled in periodically. The datasets collected in this way are further processed for classification in a number of more focused categories and table schemas. This processed data conforms to the needs for reuse data classification of incoming data streams. The main purpose of the datasets thus created is to maintain an evolving experiment data collection that is needed for:

- The development of the scientific foundations, methodologies and machine learning algorithms empowering machine mediated diversity-aware people interactions.
- The validation of the apps developed on top of the platform.
- The development of a computational sociological theory of diversity.

A basic scenario of data collection and processing by the research infrastructure is shown in Figure 2.2. The scenario refers to research which involves processing of personal data, regardless of the method

used (interviews, questionnaires, direct online retrieval etc.), where "personal data" means information relating to an identified or identifiable natural person, and "processing of personal data" means any operation (or set of operations) performed on personal data, either manually or by automatic means.

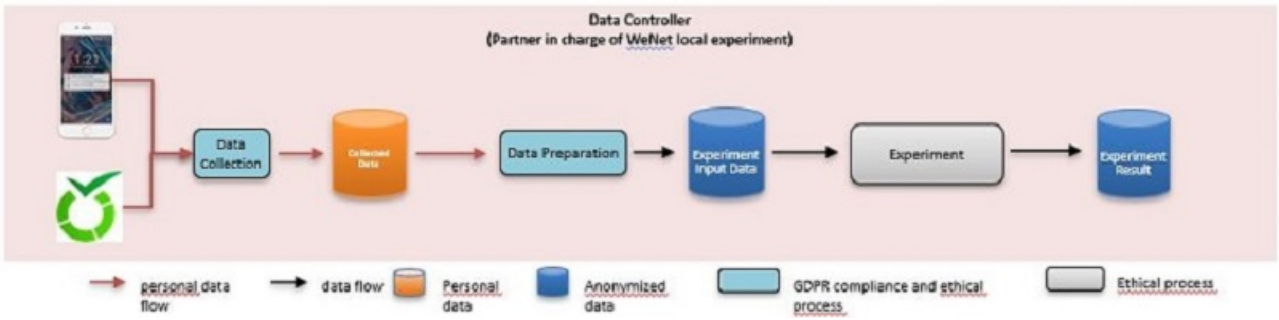

Figure 2.2: Basic scenario of the main set of operations

An example of collected data is the Smart University pilot at ABC University, which includes the collection and processing of personal data. In particular geo-coordinates, sensor data and other data provided by volunteering individuals via apps installed on their smartphone. This information is collected, for the purpose of developing a methodology that enables the quantitative characterization of behavioural patterns that correlate with health issues that may impact the lives of the students. The collected data regard personal characteristics (name, gender, age), questions about daily activities (location, mode of transport, persons being with you, mood) and sensor data (position, acceleration, temperature, etc).

2.2.1 Experiment Input Data

The collected data may be used by the Data Controller and requested by the partners for performing experiments within the project. To satisfy these requests and comply with regulations, the experiment input data will be generated from the collected data and shared. The data preparation operation to generate the experiment input data from the collected data will be carried out under the responsibility of the original data controller of the collected data. The need for personal and sensitive data will need to be duly justified and approved by the data controller before the experiment team may be given access. The process for preparing the experiment input data includes different steps towards making sure that all of the experiment data is relevant and limited to the purposes of the experiment (in accordance with the data minimization principle). The experiment datasets are then stored in IT backbone of the research infrastructure. Input experiment data is generated within the infrastructure as shown in Figure 2.3.

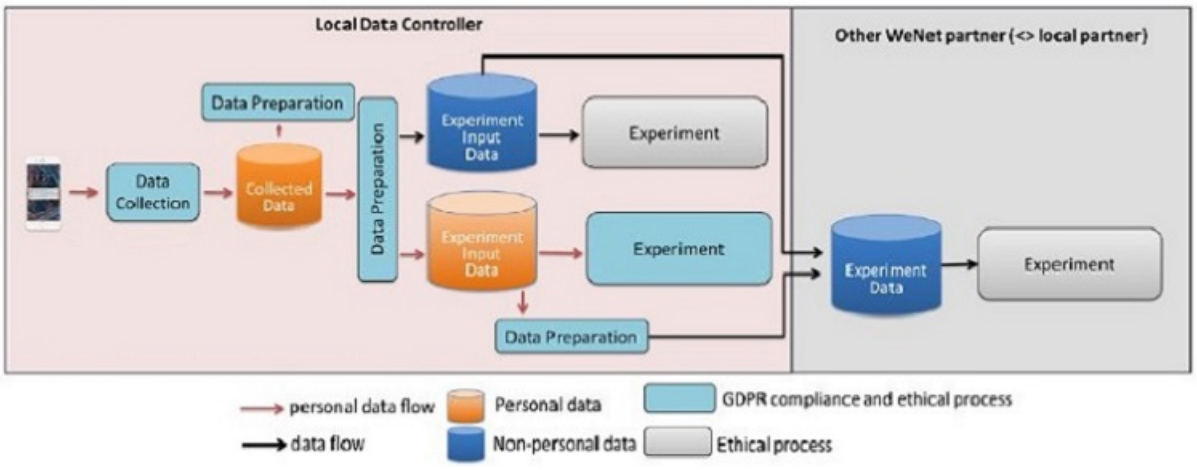

Figure 2.3: Generation of experiment input data in the local controller and in project partners

In the Smart University pilot at ABC University scenario, it is established that to achieve that

objective the processing of some personal data provided by volunteering students is necessary. And thus, to generate the experiment input data, the collected data is filtered to include only what is strictly necessary for the purposes of the experiment and also pseudonymized to limit the impact of the remaining personal data.

2.2.2 Experiment Output Data and Results

The resulting data from the different experiments that were carried out by using the previously defined experiment input data are then cataloged. The pilot trials are designed under a common generic methodology to be adapted and implemented under local conditions and the specific needs of the involved users' needs. The evaluation team will follow all piloting activities, in order to work out:

- The formative evaluation, which provides feedback to the project team, on how to improve the service in the course of the project.
- The summative evaluation, which will measure the impact of the project on society, on universities and on students' life.

The formative evaluation will run in parallel with the pilots, the summative evaluation will run once the services are tested in all the universities inside and outside the consortium. The evaluation team, in collaboration with the technical staff and service designers, will define the evaluation criteria and modalities and will identify key values to assess. The team will also define qualitative and quantitative indicators to assess the scenarios with respect to both the formative and summative evaluation. Finally, the evaluation team will create and share with the rest of the project a protocol for the result data collection. This data will be used in order to assess the impact of the carried out activities and the whole project (summative evaluation).This methodology and the organization and conduction of the pilot experiments will generate original digital content (generated output data and experiment results). The experiment results are then stored in IT backbone of the research infrastructure. Output experiment data is generated within the infrastructure as shown in Figure 2.4.

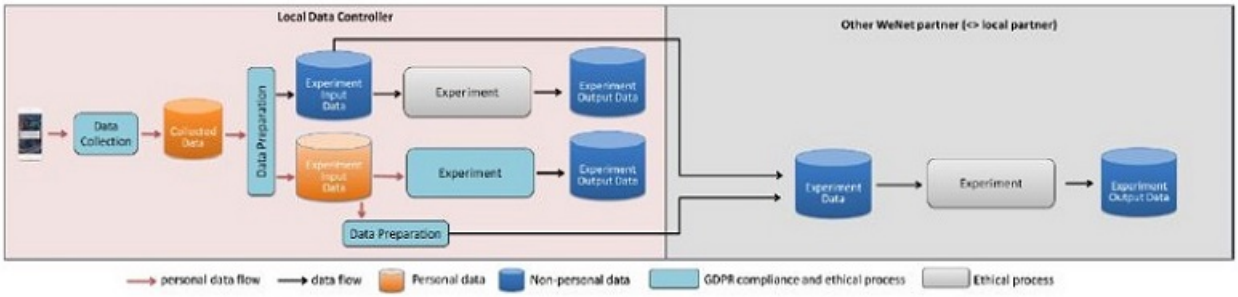

Figure 2.4: Generation of experiment output data in the local controller and in project partners

The purpose for the collection and processing of the data in the Smart University pilot at ABC University is the development of a methodology that enables quantitative characterization of behavioural patterns that correlate with health issues that impact the lives of the students. The experiment data generated in previous steps is therefore separated in 8 groups of individuals where each group is analysed using specific statistical indexes. Such analysis serves as the means by which the researches extract information from the experiment input data. Statistical analysis and calculations are used to summarize the observations, to estimate the variance and to estimate the probability that the underlying phenomena are detected or not; the output data are created by the results of calculation and the inference process.

