UNIVERSITY OF TRENTO

DEPARTMENT OF INFORMATION ENGINEERING AND COMPUTER SCIENCE

DOCTORAL THESIS

BIG-THICK DATA GENERATION VIA LIFE SEQUENCES

Author

Xiaoyue Li

Advisor

Prof. Fausto Giunchiglia

April, 2025

Acknowledgements

The PhD journey is full of challenges and joys, stemming from both interesting research and personal life. Looking back to the time, I would like to begin by expressing my gratitude to my advisor, Prof. Fausto Giunchiglia, for granting me a great opportunity to pursue my doctoral degree. His constant support and encouragement over the past four years have been invaluable. I am very honored to conduct research closely with him, and this experience will have a very positive impact on my future research career. Furthermore, I extend my thanks to the members of our Knowdive research group, who have consistently shared knowledge and offered generous assistance to one another, creating so many unforgettable experiences with them. I am also grateful to Prof. Themis Palpanas for providing me with the amazing opportunity to study with him at the LIPADE Laboratory, where I was introduced to diverse research methodologies. Thanks also to the reviewers, Prof. Gabor Bella and Prof. Wanyi Zhang, whose valuable comments helped me improve this thesis. Last but not least, I would like to express my heartfelt appreciation to my family, particularly my husband, Haonan Zhao, for their unconditional trust, assistance, and support throughout my doctoral studies.

Abstract

The large-scale personal *big data* is collected at high speeds by smart devices, mainly in the form of sensor data. While this extensive type of data allows for the analysis and inference about the environment where it was collected (e.g., locations) and certain aspects of human behavior (e.g., physical activities and transportation modes), a major limitation of this type of data is that it is not *thick*, that is it does not carry the information about the *context* within which it was generated. However, contextual information, such as personal, social and also objective interpretation reflecting upon how and why people do what they do, must be explicitly represented for the data to be self-explanatory and meaningful.

In this thesis, we introduce *big-thick data* as big data complemented with highly contextualized thick data. To generate big-thick data, we first represent it by defining *life sequences* as sequences of *observation contexts* over a certain period. Then, observation context is composed of two types of context, namely, *personal context* and *reference context*. The personal context encodes, for every single person, a *subjective* personal view of the world, while the reference context encodes an *objective* view of an all-observing third party. Finally, we model these two types of context and explore how they can be *unified* in many different ways (for specific purposes) into observation contexts. The big-thick data is generated by populating sequences of observation contexts in life sequences with various data as sequences of knowledge graphs. The life sequence representation and the flexible context unification allow for different types of *enquires* posing to big-thick data, enquiring about people's habits, daily lives and the surrounding world; and also allow for multiple different (possibly contradicting) answers (provided by different people or even the same person at different times) to the same enquiry. This is the crucial property of big-thick data which we consider key to the development of meaningful human-in-the-loop human-machine interactions.

Our work has been validated with a case study, where a SU2OSM big-thick dataset was generated by integrating the SmartUnitn Two (SU2) dataset with the Trentino region OpenStreetMap (OSM) dataset to construct students' life sequences. The observation contexts in these life sequences unified the reference context populated by the OpenStreetMap data of the Trentino region with the personal contexts populated by the SmartUnitn2 dataset (consisting of time diaries and sensor data from one hundred and fifty-eight students at the University of Trento over a period of four weeks). The results demonstrated that the generated SU2OSM big-thick dataset is capable of answering a broader range of enquiries and exhibited superior prediction performance compared to the original datasets.

Keywords: Big-thick Data, Life Sequence, Observation Context, Reference Context, Personal Context, Context Unification, Enquiries.

Contents

List of Tables

List of Figures

Glossary

Big-thick Data: *Big-thick data* is *big data* complemented with *thick data*, that is, observational data about *life sequences* of people, which allows us to reflect upon how and why people do what they do over time.

Context Unification: *Context Unification* is the process that *unifies* the reference context and a set of personal contexts to construct the observation context. Different possible unification ways can be used in this process, the selection depends on the specific purpose. This means we not only distinguish between reference and personal context but also consider the various ways to unify them, which makes it flexible to integrate datasets and allows for multiple diverse research questions even for the same data sources.

Enquiry: The *enquiry* is synonymous with *inquiry*, as a request question for acquiring information. We have two types of enquiry: *observation enquiry* and *habit enquiry*. The first distinguishes any specific possible observation purpose (as a function of what one is interested in observing) in observation contexts, such as one or multiple elements from personal contexts, or reference contexts, or both of them. The latter focuses on enquiring about various human habits, such as the spatial habit of a person always being at home.

Entity Graph (**EG**): An *Entity Graph* is generated by instantiating a teleology, as described in our previous work [55, 61]. An EG is a Knowledge Graph (KG) constructed from a teleology by populating each etype node with the specific entities of that etype, as well as instantiating the remaining components, including data properties, object properties, and data types.

Life Sequence (**LS**): A *life sequence* of a person represents a sequence of the real-world scenarios that she/he is in, which is represented as a sequence of her/his *observation contexts* during a certain period of time. The life sequences of people are the representation of big-thick data.

Observation Context (C): An *observation context* represents a real-world scenario and is built out of (a part of) the *reference context* and (parts of) one or more *personal contexts* based on shared identifying information.

Ontology: The term *ontology* builds on the Greek words *ont* (meaning *being*), and *logia* (meaning *a branch of learning*) [58]. We use the term ontology (and also *teleontology* and *teleology* below) to capture the intuition that they are constructed with a *purpose*. There is no claim of generality beyond the purpose for which it is generated. An ontology is an explicit specification of the shared conceptualization as to what exists, which is an evolving directed acyclic graph, representing the hierarchy of concepts. In an ontology graph, the nodes are a set of Entity Types (etypes, referring to types of real-world entities), which are related by the *isA* relation indicating subsumption. If one etype subsumes another, the former is the superclass of the latter (subclass), with *isA* depicted as an arrow from the subclass to the superclass. The etypes can be decorated with data properties representing their attributes.

Personal Context (C_P): A *personal context*, one for each user in C_R, encodes a person's subjective view of the world, which is any context that carries a subjective representation of the world.

Reference Context (C_R): A *reference context* provides an objective, person-independent all-encompassing view of a third-party observer, which is any context that carries an objective representation of situations within which users are operating. C_R is from C_S enforcing objectivity.

Spatial-temporal Context (C_S): A *spatial-temporal context* provides the view of real-world space in time, constructed by a *reference spatial region* (S) and a set of *spatial regions* perceived as objects within the boundary of S. Massive data sources can be used to build C_S. For example, the OpenStreetMap dataset from Geofabrik[1] provides geographical information with relatively large reference spatial regions, such as the northeast of Italy (not available for the sub-region of Trentino), where a vast number of spatial regions can be found, such as buildings, universities, and hospitals.

Teleology: The word *teleology* builds on the Greek words *telos* (meaning *end, purpose*) and *logia* (meaning *a branch of learning*) [58]. Teleology is the same as the Entity Type Graph (ETG), e.g., used in [61]. Teleology is built by properly selecting elements of teleontology based on a purpose. It is first to select a subset of etypes from teleontology and remove all their hierarchical links, then appropriately adjust the data properties and object properties for each etype due to property inheritance.

Teleontology: *Teleontology* models the purpose-specific interrelationships of etypes, inheritances, data properties, and object properties within the scope of a knowledge graph engineering project. Object properties represent the relations among etypes. Teleontology can be (partially) modeled using a top-down approach by (partially) reusing the knowledge of domain experts and existing applied ontologies. The base forms of Enhanced Entity-Relationship (EER) models can be formalized as teleontologies. There are primary representational constructs in a teleontology graph: a set of etypes as head nodes, each decorated with a set of data properties as arrows, pointing to data types as tail nodes. A set of object properties illustrate the relationships among etypes, represented as arrows linking etype nodes. The *isA* relations indicate subsumption among etypes. The etypes as subclasses inherit the data properties and object properties of their superclasses.

View: A *view* represents context as a set of facts. As observation context is a partial representation of the world, we distinguish two types of views for obser-

[1] https://download.geofabrik.de/.

vation contexts.

- *Objective view* implies the facts representing a situation from the objective view of an all-observing third party. These facts are known and agreed upon to be true by all the objects who have got to know about the situation, such as the university address for students.

- *Subjective view* implies the facts representing a scenario from the subjective view of a person. These facts are exclusively known to and influenced by people, such as the fact that Bob is my friend or I am happy.