2.2.3 Research Infrastructure's Service Data

Service data refers to information collected not directly towards experiments but towards capturing data that may be used to improve the technical or procedural details related to the correct execution of the project; this may include the production of training datasets for machine learning algorithms, and participation or engagement metrics that will help the better and more effective organization of future pilots. Service data is generated within the infrastructure as shown in Figure 2.5.

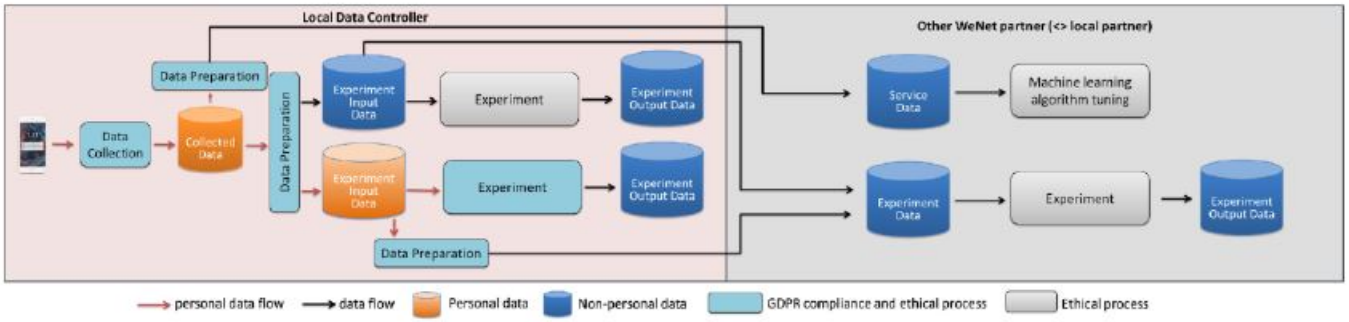

Figure 2.5: Generation of service data within the platform

The generation of service data by the local data controllers for a given dataset specification (provided by another partner) may provide additional information that will aid to the correct design of the global pilot. Furthermore, working with service data can help the partners get a better intuition of the kind of data is available and how they may be able use eventually to answer to the project requirements. The generated service data and information will be presented on the WeNet website platform and allow the evaluation of the pilot trials performance over the project's timeline. These will be accessible by the interested public in a structured visualization form but mainly they will be exploited by the project's consortium to monitor the platform's effectiveness, review and revise best practices and enhance the community building strategy. Service data must be guaranteed to not allow any inference about the collected data (it other words, it have to be summarized, randomized or anonymized) to avoid any risk of leaking sensitive information without approval or GDPR compliance. The Smart University pilot at ABC University provides some of the collected datasets so that the research infrastructure is able to reuse to train the machine learning algorithms that will be used throughout the project. Anyway, to avoid sharing personal information (that is not even necessary for this particular purpose) the Local Data Controllers from the Smart University project perform a data preparation activity to completely anonymize the collected data into service data, that other partners are able to safely use for the target training activities.

2.2.4 Research Data for Later Use

The project will run experiments with students throughout the whole duration of the project. Research infrastructure's data will be collected at several worldwide locations, namely through local pilot trials in 18 university sites carried out in several waves. These datasets will be directly obtained through data collection campaigns and then prepared for its reuse by the experiment team (Local Data Controller) and all the information for reusable data finally collected into a Data Catalogue as shown in Figure 2.6.

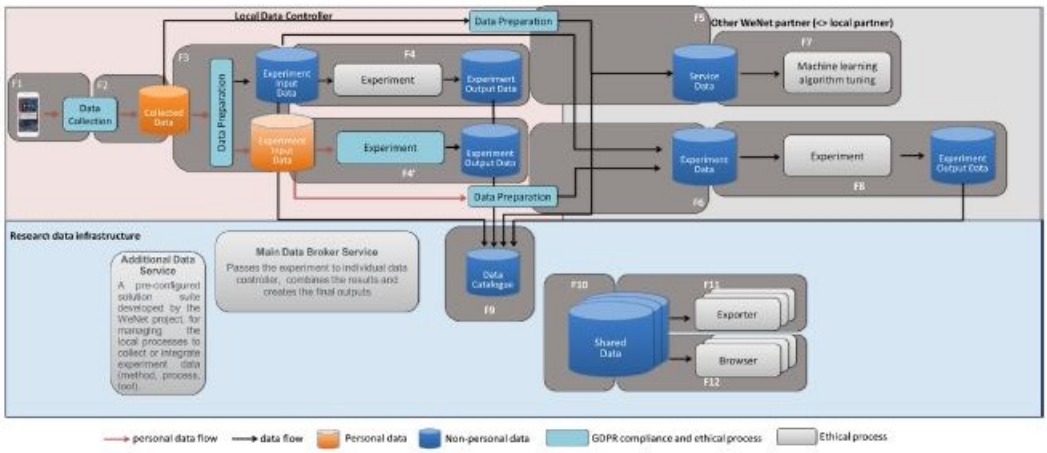

Figure 2.6: The complete diagram of flow of data

This leads to the federated Research Data Infrastructure, were the resulting datasets from the local pilot trials are used for research activities that may be beyond the original scope of those trials. This requires the defining of data sharing methods to clean, anonymize and summarize data to enable it use by researchers. Furthermore, this helps the consortium to achieve its objectives of using data as a basis to establish reproducibility for the large scale pilot trials, while ensuring value and future transfer of data toward the scientific community.

2.3 Data Preparation Services

Regarding data processing, the data controller assignments are essentially three: cleaning, pseudonymization and anonymization. As far as cleaning is concerned, this is divided into a complex of procedures aimed at improving the quality of the data collected. Data cleaning concerns several aspects. The first is the definition of data labels, so that they can be easily understandable and usable. A second aspect concerns the recognition of missing data and their reporting and imputation, in order to have a dataset as robust as possible and therefore presenting the least possible bias. The third concerns the aggregation of all the collected variables that concern the same concept, in order to be able to analyze them in statistical terms. These operations end with the creation of metadata that can be used by researchers and useful for orienting themselves and selecting and requesting the necessary data for their research. Another fundamental step is the pseudonymization procedures, the removal of the participant ID and replacement with an alpha-numeric code. Finally, various procedures take place to remove all participants' personal information (and replace it with anonymous information), in order to be able to analyze and share data in a GDPR compliant way. If it does not have sufficient resources, the Data Controller may make a possible appointment for a data processor not external to the consortium (through a specific GDPR compliant procedure).

2.3.1 Data Minimization

Data Minimization is a principle that states that data collected and processed should not be held or further used unless this is essential for reasons that were clearly stated in advance to support data privacy. Personal data, which is collected and otherwise processed, should be adequate, relevant and limited to what is necessary in relation to the purposes. Moreover, minimising the amount of collected data is positive in many aspects: it may facilitate storage, monitoring and updating and cleaning operations, and may also help mitigating or minimising other risks such as data leaks or other data breach incidents.