Chapter 1

Introduction

1.1 Motivation

Smart devices, such as smartphones and smartwatches, facilitate the collection of large-scale sensor data at high speeds, such as GPS, Bluetooth, WIFI, and accelerometer. This extensive type of data is commonly referred to as *big data* [16, 31], enabling the tracking and analysis of some aspects of people. For example, big data has been extensively used in many AI studies, e.g., Human Activity Recognition (HAR) [25, 124, 135, 141], Health Monitoring [130] and Autonomous Vehicles [95, 153] and Intelligent Transportation [69]. Nevertheless, big data is often used 'out of context', which substantially diminishes its meaning and value [18]. De-contextualized big data is difficult to use when trying to understand human behaviors, such as involving human social and personal life, which are always context-sensitive.

To address the de-contextualization of big data, researchers suggested complementing it with the highly contextualized *thick data* [12, 30, 44]. Thick data is synonymous with ethnographically collected and analyzed observational data [47]. It emerges from the observations or interviews of the contexts in which human behaviors occur, containing the contextual information around humans, e.g., social media content, people's interviews, and self-reports [16]. This enables researchers to understand and reflect upon the scenarios where people are.

However, thick data is often limited in scale due to the challenges associated with its collection, as already made evident by the work on crowd-sensing, see, e.g., [2, 89] and mobile self-reports, see, e.g., [50, 158]. Given the above, integrating the complementary big data with thick data is necessary [16, 49]. The concept of *big-thick (data) blending* has been proposed [16] and discussed its benefits in improving data analysis [4, 16], due to the fact that thick data provides a *why* observations on the *what* observations from big data [85]. Current studies related to big-thick data remain the general theories and methodologies [44], without actually generating extensive big-thick data.

The notion of *context* has a long history across many research areas, dating back to John McCarthy's seminal work [106, 107]. In the Human-Computer Interaction (HCI) community, context was defined as 'any information that can be used to characterise the situation of an entity' [33]. Within the Knowledge Representation (KR) community, contexts are modeled as knowledge bases that allow for knowledge representation and reasoning [22, 23, 57]. The *context recognition* tasks in the Internet of Things (IoT) domain apply the context to learn about various aspects of human behavior. In this research line, the possible context dimensions include locations, activities, body posture, and more [7, 142]. *Context* – what was accomplished by a person, how and why, and in which overall situation – all these factors must be explicitly represented for the data to be self-explanatory and meaningful [52]. As the boundary between IoT and KR, Giunchiglia et al. proved that context modeled in the human life sequence to represent real-world daily scenarios can be used to collect, model, and reason in personal data streams [61]. Contextual factors are also necessary for multiple predictions based on the same dataset, because the same sensor value may stand for two completely different contextual situations. For example, the same GPS value indicates a professor and a student in the same location, e.g., the university, but maybe with different purposes, such as work and study respectively. The student can have different purposes at the same university, such as meeting a friend or joining a club. Meaningful, lifelong human-in-the-loop, human-machine interactions need this level of information richness.

1.2 Problem

To address the issue of data de-contextualization, we introduce the notion of *big-thick data* as big data complemented with *thick data*, that is, *observational data about contexts evolving in human daily life which allows keeping reflect upon how and why people do what they do*. In this thesis, we formalize human *life sequences*, composing of *observation contexts* to represent big-thick data, in practice, which is further used to generate big-thick data by integrating data from multiple sources.

However, many challenges still need to be addressed to formalize life sequences:

- *Problem 1*: In real-world human life, people constantly perceive various contexts, which are not static and change in space over time, such as a traveling person being in different locations at different timestamps.
- *Problem 2*: Context is not a simple matter [37]. A context often comprises multiple elements that often interact with each other, however, the representation of context should be general and cannot be too complex for humans and machines to interpret. Schilit et al. [128] represent people's situations concerning location, identities of nearby people and objects, and changes to those objects [128]. Giunchiglia et al. propose a three-level hierarchy of objects, actions, and functions concepts [58], as well as the endurants and perdurants [45, 52] to represent the world we perceive.
- *Problem 3*: Different partial views [19, 48] exist for a context. Context can differ for the same scenario across different people because of their different cultures, subjective perspectives, and other factors. Giunchiglia

et al. [61] define a *situational context* represents a real-world scenario from the perspective of a specific person. Another partial view of context can be seen from the definition of *situation*, which is the complete state of the universe at an instant of time [108].

- *Problem 4*: Big data and thick data are highly heterogeneous, e.g., categorical, numerical, in natural language, and unstructured formats, and usually collected with different time frequencies. For example, the sensor data may be stored in CSV tabular format or structured JSON format. Furthermore, even the same thing is diverse across multiple datasets because of varying levels of abstraction or different sources (e.g., different people). For instance, a location can be described as, e.g., GPS coordinates, a university, a workplace (e.g., for professors) or a study place (e.g., for students), the city of Trento, or Italy.

1.3 Solution

Given the challenges above, the solutions for formalizing life sequences to generate big-thick data are as follows:

- *Solution 1*: We formalize the *life sequence* of a person as a sequence of *observation contexts* during a certain period. Any observation context is defined by the current *locations* and *events* in time, representing the spatial-temporal regions. The observation contexts in a person's life sequence are sequentially ordered in space and time, depicting various real-world scenarios.

- *Solution 2*: We define a general observation context as being populated by a set of *entities* representing anything with spatial and temporal extensions in the observation context, such as people, buildings, and phones. Each entity is identified by the *object* representing its spatial regions (e.g., the body of a person) as well as a set of *functions* and *actions* illustrating the entity's expected actions (e.g., professor) and changing states in time (e.g., walk).

- *Solution 3*: The observation context is constructed based on two main components: (part of) a *reference context* and one or more *personal contexts*, which are distinguished by their representation of observation context from a (partial) *objective* view and a (partial) *subjective* view, respectively.
  - A reference context provides a person-independent, objective, and all-encompassing view from a third-party observer, which describes the situation within which people are operating. It is defined in terms of a suitable *reference location* as observed (e.g., Trentino or Italy) during a *reference observation period* (e.g., one month).
  - A personal context encodes the world from a subjective view of a person, such as where she is, what she is doing, why she does it, who she is with, and her mood [52]. Personal contexts may differ across persons, and even for the same person at different times, as people's subjective views evolve. A set of personal contexts can arise from one reference context, where each personal entity has her/his own personal contexts.

We propose the purpose-driven *context unification*, which aims to compose the subjective information of personal contexts into the objective perspective of the reference context. Observation context is built out of the reference context and one or more personal contexts based on shared *identifying information*, e.g., identifiers, names, and spatio-temporal coordinates. Context unification is a flexible process, which is configured as a function of a specific purpose, as defined in [54]. The flexibility mainly comes from the different purposes that lead to selecting varied shared identifying information linking personal and reference contexts, or selecting different data to build contexts. For example, to learn about the behavior of a certain group of people, the observation context might be tuned to that group. The reference context is the key pivotal element in this process, what allows to *absorb* subjectivity into objectivity. In turn, the personal contexts show the subjective views of different people about what happened in the reference context.

- *Solution 4*: We propose a general methodology for representing life sequences, which is used to formalize and integrate different data sources into big-thick data. As components of observation contexts in life sequences, the reference context can be built using various types of spatio-temporal (big) data, including those annotated by humans, e.g., coordinates, images and labels, obtained from sources like OpenStreetMap (OSM)[1] and Italian Spatial Data portal[2]. Additionally, personal contexts can be used to generate personal big-thick data by integrating personal big-data (e.g., GPS data and social media data) with *human annotations* as the descriptions of their situations (e.g., the user answers to machine questions [64, 109], human self-reports [137, 143, 158], or information from the phone's personal contacts or agenda [144, 159]). Hence, various data sources are formalized by observation context as a set of knowledge graphs, which are ordered in time by life sequence representation, generating the big-thick data.

We propose that two types of *enquiries* can be posed to big-thick data. The first is *observation enquiry*, which is distinguished by considering the different specific *observation purposes* on observation contexts, such as the objective location enquiry from the reference context or the human mood enquiry from the personal context. The second is *habit enquiry*, which aims to learn a certain repeated activity in a life sequence, such as a person's spatial habit of always be-

[1]https://www.openstreetmap.org.
[2]https://www.agid.gov.it/en/data/spatial-data.

ing at university, identified by the subsequences involving university in her/his life sequence. Given the absence of a standard evaluation for big-thick data, we propose the metrics of *data size and connectivity* and *enquiry purpose feasibility* to assess the quality of big-thick data. We conduct an initial assessment of the proposed approach via the SU2OSM case study[3], where we unify a dataset, called SmartUnitn2 (SU2) data[4] describing the daily life data of a large sample of university students, with the OSM-Trentino data[5] extracted from the OSM data within northeast Italy[6].

1.4 List of Publications

This thesis is based on the following publications:

- Giunchiglia Fausto, **Xiaoyue Li***. Big-thick data generation via reference and personal context unification[C]//European Conference on Artificial Intelligence (ECAI), 2024.

- Busso Matteo*, **Xiaoyue Li**. A context model for collecting diversity-aware data[C]//Workshops at the Second International Conference on Hybrid Human-Artificial Intelligence (HHAI), 2023.

- Giunchiglia Fausto, **Xiaoyue Li***, Busso Matteo, Rodas-Britez Marcelo. A context model for personal data streams[C]//Asia-Pacific Web (APWeb)

[3]A description of the big-thick project can be found at the link: https://datascientia.disi.unitn.it/projects/su2osm/, which provides descriptions of the SU2OSM case study and the link to download the generated SU2OSM dataset.

[4]A description of the project for collecting the SU2 dataset, General Data Protection Regulation (GDPR) compliant, can be found at the link: https://datascientia.disi.unitn.it/projects/su2/. The link also describes the process for downloading the data.

[5]OSM-Trentino data can be downloaded from: https://drive.google.com/drive/folders/1Qt_Ik7ZYEph_9yPyjB4tJN4xYndMfJiE?usp=sharing.

[6] Northeast Italy OSM data can be downloaded from: https://download.geofabrik.de/europe/italy/nord-est.html.

and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data. Cham: Springer Nature Switzerland, 2022: 37-44.

- **Xiaoyue Li***, Rodas-Britez Marcelo, Busso Matteo, Giunchiglia Fausto. Representing habits as streams of situational contexts[C]//International Conference on Advanced Information Systems Engineering (CAISE) Workshops, 2022: 86-92.
- Andrea Bontempelli, Marcelo Rodas Britez, **Xiaoyue Li**, Haonan Zhao, Luca Erculiani, Stefano Teso, Andrea Passerini, Fausto Giunchiglia. Lifelong Personal Context Recognition[J]//The First International Workshop on Human-Centered Design of Symbiotic Hybrid Intelligence co-located with The First International Conference on Hybrid Human-Artificial Intelligence (HHAI), 2022.
- Giunchiglia Fausto*, Marcelo Rodas Britez, Bontempelli Andrea, **Xiaoyue Li**. Streaming and Learning the Personal Context[C]//Twelfth International Workshop Modelling and Reasoning in Context (MRC) co-located with IJCAI 2021.

The author has further contributed to the following publication:

- Daqian Shi*, **Xiaoyue Li**, Fausto Giunchiglia. KAE: A Property-based Method for Knowledge Graph Alignment and Extension[J]//Journal of Web Semantics, 2024.

1.5 The Structure of the Thesis

The thesis is organized as follows:

In Chapter 2, we introduce the state of the art of the fields of *big-thick data*, *context* research and *data integration*. The big-thick data research presents the study and limitation of *big data* and *thick data* and demonstrates the necessity and challenges of generating big-thick data. Given the context notion used in our method, we explore the application of context within three main research domains: Knowledge Representation (KR), Internet of Things (IoT), and Human-Computer Interaction (HCI). Finally, we introduce data integration research, especially knowledge graph integration.

Chapter 3 first defines the life sequence as a sequence of observation contexts during a certain period, representing a person's life. The life sequence paves the way for the formalization of human habits. Then, we introduce the general observation context, clarifying its basic representation and the two partial views of it, namely the objective view and the subjective view.

Based on the objective and subjective views of the observation context, Chapter 4 first formalizes the definition of the reference context, then from that, defines the personal context respectively.

Chapter 5 models the general reference context and personal context using teleontology and teleology. Based on the models, we formalize the general context unification process for unifying reference and personal contexts to generate observation contexts. This chapter provides the method for formalizing and integrating various data sources into unified knowledge graphs.

Chapter 6 introduces the observation enquiry and habit enquiry posed to the generated big-thick data and subsequently proposes several evaluation metrics to assess the quality of big-thick data.

Chapter 7 implements our methodology with the SU2OSM case study. This case study unifies the OSM-Trentino dataset with the SU2 data, generating the SU2OSM big-thick data. We evaluate the quality of the generated SU2OSM data by comparing it with the original OSM-Trentino dataset and the SU2 dataset using the evaluation metrics proposed in Chapter 6.

Chapter 8 presents the conclusion, including the summary, benefits, as well as the limitations and future work.

Chapter 2

State of the Art

The goal of this thesis is to represent the *life sequence*, composed of *observation contexts*, in order to formalize and integrate diverse data sources into *big-thick data*. This chapter introduces the state of the art in related fields.

In Section 2.1, we introduce three types of data: *big data*, *thick data*, and *big-thick data*, along with their main applications. This section presents the advantages and limitations of big data and thick data, demonstrating the necessity and challenges of generating big-thick data in current research.

Section 2.2 presents the notion of context as utilized in three primary research domains: Knowledge Representation (KR), Internet of Things (IoT), and Human-Computer Interaction (HCI). Section 2.2.1 introduces the notion of context in KR research, emphasizing its role as a fundamental tool for knowledge representation and reasoning. Section 2.2.2 discusses the interaction between context and real-world data streams in IoT research. Data streams enable people to identify their contextual information. In turn, the context can be used to formalize and model the data streams. Section 2.2.3 explores how context can be an effective tool to enhance human-computer interaction.

In Section 2.3, we initially introduce the methodologies for data integration, followed by a discussion of the knowledge graph integration. This section implies the feasibility of organizing and integrating big-thick data as sequences of knowledge graphs.

2.1 Big-thick Data

Big data: The concept of *big data* emerged with the explosive growth of large-volume digital records (e.g., sensors) collected from smart devices, such as smartphones and smartwatches. It is characterized by its *massive Volume*, *highly Velocity* and *Variety of data forms* [127], collectively known as the three *Vs*. These features of big data have driven AI research to extensively analyze such data to reveal hidden patterns and correlations, particularly related to human behaviors. As an example among many, human activities can be predicted using the motion or inertial sensors (e.g., accelerometers and gyroscopes) [25, 124, 135, 141]. In-vehicle sensors, in conjunction with roadside infrastructure (e.g., cameras and sensors) have been employed in intelligent transportation systems [69]. Additionally, sensors for light, temperature, microphones, and accelerometers can be used to infer human anxiety and autism [79].

While big data facilitates computational and predictive analytics that far exceed human cognitive capabilities, it does not undermine the importance of human observation within contexts [4]. Big data is always too poor in its contextualization to describe many interesting aspects of people [8], such as their behaviors, roles, and motivations for actions. As [8] noted, 'with many interesting variables unavailable, people are, at best, thinly described.' The 'out of context' usage of big data diminishes their meaning and value [18], making it challenging for researchers to fully analyze it. This is why Bornakke and Due, in [16], call big data as *big-thin data*.

Thick data: Differently from big data, *thick data* is derived from observations of the contexts in which human behaviors occur, e.g., people's interviews and self-reports [16]. The thick data can be the main data source represented in the situational context model [52]. The idea of thick data is derived from the concept of *thick descriptions* in Anthropology [46], which refers to ethnographically collected and analyzed observational, contextually rich. Its limitation in size is due to the practical difficulties associated with its collection, thick data typically exists in small volumes, referred to by [26] as *small-thick data*. For example, in the SmartUnitn2 project[4], researchers collect recordings of questionnaires from students, as well as develop the iLog app [156] installed on smartphones to collect the time diary data from students, which records the information of students' daily situations, such as activity, location, social people, mood, etc [56]. These (thick) data are far smaller than the collected sensor data. Despite the small volume, thick data is still rich enough in content to enable researchers to understand and reflect upon the scenario within which people act and behave.

Big-thick data: Recent works have suggested the integrated usage of big data and thick data [9, 12] and first introduced the concept of *big-thick data* in the Human-Computer Interaction (HCI) community. For mixing big data and thick data, *big-thick blending* concept is proposed by Bornakke and Due [16], they proposed a methodological framework to mix big data and thick data fully consider the complementary of these data sources. From the view of big-thick data collection, the IoT devices allow for the generation of big-thick data, as already made evident by the work on crowd-sensing [2, 89] and mobile self-reports [50, 158]. However, most studies of the generation of big-thick data remain general programmatic theories and methodologies, but there is a conspicuous absence of research that actually generates extensive and interconnected big-thick data.