As a cardinal principle, towards making the platform as unobtrusive as possible for those who will use it, the chosen approach is to minimize the collection of data as much as possible. During phase 1, the pilots that is carried out in the course of the project will determine the minimum number of variables that allow to precisely define the diversity and avoid stereotypes. Furthermore, the subset of information chosen to be collected in an attempt to define a diversity model, is already the result of a careful selection of variables deriving from the implementation of concepts typical of theories on social practices. This means that this first theoretical reflection is already able to eliminate the type of information to be collected. This principle will be enforced during the platform and the data infrastructure phases. A second minimization of the data to be collected derives from the previous experiments. The analysis of these results will allow us to select the necessary data for the model and exclude the unnecessary. Another way of minimization derives from the methodological approach to survey. For example, in the selection of suitable collection instruments that have already been tested - questionnaires and questions whose effectiveness is certainly known. This guarantees the minimization of the necessary questions in surveys and tools and, therefore, of the information gathered.

Each member (university) which participates in data collection and analysis, is defined as a Data Controller. As such, it undertakes to follow the different phases:

1. The Data Controller takes care of defining its research objectives and methods of pursuit. Being part of the consortium and being the comparison one of the key issues to define the diversity of people, in this first phase many objectives are shared by the different members. For this reason, there is a similarity in the procedures and data collected. Each Data Controller is required to present the set of objectives and procedures to pursue them at its university.

2. If the complex of procedures is approved by the university, the member will be able to participate in data collection. In other words, only if the data collected are defined as adequate, relevant and limited to the purpose (both by the ethics committee and the legal department), the member will be able to proceed in the subsequent stages.

3. The data collection is be a very delicate phase, in which the participants are particularly exposed. That's why the data protection and security measures are defined, as well as transparency and accountability measures, which also pass from the information provided to the participants themselves.

4. Once the data has been collected, each member - with the possible help of data processors internal to the consortium - takes care of processing them. This means that the data is cleaned, prepared, harmonized and correctly anonymized - with different procedures according to requirements. The data controller must perform all the measures and algorithms aimed at making the data protected and correctly anonymized, in order to be able to save, share and analyze them.

5. The data controller undertakes to maintain the data, in accordance with specific protection and access measures, including those for updating information and deleting data that are not considered adequate, relevant or limited to the purpose of the member or the consortium.

6. The correctly anonymized data may be shared between the members of the consortium, upon request containing the specification of the need to obtain the data and objectives. These can then be analyzed, following WeNet's objectives and conforming to the shared code of conduct.

The complete data management process operated by the data infrastructure is shown in the Figure 2.7.

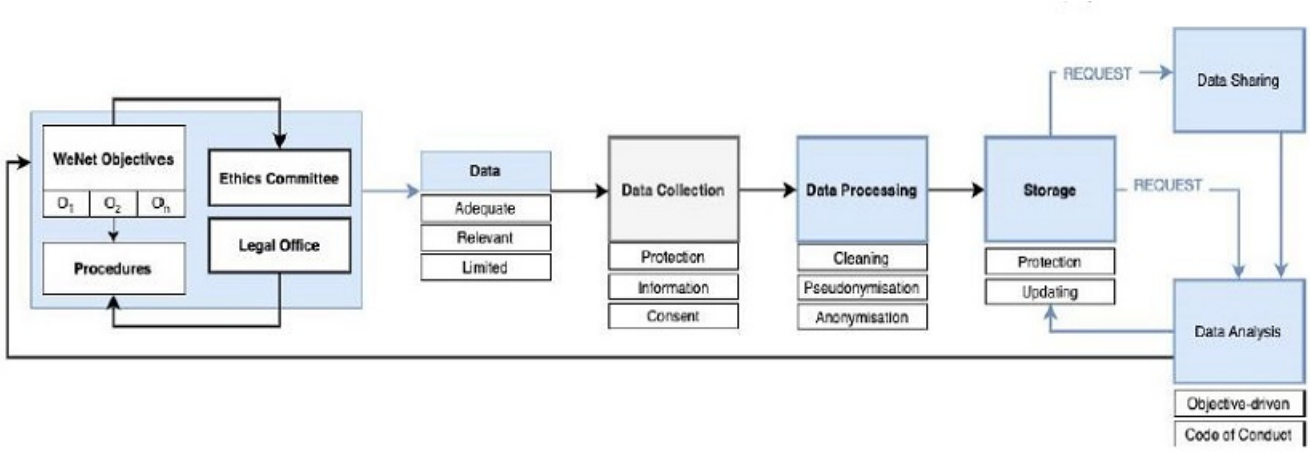

Figure 2.7: The complete data management process, complying with the principle of data minimization

2.4 Cataloging and Storage of Data

Generation, curation and preservation of datasets has to be done in compliance with current European and local legislation. Correctly processed data will be saved in the cloud. The maximum data retention period is 5 years from the end of the project. Once the set data preservation period expires, appropriate curation methods will be implemented accordingly; including archiving, final purge, destruction, anonymization, etc. The selection and storage of this information is based on the principle of minimization, considering both the risks associated with this data and the risks of having data breaches, or otherwise risk accidentally doing harm to data subject. Therefore, any personal and non-personal information that proves to be not useful for the research will be deleted and no longer collected. The security measures will be adequate, updated and include the limitation of retention periods, which will not be extended beyond what is necessary. In addition, the saved data will be

updated according to the results of the various analysis. This implies that any resulting unnecessary or obsolete data will be deleted.

To discover, access and reuse quality datasets the European Commission has launched at the end of 2013 a flexible pilot for open access to research data that aims to improve and maximize access to and the re-use of research data generated by Horizon 2020 projects, which takes the the name of 'FAIR' (Findable, Accessible, Interoperable, Reusable). The guidelines emphasise machine-actionability (the capacity of computational systems to find, access, interoperate, and reuse data with none or minimal human intervention) because humans increasingly rely on computational support to deal with data as a result of the increase in volume, complexity, and creation speed of data. The FAIR principles are:

1. Findable: the first step in (re)using data is to find it. Metadata and data should be easy to find for both humans and computers. Machine-readable metadata is essential for automatic discovery of datasets and services, so this is an essential component of the process.

2. Accessible: once the user finds the required data, he needs to know how to access it, possibly including authentication and authorisation.

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: the ultimate goal of FAIR is to optimise the reuse of data. To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings.

2.4.1 Making Data Findable

Data will be made accessible also through Zenodo or similar open data repositories. Zenodo is part of OpenAIRE (an acronym for "Open Access Infrastructure for Research in Europe") and it provides a repository for those researchers who do not have an existing institutional or thematic repositories where they can deposit their publications and data. The specific repository to store the project will be decided after a careful analysis. The considered factors for that analysis will encompass a number of qualitative issues that may completely override the numerical analysis (cost, capacity, expertise, investment, scalability, strategic importance). The following reasons will be taken into account for the selection of the repository to store the research data:

- Computer resources, scalability, replication, multiple data centers, proximity, quick data access.
- The keeping of both data files and metadata online.
- Long-term preservation.
- Digital Object Identifier (DOI) support and versioning.
- Support of a variety of types of files.
- The following of standards like the Open Archival Information System reference model.
- Support of metadata and catalogues for data discovery.

2.4.2 Making Data Openly Accessable

The consortium strongly believes in the expansion of the community generated by the project fuelled by the creation of a federated Research Infrastructure. This federated Research Infrastructure will include large amounts of data on human behaviour and social interactions and will make openly available the generated and processed datasets from the piloting activities. When personal data are involved, in order to be GPDR compliant, anonymization techniques will be utilized before the datasets are made public. The availability of the data will be ensured by utilizing an appropriate research data repository, along with relevant documentation and linked metadata for each set. Whenever possible data will be openly accessible and will be licensed under permissive licenses so will be available to everyone interested.

2.4.3 Making Data Interoperable

The consortium in its effort to make the project's datasets as accessible and interoperable as possible will seek compliance with open standards. More specifically, data will be stored using open data formats and implementing open standards that will follow best practices and guidelines for working with open data. Data stored will be accompanied by relevant metadata to ensure re-usability of the data by third parties. The consortium aims to document its research data in a way that ensures they can be interpreted, shared and reused by the scientific community. The initial assumption is that data will be preserved in the infrastructure belonging to the project partners until the end of the project. Metadata files will be created and linked with each dataset to facilitate their discoverability and usability over time.

2.4.4 Making Data Reusable

The consortium, in its effort to make the project's datasets accessible and interoperable, will publicly release aggregated and anonymized information from the pilot activities. The data re-use will be fuelled by the inception of a federated Research Infrastructure, which will include large amounts of data on human behaviour and social interactions (as collected from the pilot trials), WeNet software tools for data collection and analysis, online training material and protocols for running experiments and trials. Furthermore it will allow other institutions, innovators, and policy makers from Europe and beyond to join and reuse the WeNet software tools and research data.

3 Experiment Lifecycle

Several experiments are conducted in different locations. The experiments aim to understand how the organization of the students of different bachelor's degrees affects their academic performances. The hypothesis behind experiment is that patterns in the behavior of the students and their ability to organize themselves through time and space affect their academic performances. In fact, empirical researches have shown how students' time management ability and its translation into time allocation between academic and other daily are important aspects that have an impact on students' performance. The analysis has been conducted using personal data about the users collected from their smartphone.

3.1 Data Collection Services

The experiments conducted and the data collection processes are grounded on iLog and LimeSurvey. iLog is a tool able to collect user's personal information and to generate streams of data from smartphone's integrated sensors and attached wearable devices, while LimeSurvey allows to collect data through questionnaires compiled periodically.

3.1.1 Data Collection Design

There are 3 phases of experiment design:

- Phase 1: The actual experiment, from its design, to its preparation and actual running.
- Phase 2: The preparation of the data, to make them Big Data, passing through data cleaning (the consolidation of the data), integration with third party sources and the sharing of collected data.
- Phase 3: The analysis to be carried out on the data, obtaining a final dataset from a subset of collected data.

The 3 phases iterate during the process of experiment and give meaningful feedback about requirements to each phase, as shown in Figure 3.1. In the first iteration of the experiment, data were divided as follows:

- Smartphone-based data: all sensor data in accordance with the smartphones' specifications.

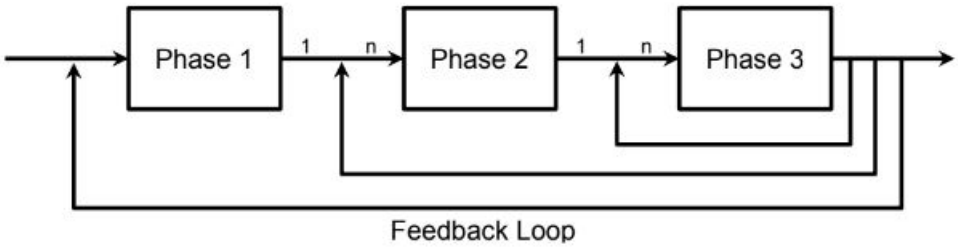

Figure 3.1: The feedback loop

- Survey-based data: psychological questionnaires and demographics obtained from the students.
- Third-party data: GPA and ECTS from the administrative offices of the University of Trento.

The experiment was conducted among the students of the University of Trento, who regularly attended classes and owned an Android device. The way to incentivize students was designed to be monetary. A reward of 7 euro was to be given to each participant. In addition, three cash bonuses of 150 euros for the three most deserving students were set. These three subjects were identified on the basis of three criteria:

- How long the application had remained in operation.
- How often did the GPS, Bluetooth and Wi-Fi sensors turned off.
- How many answers had been provided through the questionnaire.

Generally speaking, the sample of experiment consists of university students. The sample was composed of the students which reflected the general population of freshman year students of the University of Trento in terms of gender, departments, and their economic status. According to the research goal and research hypothesis, the experiment leader defined the following requirements of the sample to correctly represent the population of students:

- Students of the University of Trento enrolled in the academic year 2015-2016.
- They must attend the lessons regularly during the experiment.
- They must own an Android smartphone with 5.0.2 or higher operating system.
- They must have participated in the surveys of the "Observatory on training careers and on professional destinies of the students of the University of Trento".

The main configuration of the platform concerned the security policy concerned the collection, storage, management and analysis of the data collected throughout an experiment, in addition to the APIs for synchronizing the data.

- Pseudo-named the participant‘s data by assigning a unique code and decoupling the profile and data collected by phone from the identification data.
- The access to the information system for the purpose of the project was managed by an authentication and authorization system.
- All the data were processed and stored on the central server of the university and therefore subjected to the security measures provided by the server itself.

In order to be privacy compliant, the experiment leader defined a disclosure of personal data, a document which provide the students with information about the project aims, the type of data collected from the students, the storage and management of these data and the students' rights as data subjects. After the presentation of the project and the installation of the application on user's devices, the experiment - lasting two weeks - started. The duration of the experiment was two weeks for

two phases: during the first week, students responded to the time diaries and gave the explanation of the data that iLog collected, while during the second week students only had to leave the app running, with only 5 time diaries per day. The two main tools dedicated to data collection are described below.

- iLog Data Collection: the objective of the service is to collect information on the activities of persons via mobile devices. The reason the focus is on general purpose devices is that it can generate truthful readings, thanks to their easy integration with day-life activities, while invasive dedicated logging devices can alter normal routines. The system consists of a Mobile Application that collects sensor data from the smartphone and from additional external wearable devices through a Bluetooth connection. It's designed to be user friendly, transparent and unobtrusive. iLog consists of two components, a front-end part, which collects data from user's smartphone, and a back-end part, which stores the data collected in this way. The front-end component of iLog is a mobile application designed for Android devices, which is able to log sensor data and generate timestamped streams. Information is collected directly from both physical and virtual sensors. A virtual sensor could be a sensor aimed at collecting information about the people surrounding the user by capturing others devices in range through Bluetooth, or it could be an audio sensor able to extract audio features in real time from the microphone. Physical sensors are accelerometer, gyroscope, microphone, thermometer, GPS, gravity, magnetic field, orientation, proximity, light, pressure and humidity detectors. The application has been designed to be as unobtrusive and transparent to the final user as possible. In fact, once the user starts it and manually enables the logging mode, the application keeps running in the background as an Android Service. An overview of the sensors used by iLog is shown in Figure 3.2.

Figure 3.2: Sensors used by iLog

The back-end component of iLog is a persistence system based on Cassandra; it stores data collected by the front-end part of the application. The system has been designed to be scalable, because of the amount of data and the speed at which they are generated are huge. Furthermore, a query system permits to interrogate the database for later analysis. In accordance with the

GDPR standards, the collected data are kept exclusively for the period of time necessary for the research.

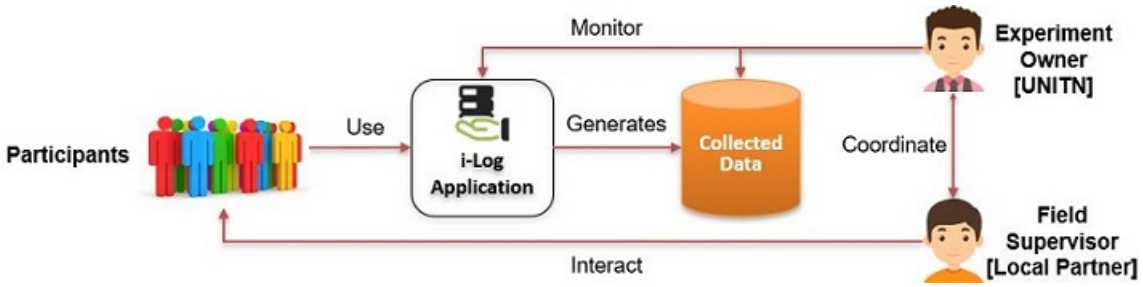

Figure 3.3: The process of iLog data collection, with the supervision of a local partner

- LimeSurvey Data Collection: the other way to collect data is though time diaries, which are logs where respondents are asked to detail how they allocated their time during the day. i-Log can administer the time diary from as a question composed of three sub-questions on activities, locations and social relations of participant every 30 minutes. Every triple of questions can be answered within 150 minutes from its notification, with a maximum of 5 questions stacked in a queue; otherwise, it expires and treated as null. More in detail, the time diary can collect information via 4 questions with closed entry questions: what the user is doing, where he is, who he is with and what is his mood. The complete questionnaire is shown in Figure 3.4.