Personal data is beneficial for human-centered artificial intelligence. A collection of personal datasets (e.g., Reality Mining project [39], Kraken.me [110], Labels [111], ExtraSensory[1]) have been collected and utilized to analyze intriguing insights into human life, such as their daily stress [13], moods [77, 6, 98], depression [151] and health [119, 154]. For instance, the Reality Mining project [39] utilized smartphone sensors (e.g., Bluetooth proximity logs, GPS, and cellular tower IDs) and people labels (e.g., locations) to examine students' daily activities, including their social networks and significant visited locations. Similarly, the StudentLife project [149, 71] employs smartphone sensors in conjunction with human annotators to infer, under varying workloads and term progress, students' behaviors, e.g., mental health, usage of apps, and academic performance.

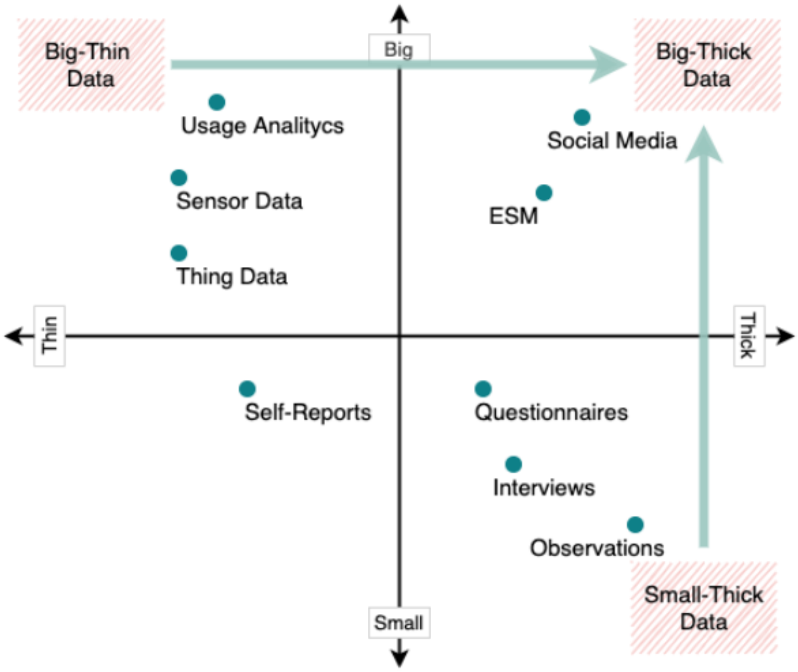

Figure 2.1: Big-thick data [26].

Figure 2.1 [26], as an evolution of Figure 2 in [16], puts various types of data in the dimensions of thin/thick and extensive/small, which well demonstrates

[1]http://extrasensory.ucsd.edu/.

the intuition underlying the idea of big-thick data, namely big-thick data as the blending of: Big-thin data, e.g., usage analytics, sensor data, thing data (from Internet-of-Things (IoT)); Small-thin data, e.g., self-reports; Small-thick data, e.g., questionnaires, interviews, and observations; and Existing big-thick data, e.g., social media data and data collected by the Experience-Sampling-Method (ESM) [134, 92, 109]. In this thesis, big-thick data is the *(multi)context-aware* data. The notion of big-thick data, together with its computational definition, design of knowledge reasoning, and implementation in terms of life sequences, extends and operationalizes the idea of big-thick data blending.

2.2 Context Research

2.2.1 Context in Knowledge Representation

The notion of context has a long and storied history, which has been extensively studied across multiple research areas. To the best of our knowledge, it was first introduced in the field of Knowledge Representation (KR), beginning with John McCarthy's Turing Award lecture [106]. McCarthy observed that the results of axiom-based reasoning are significantly influenced by the generalization of context. He proposed formalizing the concept of context and incorporating it into axioms and non-monotonic reasoning to better manage this variability. In this framework, the context was proposed as a key element for the formalization of commonsense reasoning. As an application of the context defined by McCarthy [106], the Multi-Context Systems (MCSs), as a heterogeneous framework, are widely used for formalizing and reasoning in different contexts. Giunchiglia et al. formalized the context and defined a set of reasoning inside a context in their monotonic multi-context system [57, 62]. Later on, Brewka et al. [21, 23, 22] develop the MCSs, where heterogeneous contextual information and knowledge are formalized as multiple contexts (i.e., components of agents or modules), with information flows (commonly represented as bridge rules) allowing for formalizing and reasoning across these contexts. In nonmonotonic MCS, existing knowledge can be retracted or revised when new knowledge is added [19, 20]. This is particularly valuable when the scenarios are incomplete or change over time. Meanwhile, the monotonic MCS only allows for adding new knowledge and not removing and changing existing knowledge, which is simpler, more stable, and predictable.

In [57], the context was introduced as personal representations of real-world situations. This work coincides with the following *magic box* illustration proposed by Giunchiglia et al. [48], later recollected in Brewka et al. [19]. The magic box illustration revealed that the underlying intuitions of the formalization and reasoning in context are from different partial views. This work [48] first introduced the idea that multiple subjective views exist in the same objective reality, where a *view* was defined as a partial *representation* of the world consisting of a set of facts describing the people's current perspective. Motivated by that, the context notion is used to model the world around people from their subjective views [52, 63, 121, 55, 158, 61]. A set of case studies [157, 158, 121, 55, 61] have proved that such context is a useful tool for collecting personal data streams [61], as well as is a benefit for understanding and predicting human behaviors. For instance, Shen et al. [121] show that considering subjective rather than only objective features (e.g., situational information about time, location, activity, and social relations) in context can improve the performance of machine learning methods. In our recent work [55], we propose a streaming context to represent a comprehensive personal context of real-world human lives, such as locations, people's internal states, and actions, which can be well integrated with machine learning methods by designing Direct Acyclic Graph (DAG).

However, as defined by Guarino and Guizzardi [67], a *scene* is as everything happens in a suitably restricted spatio-temporal region. As in [108], a *situation* is defined as the complete state of the universe at an instant in time. We observed that, although there are many benefits to using subjective views to describe the context, as we discussed above, we cannot ignore objective perspectives to describe the context. Because subjectivity is always partial, may be incorrect, and even with human bias, as discussed in [157], 'humans can make mistakes and more often than not unknowingly or unwittingly, due to their biases, which alter their perception and judgment of reality'. This thesis proposes that the *personal context* notion represents the context from the people's perspective, while the *reference context* notion represents the situation (from objective views) as it actually is. The key difference, using the terminology in [67], is that a scene is a perdurant (which is composed of temporal parts [52]) while the reference context is an endurant (which is wholly present whenever it is present, and it persists in time while keeping its identity [52]). The connection between the two notions is in the fact that a scene can be seen as the perdurant describing the evolution in time of the reference context.

2.2.2 Context in Internet of Things

In the Internet of Things (IoT) field, *context recognition* task aims to identify people's contextual situations from the data collected from smart devices such as sensor time-series data using AI methods such as machine learning and deep learning algorithms [76], where context is then used to learn about the user behavior. In this work, the possible context dimensions include locations, activities, body posture, and more [142]. Sensors are explored to identify people's specific activities, e.g., accelerometers can be used to detect people's physical activities, e.g., walking, running, or steps [142, 131, 7]. Multiple sensors can detect more complex contexts, e.g., people's locations [115, 120] and activities [7, 65, 17]. For example, a Multi-Layer Perceptron (MLP) [142] is designed to use multi-modal sensors (e.g., smartphone accelerometer, smartwatch accelerometer, phone gyroscope, phone audio) as features to simultaneously predict many diverse context labels, including people's body-states, home activities and environments. Umapathi et al. [140] apply the GPS data attached to student ID cards to identify student locations via a SQL-coded database. In addition, many studies use human-provided labels together with sensors to infer the contextual information around people, such as human behaviors, including physical activity [105], assessment personality states [82, 126], and visiting points of interest [36, 114]. For example, the ExtraSensory used a variety of sensors (e.g., acceleration, gyroscope, and audio) as well as ground truth context labels (e.g., activity labels and mood indicators) to predict human activities [141, 131]. (For similar research see the Reality Mining project [39, 38], StudentLife [150, 151], Device Analyzer [148, 72]) and GLOBEM [152]).

At the boundary between IoT and KR, Giunchiglia et al. keep working on modeling context to formalize and organize personal data streams [52, 63, 14]. Our recent work [55, 61] builds upon these previous works, proposing that define the notion of *streaming/situational context* to formalize and collect personal data streams. This thesis learns from them and defines the life sequence composing of context streams with more comprehensive views, e.g., involving the context unification of reference and personal contexts; it makes the data integration process more flexible, hence allowing multiple types of enquires from the generated data. All the works are explored to generate high-quality big-thick data that is needed to track human daily life and support meaningful human-in-the-loop human-machine interactions.