What are you doing?	Where are you?	Who is with you?
Lesson	Class	Alone
Study	Study Hall	Classmate(s)
Eating	Library	Friend(s)
Personal Care	Other University place	Roomate(s)
En route (*)	Canteen	Partner(s)
Social life	Bar/ Pub/etc...	Relative(s)
Social media & internet	Home	Colleague(s)
Cultural Activity	Other Home	Other
Sport	Workplace	
Shopping	Outdoors	
Hobbies	Gym	
Other Free TIme	Shop	
Work	Other Place	
Housework	(*) How are you travelling?	
Volunteering	By Foot	
Other	By Bus	
	By Train	
	By Car	
	By Motorbike	
	By Bike	
	Other	

Figure 3.4: The complete questionnaire

Questions appear as silent notifications, in order to avoid bothering students and disrupt their activities too much. During the period that iLog was running, there were two notifications in the notification bar. The upper one showed the number of questions to be answered, while the bottom one notified that the application was running. The reason for this style of notifications was to enforce a transparency policy for the user so that she was always aware of the app behaviour during the experiment.

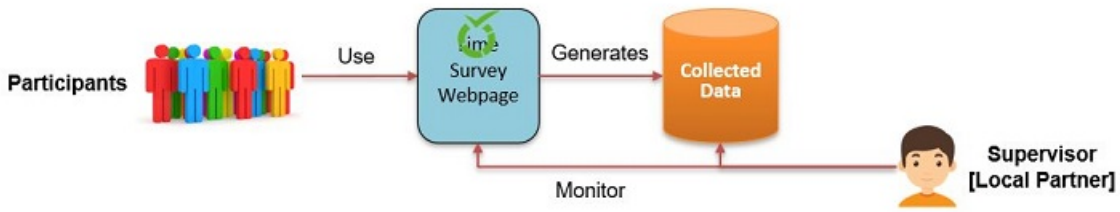

Figure 3.5: The process of LimeSurvey data collection, with the supervision of a local partner

3.1.2 Data Collection Execution

There are 3 main categories in terms of service activities during the data collection phase:

1. Campaign set-up: this activity covers the initialization of the campaign, and it consists in identifying user input requirements, the customization of iLog in terms of its parameters, the technical set up and configuration of the mobile application and the back end on cloud, and the registration of the users.

2. Data collection campaign: this activity covers the actual running of the data collection campaign in terms of user support and campaign monitoring to control the quality of the data throughout the campaign.

3. Campaign closing: this activity concludes the data collection campaign and covers the steps to consolidate and check the data.

Within the collaboration, the University of Trento would focus on the previous services, in particular, it will manage the customization of iLog, the technical configuration for iLog back end, the monitoring of user registration during the campaign set-up. There are several services operating during the campaign, including:

- System operation: the administration and operation of the data-collection system (front end and back end).

- Daily report: the University of Trento provides a daily report at the end of each day that consists in a table where, for each user, the research team indicates whether any data has been collected from each sensor at the end of the day.

- Help Desk, organized in different levels:

  - The first level is based on a FAQ data base provided by the University of Trento. The first level Help Desk is able to solve easy operation requests directly.
  - The second level Help Desk is a single point of contact; the e-mails from the user are recorded and for each one a ticket in the trouble ticket tool is created.
  - The second level Help Desk is able to solve medium operation requests; in this case it solves the ticket.
  - If the second level is not able to solve the request, it forwards the ticket to the third level.
  - The key assumption is that the iLog application is stable, and the issues are mostly due to biases, lack of knowledge or human mistake.
  - The Help Desk service manages a total number of tickets up to 10 percent of the participants' number (max 15 requests), expecting to be concentrated in the first days of the experiment.

3.1.3 Data Collection GDPR Compliance

The main source of the data for the project will essentially be students. Data will be generated by students through three main means:

- Data from sensors active on their smartphones.
- Data concerning social practices from questionnaires.
- Other possible administrative data sources to be defined in agreement with universities, local laws and the GDPR.

As for the personal data in the contents, that is, which allow to trace directly to a natural person, without the need for further correlation steps, these are essentially 5, listed in Figure 3.6. Once

DATA	Objectives	Risk Reduction
Home address	Data cleaning and harmonization	Location accuracy reduced via anonymization; limited access to authorized partners
Email address	Contact the user	Anonymization; not shared
Location data	Analysis of behaviour	Location accuracy reduced via anonymization; limited access to authorized partners
Contacts	Analysis of interactions	Anonymization; limited access to authorized partners
Internet Protocol (IP)	Install the app	Anonymization; not shared

Figure 3.6: Personal data collected and the way they are modified, in accordance with GDPR standards

collected, each of these data will be removed from the main dataset and, where necessary, replaced with a pseudo-anonymized or anonymized data. Specifically, the data concerning the email address (declared by the participant) and the IP, collected in order to be able to contact the participants and install the app on their smartphone (subject to explicit consent) will be replaced with an alpha-numerical identification (pseudonymization). The content of this data will not be shared with other project members. The participant's telephone contact names are collected as they are useful for analysing the network and interactions. These could contain the names of people, will be removed and replaced with alpha-numeric id (pseudonymization). Location data have multiple purposes and uses: from the validation of the information collected with the questionnaire, to the analysis of mobility routines, to the identification of points of interest in a city. For example, in the case of the home address, this is collected as the GPS signal is highly inaccurate when considering the desire to identify a building. In order to understand the movement of student populations, such as commuters, it is used to refine and validate observations and decrease errors and biases. In any case, these data will be separated from the main dataset and will be made anonymous in multiple ways (depending on the objectives with which they are analyzed).

Regarding the pilot and data collection phase, in the surveys previous to the project, the selection of participants was done using the lists of students enrolled in the University. The project includes neither children nor other vulnerable people. The initial contact with the participants is limited to an invitation to participate in the survey and the explanation of the pilot procedures. During the fieldwork, the consenting participants will be contacted for verification and assistance towards the effective and efficient carrying out of data collection activities. They will have the possibility to contact the helpdesk for any questions or details regarding the project and the pilot. Communication with

the participants will always be limited to the purposes of the pilot. Any communication outside the nature of the project will not be carried out through the project channels. As stated in the previous chapter, and according to the GDPR, every partecipant has the possibility to stop his partecipation in the experiment and to request the deletion of their personal data, by contacting the helpdesk.

3.2 Data Elaboration Procedures

The collected data may be used by the Data Controller and requested by the partners for performing experiments within the project. To satisfy these requests and comply with regulations the experiment input data will be generated from the collected data and shared. The process for preparing the experiment input data includes different steps towards making sure that all of the experiment data is relevant and limited to the purposes of the experiment (in accordance with the data minimization principle). The data preparation process follows models and techniques in accordance with the GDPR standards. The data are collected following the principle of data minimization, and are subsequently anonymized, so that it is not possible to trace them back to the user. Particular attention is paid to the anonymization of the GPS coordinates, obtained either by rounding or by substitution with Point of Interest, and to the anonymization of the user's personal data, obtained by associating each user with a key which will then be replaced in the data preparation phase. Anonymization is achieved by two different operations:

- User Id anonymization: replacing the unique attributes for each of the participants (name, email address) with another unique identifier so it cannot be linked back to the original dataset. The replacement technique used is one of the following:
  - A hash function applied to the original iLog id: with a function that cannot be reversed (SHA-256 cryptographic function is used to perform the hashing). This ensures that it is computationally very hard to obtain the original Id by reverting the hashing function. All other unique attributes are removed.
  - An encryption with secret key: in this case, the key will be deleted at the end of the anonymization process before distributing the anonymized data to avoid the possibility of tracing the identification of each data subject by decrypting the data set if one is aware of the key.

  Both of the operations will produce numeric identifiers that are collision free (no two users are able to have the same identifier) and cannot be traced back to the original information.

- GPS anonymization: raw GPS is normally considered personal information under GDPR, so we provide two options for anonymizing it.
  - GPS Rounddown: precision is deliberately truncated from the location sensor so it becomes anonymous but it is still useful for certain scientific purposes. This is achieved by truncating the year in the GPS timestamps, changing its format and rounding down latitude and longitude values.
  - GPS transformation to Point of Interest: if latitude and longitude do not change for a period of time then a Point of Interest tag is added to the stream. For each Point of Interest the elapsed time in seconds is added to the stream. GPS longitude and latitude readings are removed. The Point of Interests corresponds to the general area (suburb, city, region) and the closest general places (bar, restaurant, lake, etc).

The anonymization process leads to the generation of a full anonymized dataset, stored in IT backbone of the research infrastructure.

3.3 Datasets Management

Generation, curation and preservation of datasets has to be done in compliance with current European and local legislation. In the first stage, the project will identify the overall storage guidelines

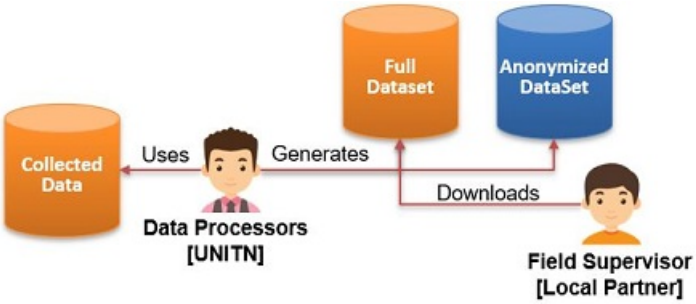

Figure 3.7: The process of iLog data preparation, generating a full dataset and a partial anonymized dataset

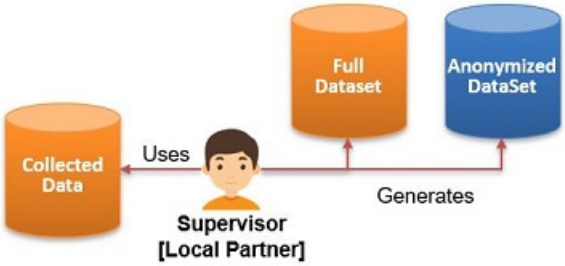

Figure 3.8: The process of LimeSurvey data preparation, generating a full dataset and a partial anonymized dataset

and limitations to drive the platform and research data infrastructure implementation and the data management processes design and operation. Once the set data preservation period expires, appropriate curation methods will be implemented accordingly (including archiving, final purge, destruction, anonymization, etc.). Collected or generated data streams for the project are stored in the cloud facilities to protect data from being lost or destroyed and to contribute to the reuse and progression of the data. The datasets related to the data collection will be kept by the local partner; the experiment generated data and experiment results will be made available for further research through the research infrastructure. No personal or sensitive data - which is to be collected according the experiments' needs - will be shared with third parties outside the project or offered on open access repositories. Data subjects will have the right to request the erasure of their personal data to the project data controller holding their information. The deletion will be then carried out without undue delay, as defined by the specific details of the experiment.

4 Procedures for GDPR Approval

The research project complies with the General Data Protection Regulation (GDPR) of the European Union and adheres to high ethical standards for data handling. A particular ethical challenge is the transfer of data from the European Union to non-EU countries and vice versa. During the first phase of pilots, data will be collected both inside and outside the EU. As an EU-funded research project, the data processing in the research infrastructure complies with EU and national data protection laws, both in the foundations of the project and in its realization. This applies to the phases of collection, storing and sharing; where the principles of safeguarding the rights and freedoms of the data subjects and the principles of data protection are followed according to the internal regulations of both EU and non-EU members. Non-EU members, as stipulated in the signed Grant Agreement, will be legally bound to comply with European regulations.

The project will collect sensitive data, which is commonly referred to as "special categories of data". These may be: name, email, phone number or other data that allow to identify the data

subjects with their own measuring instrument (smartphones) and their own data. It is important to note that this will be collected only for the purpose of installing and using the apps to be used as data collection tools by the project. Other personal data collected includes information about gender, age, level of education, enrolment in a department, geographic location, and some attitudes regarding food consumption, linked to lifestyle and, indirectly, to health.

The data will be collected by members designated as Data Controllers (the natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data). They will be in charge of the data processing, the specification of the scope/purpose of the data collection and managing related experiment's field maintenance and data storage processes. The following typology for data are defined: collected data, experiment input data, experiment output data, result data, and service data. Each data collection and processing in the project is required to be approved by a local ethics commission, given the involvement of people and the sociological nature - within the interdisciplinary context - of data processing. In the event of absence of an ethics commission locally to a specific data collection activity, the members designated as Data Controllers will be responsible for identifying and nominating an ethics commission.

4.1 Ethics Operation Steps

This chapter describes the mandatory process for handling ethics and privacy law compliance when performing data collection or processing activities within the project, shown in Figure 4.1.

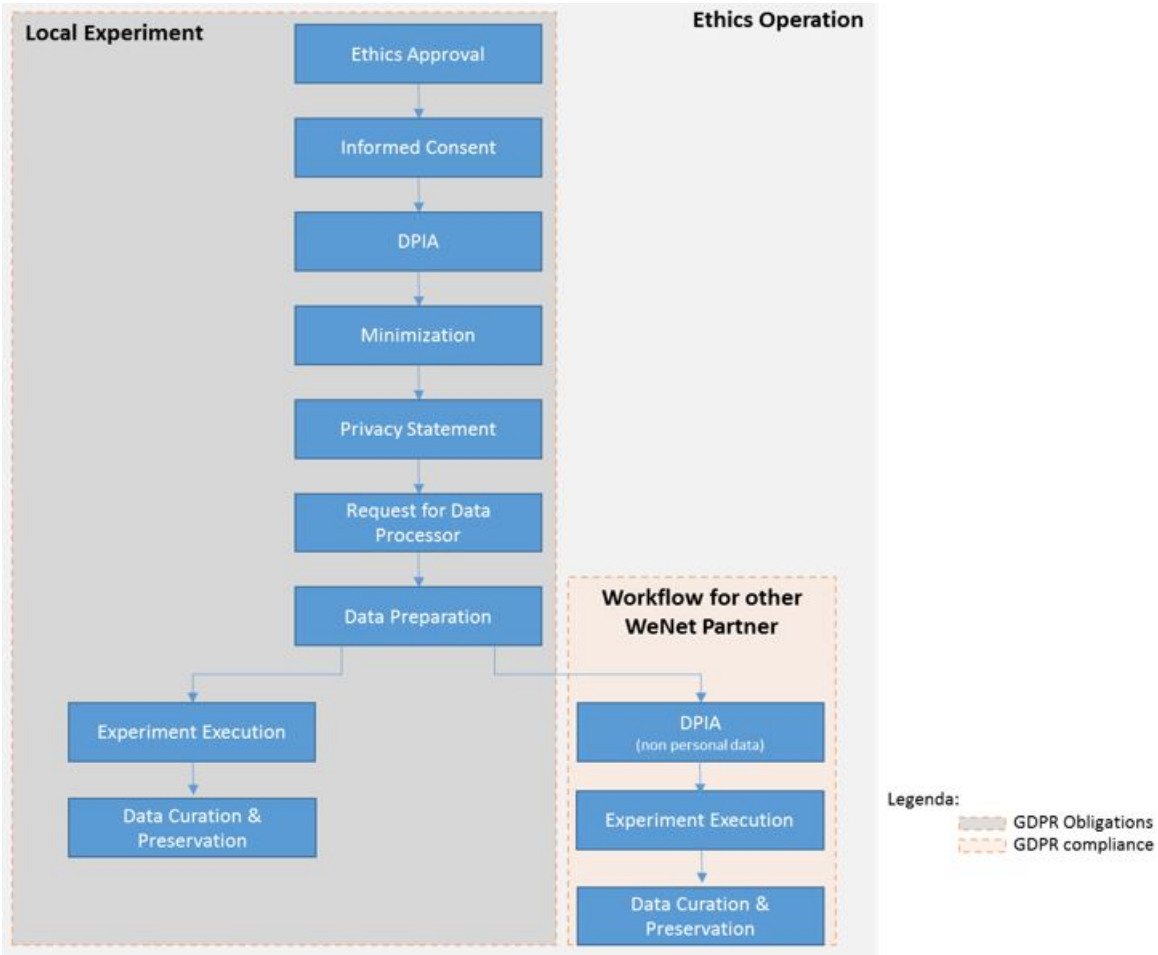

Figure 4.1: The project's ethics operation steps conceptual model

4.1.1 Ethics Committee Approval

At the start of any trial, pilot, experiment or study for all partners, the coordinator of the research project and pilot trial must submit the request for review to the identified Ethics Committee and get its approval. In the absence of the Ethics Committee figure at the University, the Data Controller