2.2.3 Context in Human-Computer Interaction

The notion of context has also been extensively studied by the Human-Computer Interaction (HCI) community. Dey et al. provided a widely accepted definition of context as 'any information that can be used to characterise the situation of an entity' [1, 33]. In this field, the context notion is applied for context-aware applications to facilitate richer and easier human-machine interactions [34]. For instance, Aseniero et al. [5] developed a personal visualization tool called the

Activity River, which enables individuals to visualize historical and contextual data such as activities, compare their goals and achievements, reflect on their self-defined activities, etc. The Smart Agora software [118] was developed to visualize complex experimental scenarios on smartphones using geolocated survey and phone sensor data from participants, which highly improved crowd sensing. The significance of this work lies in the necessity for high-quality interfaces for human-in-the-loop human-machine interactions, while the HCI community focuses on developing such high-quality interfaces. Although these studies are less relevant to the methodology of defining the life sequence by unifying contexts in this thesis, they offer us the popular applications and strong motivation for the big-thick data generated in the real world. The big-thick data applications in HCI will be one of the subsequent focuses following the generation of big-thick data.

2.3 Data Integration Research

Data integration refers to the process of combining heterogeneous data and providing the user with a unified view of these data [28, 94]. Heterogeneous data integration is classified into two main types of integration: logical integration and physical integration [122, 101]. Logical integration simply provides common virtualization access to diverse data sources. For instance, the Ontology-Based Data Access (OBDA) approach [87, 90, 91] establishes a high-level global conceptual schema, namely, an ontology, of already existing data sources. The ontology is linked to the data sources through declarative mappings. OBDA separates users from the data sources and provides them with a convenient query vocabulary. However, Logical integration benefits query information for users, while it relies on knowledge representation expertise and not extracting and transforming from data. Physical integration, commonly using data warehouse techniques, e.g., Extract-Transform-Load (ETL) [146, 145], migrates and integrates heterogeneous data sources into a global data schema, in which various data sources are stored in a centralized database. As discussed in [101], although physical integration is more costly than logical integration, it benefits a lot in the long term because it provides not only efficient data access but also data management. This is a motivation for this thesis to integrate real big-thick data in a physical integration manner.

Knowledge-based integration is regarded as one of the efficient integration methods because of the excellent semantic interoperability of the knowledge bases (e.g., WordNet [112], Ontologies [88]), which is due to the schema alignment and mapping tasks[100]. Knowledge Graphs (KGs), effectively extract and organize complex information from heterogeneous data into knowledge in the real world [139]. KGs typically achieve physical data integration, wherein information from multiple diverse sources is combined in a new graph-like representation [75, 73]. Learning and representing human knowledge are vital tasks in the AI area; hence, knowledge graphs are widely applied in AI research, e.g., recommender systems, question-answering, and information retrieval to benefit human life and society [117]. Due to the great significance of knowledge graphs in processing heterogeneous information within a machine-readable context [117], KG construction and integration is a useful way to integrate heterogeneous data in the KR area. For example, a novel automatic framework is proposed for heterogeneous learning data sources fused into knowledge graphs, which improve students' and teachers' cognition of courses and disciplines [97]. Similarly, the *iTelos* [11, 54, 53] was developed as a general stratified data integration methodology, which involves building the knowledge graph schema to integrate diverse data sources into knowledge graphs.

This thesis represents observation contexts with KGs, both at the schema and at the data level. This is achieved by implementing *context modeling* (reference and personal contexts) and *context unification*. Context unification is a form of configurable KG integration process [51], implemented according to the *iTe-*

los methodology [11, 54]. Different from the *schema alignment* task aiming to determine the correspondences between concepts in knowledge graph schemas [41, 132]. Here, we use a novel concept of context unification[2]. Compared with the alignment task focusing on recognizing alignment relations, unification focuses on a unified result, i.e., we finally generate a set of unified observation context KGs as the elements of big-thick data. Context unification indicates not only how we distinguish between reference and personal context but also the way we do context or graph alignment. It is a flexible process integrating multiple KGs into a single KG and allows for multiple diverse research questions based on the same data sources.

[2] As Webster's Dictionary states: 'Unification is the act, process, or result of unifying: the state of being unified'.

Chapter 3

Life Sequence as Observation Contexts

As outlined in the state of the art of context research in Section 2.2, the *context* is an efficient concept for generating big-thick data, because of its strong capabilities in knowledge representation and reasoning, formalization of the real-world data streams and enhancement of human-computer interaction. This chapter aims to represent a person's life by the notion of *life sequence*, which is a sequence of *observation contexts* observing the scenarios in real-world life.

In Section 3.1, we formalize the life sequence of a person as a sequence of observation contexts she/he is in time, which represents her/his real-world life. The life sequence definition opens the possibility of learning human habits, we formalize the *habit* definition to learn different human habits. Similarly, the life sequences of a specific group of people can be used to learn the habits of this group, e.g., the points of interest of students.

Section 3.2 represents a general notion of observation context. Specifically, Section 3.2.1 represents the observation context as *locations* and *events* with some fundamental populations, including *entities*, *objects*, *functions* and *actions*. We explain these concepts and show their relations. However, the underlying intuitions of the formalization and reasoning in context are from different partial views [19, 48]. Section 3.2.2 introduces the *objective view* and the *subjective view* for the observation context, which serves as the basis for generating the reference context and personal context in Chapter 4.

3.1 Life Sequence

3.1.1 Life Sequence Definition

An *observation context* (C) represents a real-world scenario a person is in, a sequence of such observation contexts happens over time in her/his life. We define the *life sequence* (LS) of a person P as a sequence of *observation contexts* during a certain period of time as follows:

$$LS(P) = \langle C_1(P), C_i(P), \ldots, C_n(P) \rangle; \quad 1 \leq i \leq n \tag{3.1}$$

where C_i is the i_{th} observation context of a person P. Considering that a person can only be in one location at one time, we assume that a person can be in only one observation context at any given time. Therefore, LS is a sequence of her/his observation contexts, occurring one after the other in strictly sequential order. The LS represents a strict sequence of observation contexts of a person in time, covering the entire period under consideration based on the specific purpose. For instance, the case study in Chapter 7 considers a four-week period as the time period of student life sequences, which aligns with the time diaries data collection period in the SmartUnitn2 project. Another example is the complete life sequence of a person, spanning from birth to the present and extending to death.

Figure 3.1 shows the life sequences $\{LS(P_1),\ LS(P_2),\ \ldots,\ LS(P_n)\}$ of a set of persons $\{P_1, P_2, \ldots, P_n\}$ from time T_a to T_b, where the length of these life sequences is $\{l_1 + 1, l_2 + 1, \ldots, l_n + 1\}$. Life sequences of a group of people, e.g., students and workers, can be used to learn from them in a certain period, e.g., student behavior patterns, habits, and lifestyles during the COVID-19 pandemic period. Additionally, by analyzing the life sequences within the same temporal system or spatial system, it is possible to identify time-related or

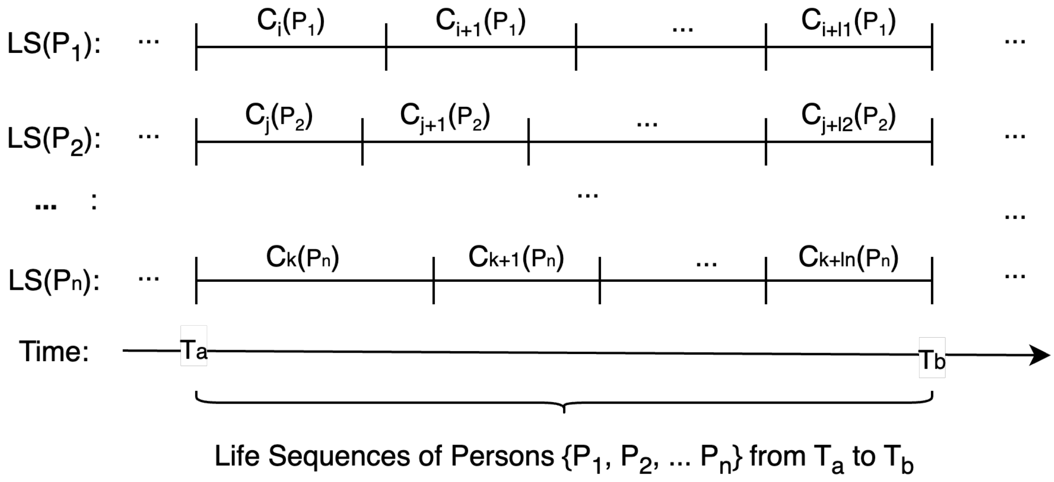

Figure 3.1: Life sequences of a group of persons.

spatial-related patterns among the involved individuals, such as students' points of interest during weekends or the frequency that students visit the University of Trento.