must identify the alternative suitable figures to perform this task. He needs to:

- Identify the relevant Ethics Committee to which submit the request of approval.
- Compile and submit to the Ethics Committee the "template to request the approval of a research protocol" and all relevant documents. In case the request needs to be made in a non-English local language, an English version of the request must also be prepared and stored as part of the documentation of this process.
- Obtain the approval from the Ethics Committee in written form. Otherwise, if more changes are needed, a revised submission is prepared and submitted.

4.1.2 Data Protection Impact Assessment

Once obtained approval by the ethical committee, a Data Protection Impact Assessment (DPIA) needs to be written. This is a necessary step whenever a Data Controller plans to collect or process data from any trial, pilot, experiment or study within the project. A DPIA is a document that helps the project assess, identify and minimise risks that may result from data processing. Data Controllers are responsible for ensuring the DPIA is carried out for each processing operation of personal data. Additionally all the other partners responsible for data processing should also conduct a DPIA. The required procedures are:

- Identify the institution's DPO (Data Protection Officer) or equivalent figure that will eventually assist in the compilation of the DPIA.
- Compile the DPIA using the WeNet template.
- As an optional step, a submit to DPO the DPIA and all relevant documents. In case the request needs to be made in a non-English local language, an English version of the request must also be prepared and stored as part of the documentation of this process.
- As an optional step, the obtaining of the review from the DPO and the implementing of any comments into a finalized version of the DPIA document.

4.1.3 DPIA for Data Processors

One of the best ways to mitigate the ethical concerns arising from the use of personal data is to anonymize it so that it no longer relates to identifiable persons. Data that no longer relates to identifiable persons, such as aggregated and statistical data, or data that has been otherwise rendered anonymous (the data subject cannot be re-identified), are not personal data and are therefore outside the scope of data protection law. However, even if a consortium partner uses only anonymized datasets, the partner must specify the source of the datasets he intends to use in its research by the preparing a DPIA document. In this case the partner will complete only part 1 of the template by answering "no" to all the questions to justify the decision of not to carrying out a full DPIA (answering yes to one or more questions is an indication that a full DPIA is necessary). The partner is responsible for ensuring the DPIA is carried out for each processing operation of personal data (anonymization, pseudo-anonymization, etc). In order to do that, he must compile the part 1 of the DPIA using the WeNet template. If, after completing the screening questions in part 1, the partner decides a full DPIA is not necessary, he has to mantain a record of answers to the screening questions of the template. After have received the finalized compiled DPIA document, the project maintains a record of answers to the screening questions, in order to document that the decision on whether to carry out a DPIA was properly considered.

4.1.4 Informed Consent

Personal data processing requires free and fully informed consent from the involved persons. In the project, all participation must be voluntary. As such, the data controller and related partners must obtain (in advance) and clearly document the participants' informed consent. This step is necessary whenever a partner plans to involve research or study participants in any trial, pilot, experiment or study within the project. The data controller (aided by the participating partners) is responsible

for providing the information related to informed consent and to ensuring informed consent from all participants. The informed consent forms:

- Are written in a language and in terms the participants can fully understand.
- Describe the aims, methods and implications of the research, the nature of the participation and any benefits, risks or discomfort that might ensue.
- Explicitly state that participation is voluntary and that anyone (at any time and without consequences) has the right to refuse to participate and to withdraw their participation and generated data.

The mandatory outcomes from this step are the text used to provide information to the users, an information file containing a trace of the informed consent given and the Informed Consent uploaded in pdf format (in the event that data collection is conducted with a paper-based alternatives).

4.1.5 Data Minimization

The project should collect only the data needed to meet its research objectives. Collecting unneeded personal data may be deemed unethical and unlawful. Before start any trial, pilot, experiment or study within the project the partners should conduct a data minimization review to ensure that data are collected on a "need to know" basis and report it in order to document that the principle was properly considered. The researcher leading the experiment is responsible for ensuring the review is carried out and confirmed by the Data Controller. The main procedures implemented for this step require:

- Conduct the review before any trial, pilot, experiment or study is activated.
- The review applies not only to the amount of personal data collected, but also to the extent to which they may be accessed, further processed and shared, the purposes for which they are used, and the period for which they are kept.
- Perform the review according to the project guideline guideline described in D11.2 POPD (Protection of Personal Data) document.
- Maintain the summary for the checks done in the project repository, so that it can be documented that the data minimization principle was properly considered.

4.1.6 Data Preparation

The partners should only use during data processing activities as much data as is required to successfully accomplish a given task. Data Preparation is a necessary step at the end of each data collection campaign or before starting any subsequent study that involves use of a dataset within the project, therefore the partners should conduct a data preparation task to ensure that the data minimization principle is properly applied. The Data Controller is responsible for carrying out this step, he may enlists the help of a Data Processor for the specific data process activities. The procedures implemented are:

- Conduct the data preparation operation to generate the experiment input data from the collected data (the task is carried out under the responsibility of the original data controller of the collected data).
- Personal and sensitive data will be removed. The project protocol states that partners that are going to use that data for an experiment should never have access to personal and sensitive data originally in the dataset (it implies that anonymization and pseudonymization techniques are applied during data preparation).
- Additionally, the process could optionally include steps towards making sure that all of the experiment data is relevant and limited to the purposes of the experiment (in accordance with the data minimization principle).

- Document a summary report detailing the tasks performed as part of the Data Preparation process, maintained by the project.

The project maintains a summary report about the data preparation tasks in order to document that the data minimization principle was properly considered.

4.1.7 Privacy Statement

The partners must inform the potential participants on the technical and organizational measures to safeguard the rights of the research participants. To accomplish this, the Privacy Statement is given to individual participants to explain how their personal data is processed; it has two purposes: to promote transparency and to give individuals more control over the way their data is collected and used. The Data Controller is responsible for the Privacy Statements. They are a legal requirement under the GDPR to ensure that individuals are aware of the way their personal data is processed. The procedures implemented are:

- Identify the institution's DPO (Data Protection Officer) or equivalent figure that will eventually assist in the compilation of the Privacy Statement.
- Compile the Privacy Statement using the WeNet template.
- As an optional step, submit to DPO the Privacy Statement and all relevant documents. In case the Privacy statement needs to be made in a non-English local language, an English version of the statement must also be prepared.
- As an optional step, obtain the review from the DPO and implement any comments into a finalized version of the Privacy Statement.
- Provide a privacy notice whenever a data subjects' personal information is obtained.

The Privacy Statement and other consent documents must be kept on files. All consents will have to be exported to pdf, with the necessary information (date, time, etc) for each user. In the event that data collections are conducted with paper based alternatives, Privacy Statement must be uploaded in pdf format.