3.1.2 Habits from Life Sequence

Merriam-Webster defined habit as *a behavior pattern acquired by frequent repetition or physiologic exposure that shows itself in regularity or increased facility of performance*, e.g., smoking and drinking habit. Habit is a learned behavior by practice, which is regular or predictable, while performed almost automatically [125]. Based on the life sequence definition in Equation (3.1), we define *Habits* of a person, $H(P)$, as:

$$H(P) = \langle LS(P)_{a_1} : a_1, \ldots, LS(P)_{a_n} : a_n, f \rangle; \quad 1 \leq i \leq n \tag{3.2}$$

where $LS(P)$ is a life sequence of a person P, and a_i is an *activity* of P, such as *being in a location*, *being involved in an event*, *being together with a person*, *using an object*, *performing an action*. The $LS(P)_{a_i} : a_i$ is a subsequence from $LS(P)$ involving a_i, which may have time gaps between its contexts. Our previous paper [96] shows five basic types of habits: *spatial habits*, *temporal habits*, *social habits*, *material habits*, and *action habits*. Examples of these types of habits are always at home, cooking, being together with parents, carrying a smartphone, and walking, respectively. The f in Equation (3.2) is the *frequency* threshold that determines whether H occurs. There are various levels of flexibility in defining f, such as using a number or a percentage. For example, f being five means that the activities occurring no less than five times in the life sequence are considered habits. The different settings of f influence the results of habit identification, even from the same life sequence. For example, consider the life sequence of Bob is:

$$LS(Bob) = \langle C_1(Bob), C_2(Bob), C_3(Bob), C_4(Bob)\rangle$$

where Bob is with parents in all contexts, and with his relatives in $C_1(Bob)$, $C_2(Bob)$, and $C_4(Bob)$. If the habit frequency is set to 4, a social habit can be derived as:

$$H(Bob) = \langle\langle C_1(Bob) : Parents, C_2(Bob) : Parents, C_3(Bob) : Parents, C_4(Bob) : Parents\rangle, 4\rangle$$

indicating Bob is always together with her parents. While supposing the habit frequency is set to 2, the two social habits can be identified as:

$$H(Bob) = \langle\langle C_1(Bob) : Parents, C_2(Bob) : Parents, C_3(Bob) : Parents, C_4(Bob) : Parents\rangle, \langle C_1(Bob) : Relatives, C_2(Bob) : Relatives\rangle, \langle C_4(Bob) : Relatives\rangle, 2\rangle$$

showing Bob is often being together with her parents and relatives.

3.2 Observation Context

3.2.1 Observation Context as Locations and Events

Guarino et al. [67] define a *scene* as everything that happens in a suitably restricted spatio-temporal region. Motivated by that, let us think of the *world* as a set of three-dimensional spatial regions that we call *locations*, e.g., universities, supermarkets, forests, mountains, etc., with their GPS coordinates or proximity. These locations are infinite and continuously evolving slowly in time. Inside each location is an infinite set of, continuously evolving over time, mono-dimensional temporal regions that we call *events*, e.g., celebrations, presentations, etc. Each event has a *duration*, defined by a *start-time* and an *end-time*, which are always recorded by timestamps. Events intuitively represent how that location internally evolves in the durations. We represent the world using *observation context*, as locations and events.

Observation Context Populations: Motivated by the former teleology work used in knowledge representation [58], we define the observation context as populated by a set of *entities*, which are anything that we can think has spatial and temporal extensions, such as people, cars, buildings, cities, etc. We think entities are conceptualized as being identified by the following two fundamental components:

- *Objects* define the spatial regions occupied by entities, which are characterized by their own properties. For example, a student entity corresponds to her/his body object, which has properties of height and weight. A university entity corresponds to a building object, which has properties of color, height, and coordinates.

- *Functions*[1] define a specific set of *expected actions* for entity. Entities may

[1]The notion of function is somehow related to Guarino et al.'s notion of *role* [68], with the key difference that functions are associated with a set of characterizing actions.

dynamically change their states in time, e.g., their positions, shapes, speed, or proximity with other objects. A change of state by an entity is called an action (of that entity), e.g., a person talks (as an action) with a dog, or a car moves (as an action) on a road. Functions of entities represent the expected actions of entities, these expected actions or behaviors can be predicted or reasoned by the repeated actions of entities. For example, a person is a customer of (as a function) a supermarket because she/he always conducts customer actions such as selecting goods and paying a bill. The actions of entities evolve and change; hence, the functions of the same entities change over a long period but commonly are static over a short period. For example, a person is a supervisor of (as a function) another person in four years and also can be a colleague for (as a function) each other in the later ten years.

There are some key observations of the mapping among objects, entities, functions, and actions:

- One object may correspond to one or more entities. This is because there exists a many-to-many relationship between objects and functions. Any such combination of objects and functions defines an entity, but not always the same. We can also think of (the body of) a person as an object supporting many functions, corresponding entities and actions, e.g., a student reading a book, a driver driving a car, a customer paying the bill, and so on.

- Different entities defined by different objects may have the same function or action. For example, a car and a bus are two different entities associated with two different objects, and both of them implement the same function of a vehicle, as well as perform the action of carrying people traveling.

Observation Context Representation: Figure 3.2 illustrates the relations among populations of observation context, defined by locations and events. We derive some observations within the figure. An object of an entity represents the spatial region of the corresponding entity. We say that an entity `PartIn` (or `Populates`) a location if its object is `Inside` a region, which identifies the location. For example, the body of a person (as an object) of a student (as an entity) is located in Trento (as a location). An entity or object exhibits `Spatial Relation` with others, such as a person near a university or the body of a person in front of a building. As previously indicated, an entity possesses functions and actions. These actions are the cause of the mechanism by which events change in the world. At any given moment in time, an entity `PartIn` (or `Populates`) an event if it exists within the spatial region of the location of the event and also inside its temporal region. For instance, a student traveling (as an event) in Trento on a Sunday.

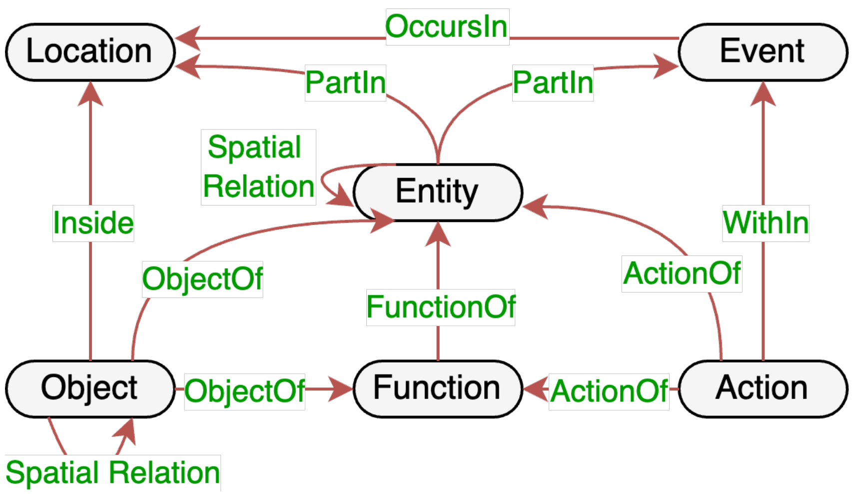

Figure 3.2: Observation context as locations and events, populating with key components.

Entities are associated with the *spatial coordinates* and *temporal coordinates* for each location and event they populate. These coordinates, derived from the locations and events, are typically not the only possible choice, such as GPS coordinates and the local time of a time zone. This provides the spatio-temporal coordinates for any possible triple $< location, event, entity >$. These coordinates provide relatively precise positions for any entity in space and time. Notice that, according to the definitions above, the positioning of entities is only inside the involved locations and events, without the necessity of being located by the external spatio-temporal coordinate system. This reflects the idea that, for instance, supposing you are at home now, what you do depends only on the entities within the home, such as the television, the refrigerator, and the stove, based on where they are positioned, and their evolving states (such as the television is turned on, the refrigerator is getting empty, the stove is lit), rather than on the location of the apartment in the city. And similarly for time.

3.2.2 Objective and Subjective Views

People are entities in contexts that have an internal representation of the world and use it to reason about it and take action. Obviously, different people possess different representations of the world, even when observing the same scenario, because of their different experiences, different subjective perspectives, and so on. For example, when faced with the same subway scenario, a thief is more likely to notice passengers' valuables and wallets, whereas a passenger might focus on intriguing magazines, notifications on their phones, and people acting suspiciously. This internal representation of the world from people is organized in *contexts*, 'a context is a theory of the world which encodes an individual's subjective perspective about it' [57]. This type of representation of contexts has been widely applied at the beginning of our work [27, 61, 96], where we defined that a *situational context* represents a real-world scenario from the perspective of a specific person.

However, we observed that the representation of context is not limited to this partial subjective view. Those contexts encoding people's internal representation are not *situations*, as from [108], '... a situation is the complete state of the universe at an instant of time'. In other words, people have partial and potentially incorrect perspectives of the world, situations establish what actually is the case. Thus, situations are able to provide a single objective reference point for comparing the contents of contexts and measuring what is the case and what is not.

We introduce the *objective view* and *subjective view* of the observation context, which are the two different representations for perceiving the world. The objective view is what happens in a situation from an objective view of an all-observing third party. From this perspective, a specific fact that is not known or agreed upon is not objective but subjective, such as the fact that two people are friends or, who is right or wrong in a discussion. It is worth noting here that our discussion of subjectivity and objectivity is confined to a relatively limited form. Specifically, the objectivity with reference to the fact that people have *partial knowledge* of the world, thereby excluding the issues related to the subjectivity, such as opinions or sentiments. Thus, we claim to have objective knowledge if everybody in the target audience knows or has the means to know about it. Thus, for different observation contexts of different target audiences, there may be different objective views, even for the same scenario. The subjective view is what happens in a situation from the subjective view of a person in this scenario. Different objects in a situation can perceive different personal views of the same scenario. We say subjective knowledge is known only by a few specific subjects while others may or may not be unaware of it unless informed by those who possess it. Thus, for instance, my home address is my personal knowledge, while the name of the street where I live is objective knowledge. The objective view and the subjective view of observation context are the motivations to define the notions of *reference context* and *personal context* in the next Chapter 4.

Chapter 4

Reference and Personal Context Definition

As the *objective view* and the *subjective view* for the observation context introduced in Section 3.2.2, this chapter aims to precisely define two types of context: the notion of the *reference context* (C_R) and that of *personal context* (C_P), respectively.

Section 4.1 aims to define C_R according to the objective view of the observation context. A C_R is any context that carries an objective representation of situations, as defined in [108]. C_R represents the world from an objective, person-independent, all-encompassing view of a third-party observer. Initially, Section 4.1.1 defines the notion of *spatial-temporal context* C_S, characterized by a *reference spatial region* S, which has a set of (relatively) static *spatial objects*. Subsequently, Section 4.1.2 defines C_R as a projection function from C_S enforcing objectivity and is represented as the *location entity* L_R (associated with the spatial object of S) as well as a set of entities in C_R.

Section 4.2 aims to define C_P according to the subjective view of the observation context. A C_P is defined as any context that carries a subjective representation of the world, in accordance with the definition of context from [57]. Section 4.2.1 demonstrates how to go from C_R to C_P, wherein C_P is initially defined by identifying C_R and a specific person me. C_P carries a subjective representation of the world, encoding a subjective view from me. Section 4.2.2

then refines and finalizes the definition of C_P in Section 4.2.1 by formulating the events, locations, entities, functions, and actions from a subjective view from me of C_P. This section also shows the key processes and observations involved in constructing C_P.

4.1 Reference Context

A reference context provides an objective, person-independent, all-encompassing view of a third-party observer, which represents the situation within which users are operating. As we said above, it is 'an objective' representation of the situation. Notice that, we wrote about 'an objective', not 'the objective'. This is due to the fact that it is inherently impossible to construct a complete description of reality, in addition, objective representations of the same situation can differ in multiple dimensions, such as the level of detail, the level of partiality, the viewpoint, the entities being considered, and so on [58, 24].

We build the reference context based on a specific *purpose*, following [54]. This purpose assigns a specific target use, such as recognizing a specific action or answering the query about the functions of people in natural scenes. Different purposes result in the construction of distinct reference contexts, even within the same situation. The process is similar to the manner in which a person might be posed multiple questions, as a result, she/he focuses on different fragments of her/his knowledge. For example, when asked 'Does the current location usually host similar events?', the person reflects on her/his experiences of events that occurred at that location. When asked, 'How does her/his mood correlate to one or more people, or their roles, when being together in a certain location at a certain time?', the person focuses on her/his past mood states and the people interacted with in that location.

4.1.1 Spatial-temporal Context Definition

We begin by selecting a *reference spatial region* S, defined as the set of points located inside the *boundary* of S. For example, Trento is defined by a set of points with latitudes, longitudes, and altitudes, (x, y, z), annotated by a map. The boundary of S delineates the *inside/outside* of S, with reference to any location out of the boundary, is excluded from consideration in building C_S, because it is not purpose-relevant. For S, we have:

- A *reference observation period* ΔT_R is set for S, such as a month or two weeks, the time $t \in \Delta T_R$ measures how change happens within S.
- The S contains a set of spatial regions, perceived as *objects* $(O_1, ..., O_n)$, such as the University of Trento, BUP library, and MUSE museum, where each object is defined by its boundary. Hence, we have:

$$O_i \subseteq S \quad \text{with} \quad 1 \leq i \leq n, \tag{4.1}$$

Then we define a *spatio-temporal context* C_S as:

$$C_S(t) = \langle S(t), \{O(t)\}_S \rangle \quad \text{with} \quad t \in \Delta T_R \tag{4.2}$$

Where $\{O(t)\}_S$ is a set of objects O_i that satisfy Equation (4.1). Intuitively, Equation (4.2) demonstrates that C_S consists of a set of objects located within S, and the region S and the objects O_i change during the time period ΔT_R, with $t \in \Delta T_R$. For any C_S, the selection of time period, objects, and objects' spatial properties is determined by the intended purpose.

Time variance is a default fact of life, as everything continuously changes and evolves. Actually, this is a major source of subjectivity, as the same object is different at different times, perceived by people. To deal with this problem in the objective representation, we require that, during ΔT_R, we have *time invariance* for defining the spatial-temporal context. The intuition of this setting is that the endurant concept in [45, 52]. Roughly speaking, an endurant is an individual wholly present and persists in time while keeping their identity, while lacking temporal parts, e.g., a building. This implies that ΔT_R must be chosen such that the location associated with the reference spatial region S and the selected objects within it remain unchanged. In other words, they must keep the same *selected spatial properties* based on the purpose, such as their positions, shapes, and colors.

4.1.2 Reference Context Definition

As we presented in Section 3.2.1, an object may correspond to one or more entities. We define e_i^j (with $j = 1, ..., m$) is the j-th entity associated to the object O_i. Based on the C_S, we define the *reference context* C_R as:

$$C_R = f(C_S(t)) = \langle L_R, \{e\}_R \rangle \quad \text{with} \quad t \in \Delta T_R \tag{4.3}$$

where L_R is the *reference location entity* (e.g., Trento as an entity), as an entity associated with the object of reference spatial region S (e.g., Trento as a region). The $\{e\}_R \subseteq \{e(O)\}_S$ is a set of entities of C_R, where $\{e(O)\}_S$ is the set of all entities associated with the objects of C_S. The $f()$ is a projection function from C_S to C_R enforcing the *objectivity*. The precise definition of $f()$ should be determined by the modeler. However, it is crucial to adhere to the following set of constraints when defining C_R, which also play a key role in the enforcement of objectivity.

- *Time invariance.* As C_R is defined from C_S, C_R is independent of time, with no dependency on t. This constraint implies that what is in C_R remains static, not changing over time. Such time invariability facilitates the establishment of a robust form of objective knowledge, simplifying the processes of sharing, managing, and scaling this knowledge. This constraint is feasible in practice based on two primary considerations. Firstly, the fact is that most spatial entities, such as buildings, mountains, streets, and cities, change very rarely. Secondly, the specific reference observation periods in the definition of reference contexts can be acquired by segmenting an extensive temporal framework into relatively smaller periods. Each segmented observation period corresponds to a time-independent reference context, via a time-aware versioning mechanism. For example, for a residence, two distinct reference contexts might exist: one spanning from the acquisition of the apartment to the commencement of renovation, and another from the completion of renovation to now. Actually, the time invariance can be extended to the management of what may be termed *objective events*, i.e., the events that are universally recognized and acknowledged, such as a new year concert has been organized long before its occurrence, but it is out of the scope of this thesis.