4.1.8 Request for Data Processor

The Data Processor processes personal data on behalf of the controller. Within the project, the procedure must ensure that any partners that process research data at the Data Controller request (and on his behalf) comply with the GDPR ethics standards. The Data Controller is responsible for engaging the Data Processor with a prior written authorization; according to this written authorization the data processor also may be held liable, along with the controller or other processors, in case of GDPR infringements. When the Data Controller wishes to subcontract certain activities (which imply the processing of personal data) to a Data Processor, a proper drafted data processing request is needed to adequately protect Data Controllers. The Data Controller has to draft a data processing agreement (or a letter of appointment as Data Processor).

4.1.9 Experiment Execution

A Code of Conduct of the project sets forth the principles and ethical standards that underlie the members' scientific and professional responsibilities and conduct. The partner leading the experiment is responsible for ensuring the declaration of commitment of the researchers. By signing the statement all involved researchers in the experiment shall comply with the provisions of the ethical rules of professional conduct for treatments for statistical purposes or scientific research. Every signed declaration of commitment has to be kept on file.

4.1.10 Data Curation and Preservation

The collected data must be kept in a form that enables the data subjects to be identified for a period not exceeding what is necessary for the purposes for which they are processed. Data retention occurs for a specified time period, based on the legal agreements in force and the business needs. The appropriate curation methods are implemented accordingly: archiving, final purge, physical destruction,

anonymization, etc. This step is mandatory in any trial, pilot, experiment or study for all partners of the project. To ensure data curation and preservation, a Data Controller is needed.

5 Conclusions

This thesis has presented a diversity-aware paradigm for the design of social media platforms and the relative research infrastructure. The adopted design approach leverages the diversity of users to their benefit. In the case of the presented platform and app, the approach is used to design a platform and corresponding app which fosters students' social relations and well-being. The project has laid out the theory and practice of framing and operationalizing "diversity", as well as collecting data regarding diversity among the students, in compliance with the GDPR and its strict data protection regulations. In order to define the diversity of a community, the project proposes to look at the social practices present in said community. These practices can be learned from data regarding diversity, which are collected from surveys and the interaction of students with an application.

The project described in this thesis has the fundamental objective of bringing humans and the protection of their privacy at the center of research. The project is born in a context of change and debates related to ethics and privacy, in particular following the recent introduction of the GDPR. In this context, this thesis describes how the research infrastructure adopts the minimization strategy, selecting data and tools that are relevant, adequate and necessary for the purpose of the project. This is both through the guidelines underlying the academic research projects and the principles on which the project is based. This strategy is accompanied by a careful selection of data exchange methods, between EU and non-EU countries, in accordance with current regulations, as well as operational data protection and impact and risk analysis choices.

The integration of ethical principles is of central importance in the development of such a platform and research infrastructure. Ethical values and principles must be observed during the development and design phase and by the technology itself. The entire project is based on the following general principles: transparency, fairness, purpose limitation, data minimisation, accuracy, storage limitation, data security, accountability, privacy-by-design and by-default, protection of minors, monitoring on a regular basis and iterated evaluation as well as support of privacy-literacy. These general principles were, then, applied to all central components of the research project.

Bibliography

[1] WeNet - Internet of Us. *https://www.internetofus.eu*.

[2] R. Bartolacelli, D. Ben Zaken, R. Bona, M. Britez, C. Caprini, S. De Cristofaro, W. Droz, B. i Gui, C. Michael, D. Miorandi, I. Perikos, N. Pomini, A. Segal, and S. Tavonatti. D6.2 WeNet Platform and Research Infrastructure 1.0. *WeNet Consortium, Work Package 6*, 2020.

[3] A. Baur, M. C. Campodonico, A. de Götzen, J. Heesen, D. Miorandi, K. Reinhardt, and C. Rodosthenous. D9.3 A Revised Guideline Concerning Privacy - Standards for WeNet. *WeNet Consortium, Work Package 9*, 2020.

[4] A. Baur, J. Heesen, K. Reinhardt, and L. Schelenz. D9.2 A Preliminary Guideline Concerning Privacy - Standards for WeNet. *WeNet Consortium, Work Package 9*, 2019.

[5] M. Bidoglia, I. Bison, M. Busso, M. Britez, R. Chenu, M. Cvajner, G. Gaskell, G. Sciortino, S. Stares, and G. Veltri. D1.3 Final Model of Diversity: Findings from the pre-pilots Study. *WeNet Consortium, Work Package 1*, 2021.

[6] M. Bidoglia, I. Bison, M. Busso, L. Cernuzzi, A. Chagnaa, R. Chenu, A. de Götzen, A. Ganbold, G. Gaskell, F. Giunchiglia, C. Gunel, A. Hume, P. Kun, M. Rodas, S. Stares, G. Veltri, J. L. Zarza, and M. Zeni. A Worldwide Diversity Pilot on Daily Routines and Social Practices (2020). *Department of Information Engineering and Computer Science, University of Trento*, 2020.

[7] E. Bignotti and M. Zeni. i-Log for Eurostat Summary - i-Log service within the European Big Data Hackathon 2019. *Department of Information Engineering and Computer Science, University of Trento, KnowDive Group*, 2018.

[8] I. Bison, M. Busso, D. Gatica-Perez, F. Giunchiglia, A. de Götzen, L. Meegahapola, S. Ruiz-Correa, and L. Schelenz. The Theory, Practice, and Ethical Challenges of Designing a Diversity-Aware Platform for Social Relations. *AIES '21: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 905–915, 2021.

[9] I. Bison, M. Busso, and F. Giunchiglia. Experiment Design - SmartUnitn(Two). *Department of Information Engineering and Computer Science, University of Trento, KnowDive Group*, 2019.

[10] R. Boardman, A. Mole, and J. Mullock. Guide to the General Data Protection Regulation. *Bird and Bird*, 2016.

[11] R. Bona. Considerazioni relative al processo di anonimizzazione. *Department of Information Engineering and Computer Science, University of Trento, KnowDive Group*, 2020.

[12] R. Bona. WP6 Research Infrastructure. *WeNet Consortium, Work Package 6*, 2020.

[13] R. Bona, M. Busso, R. Chenu, F. Giunchiglia, and L. Schelenz. D11.2 POPD - Requirements n.6. *WeNet Consortium, Work Package 11*, 2019.

[14] R. Bona, M. Busso, R. Chenu, and R. Job. Ethics and Privacy WeNet Operating Procedures. *WeNet Consortium, Work Package 11*, 2019.

[15] R. Bona, C. Caprini, R. Chenu, S. De Cristofaro, W. Droz, B.R i Gui, T. Mantadelis, D. Miorandi, and A. Segal. D6.1 WeNet Platform Architecture Specifications. *WeNet Consortium, Work Package 6*, 2020.

[16] R. Bona, R. Chenu, and F. Giunchiglia. D10.2 Data Management Plan. *WeNet Consortium, Work Package 10*, 2019.

[17] M. Busso, R. Chenu, G. Sciortino, and D. Song. SmartUnitn(One) - Experiment Design. *Department of Information Engineering and Computer Science, University of Trento, KnowDive Group*, 2019.

[18] Y. Demchenko, C. de Laat, P. Grosso, and P. Membrey. Addressing Big Data Issues in Scientific Data Infrastructure. *2013 International Conference on Collaboration Technologies and Systems (CTS)*, pages 48–55, 2013.

[19] Y. Demchenko, C. de Laat, P. Grosso, A. Wibisono, and Z. Zhao. Addressing Big Data challenges for Scientific Data Infrastructure. *4th IEEE International Conference on Cloud Computing Technology and Science Proceedings*, pages 614–617, 2012.

[20] F. Giannotti, V. Grossi, D. Pedreschi, and B. Rapisarda. Data science at SoBigData: the European Research Infrastructure for Social Mining and Big Data Analytics. *International Journal of Data Science and Analytics*, 6:205–216, 2018.

[21] F. Giunchiglia, I. Zaihrayeu, and M. Zeni. Multi-Device Activity Logging. *Department of Information Engineering and Computer Science, University of Trento*, 2014.