- *From objects to entities.* Similarly to define C_S, in the construction of C_R, the selection of C_R entities (i.e., L_R and $\{e\}_R$) along with their properties (including functions and actions) should fit the intended purpose. This constraint makes explicit the fact that any representation in C_R has a purpose and the C_R should be tuned for that. For instance, the city of Trento (as the L_R in C_R) can be used to define multiple reference contexts based on different purposes, where it can be as a traveling city providing a set of points of interest (as $\{e\}_R$) or a health-related city providing a set of health-related facilities (as $\{e\}_R$) [10]. This means, furthermore, allowing us to compute the relations among the entities in C_R, such as computing their spatial relations (e.g., a mountain near a river), distances, and reachability via their coordinates.

- *Completeness with respect to users.* The data describing C_R should be shared across all users and be equally accessible to all of them, such as via the Application Programming Interface (API). C_R should provide enough detail to ground all the diverse subjective views of all the different users, which allows users to ascertain whether their representation is consistent or inconsistent with that of the reference context, as well as with that of other users. This completeness constraint provides a general mechanism for comparing the multiple contexts of different users. In this case, the reference context serves as an oracle/reference coordinate, capable of discerning what is true and what is false among the facts in users' personal contexts. A practical consequence is that any representation that does not hold completeness property should not be considered in C_R.

- *Localization.* C_R should describe the smallest possible reference location satisfying the intended purpose. Localization is a key requirement for the reference context to work in practice, which avoids involving irrelevant and redundant information. For instance, if the focus is on what is happening at home, then the reference context should not include what is outside of the home. Notice that this localization constraint has consequences on the entities being involved in C_R. For instance, if one is interested in mobility in Trento, only some of the roads in Trento, not in Italy, will appear in the reference context.

4.2 Personal Context

4.2.1 From Reference Context to Personal Context

People are one type of entity in the reference context C_R, each of them possessing a unique subjective view of the world. We refer to a generic person as me, when we focus on how the person represents the world from her/his subjective view. We define a personal context, C_P, one for each me, encodes the person's subjective view of the world, which is any context that carries a subjective representation of the world. We say, C_P is built from C_R, but compared with C_R, it has some distinguishing features.

- $C_P = C_P(me, C_R)$. C_P depends on the both me and C_R. Thus, different people, each as me, may generate different personal contexts, even if they are in the same C_R. For instance, a student and a teacher in a classroom have distinct personal contexts focusing on studying and presenting, respectively. Similarly, a patient and a doctor in a hospital have different personal contexts concentrating on treatment and diagnosis, respectively. Moreover, different reference contexts may generate different personal contexts, even for the same me. For example, a person, as me, may have different subjective views at home when with parents in a rest-driven reference context or with friends in a party-driven reference context, which may build different personal contexts.

- *From inside to outside.* While C_R describes the entities that are *inside* the reference spatial region S, such as a street or a university in Trento. C_P describes the entities which are *outside* the object of me, but still *inside* S.

- *From no change to change.* Differently from C_R, C_P considers entities that change over time. This is evident from subsequent explanations, particularly concerning events $\{E(L_R, t)\}$ in Equation 4.4, and actions $A_{F_e}(t)$ in Equation 4.6, among others.

4.2.2 Personal Context Definition

Given the reference context C_R, the *personal context* C_P from a subjective of a person me, can be initially defined as:

$$C_P = \langle C_R, \{E(L_R, t)\}\rangle \quad \text{with} \quad t \in \Delta T_R \tag{4.4}$$

where $E(L_R, t)$ is a time-varying *event* perceived by me, involving me and occurring inside L_R. $\{E(L_R, t)\}$ is a set of such events. Events in C_P allow us to represent the time variance of the real world. In C_P, any two events may occur sequentially or in parallel, still within ΔT_R as well as in L_R. These events model how situations, as subjectively perceived by me, evolve over time. For example, a parade and the corresponding policing occur simultaneously in Trento, followed by a fireworks display in Trento. It is worth noticing that the most distinguishable difference between C_R in Equation (4.4) and C_P in Equation (4.3) is that C_P is populated by a set of continuously changing events. This distinction implies that, unlike the *time invariance* constraint of C_R, C_P is the formalization and generalization to a time-variant real-world setting, of the idea of *subjectivity as view point* of the *magic box* [48]. This setting is also in line with the concept of *perdurant*, as the individual that is composed of temporal parts in [45, 52]. Consequently, we recognize the existence of an unchanging, fully known objective reality, represented by C_R, and multiple time-varying subjective perspectives of this reality, each modeled by the set of events in which me is involved.

We refine the definition of C_P in Equation (4.4) in accordance with the requirements of *subjectivity* of events. We have two main considerations as follows:

- The first source of the subjectivity of event $E(L_R)$, is from the location L ($L \subseteq L_R$) where an event occurs, as well as the time interval ΔT ($\Delta T \subseteq \Delta T_R$) when an event happens. For instance, considering the city of Trento (as L_R) in the last four years (as ΔT_R), where a person as me can visit (as $E(L_R)$) the MUSE museum (as L) in Trento during the afternoon of yesterday (as ΔT). We can notice that when me is in a location at a certain time, she/he may just partially know about that location if the location is big enough (e.g., only the collection display area for visitors), and at the same time, she/he cannot access other locations, e.g., the street in front of the museum or the tree in a mountain in Trento. Hence, we refine Equation (4.4) into Equation (4.5) as:

$$C_P = \langle L_R, \{E(L, t)\} \rangle \quad \text{with} \quad t \in \Delta T \subseteq \Delta T_R,\ L \subseteq L_R \tag{4.5}$$

- The second and most important cause of the subjectivity of events involving me is that events are the result of the interactions among entities, these interactions occur due to the actions of entities, motivated by entities' mutual functions (where entities, functions, and actions have been introduced in Section 3.2.1). The subjectivity of events involves the subjective views of me on the interacted entities, perceived functions, and actions in the events. Thus, for example, for a questioning and answering event, different people as me, perceive different events, namely, a student as me asks questions to her/his teacher in a classroom, while a teacher as me explains questions to one of his students in his workplace. (This is also crucial contextual information for the interpretation of the actions occurring in events.) We model this form of subjectivity by defining $E(L, t)$ in Equation (4.5) as:

  $$E(L,t) = \langle L, \{\langle e, F_e, A_{F_e}(t)\rangle\}\rangle \text{ with } t \in \Delta T \subseteq \Delta T_R, L \subseteq L_R \quad (4.6)$$

  where $e \in \{e\}_E$ is any *entity*, including me, involved in $E(L, t)$, with $\{e\}_E \subseteq \{e(O)\}_S$ is the set of such entities, $\{e\}_E \cup \{e\}_R \neq \emptyset$. $F_e = \{f\}_e$ is the set of functions, the f represents the *function*, any entity $e \in \{e\}_E$ has functions with respect to any other entity $e \in \{e\}_E$. $A_{F_e} = \{a\}_{F_e}$ is the set of actions, a represents the *action*, each action with its time coordinates, that of any entity $e \in \{e\}_E$ performs towards any other entity $e \in \{e\}_E$, because of a function of them in F_e.

Thus, by combining Equation (4.5) and Equation (4.6), the final personal context C_P of me is defined as:

$$C_P = \langle L_R, \{\langle L, \{\langle e, F_e, A_{F_e}(t)\rangle\}\rangle\ \}\rangle \text{ with } t \in \Delta T \subseteq \Delta T_R, L \subseteq L_R \quad (4.7)$$

According to the equations above, the personal context of me can be constructed based on the following steps, each of which should be performed in line with a given specified purpose.

1. *Selection of S.* Select an identified reference spatial region S as shown in Equation (4.1), such as a city, a university, or a building.

2. *Selection of C_R.* Select a previously identified reference context C_R, that is, reference observation period ΔT_R, reference location L_R and entities $\{e\}_R$, as indicated in Equation (4.3). For example, the University of Trento reference context involves the University of Trento (as L_R) and a set of students (as $\{e\}_R$) during January 2024 (ΔT_R).

3. *Selection of me in C_R for C_P.* Select a set of persons as me, where each me with a corresponding personal context $C_P(me, C_R)$. For example, the students in the University of Trento are as me, each student is in the University of Trento reference context.

4. *Selection of events for C_P.* For each C_P, select the set of events, each event with its set of entities and corresponding objects, functions, and actions. For instance, for a student as me, one of her/his C_P involves taking a mathematics class event from 11 am to 12 am (as the time interval ΔT when the event happens), where he asks questions (as an action) for his teacher (as a person entity with the function of teacher) together with his friend (as a person entity with the function of classmates) in a classroom (as the location L where the event happens).

Here are some observations of the construction of C_P.

- The first observation is about how *subjectivity* is modeled. Notice that, in the general case, very few me know about the events involving the other me, especially because of their different locations and time periods. Actually, this consideration also applies to the entities involved in the same event. Specifically, most me do not know about the functions and the actions relating people, as well as the entities in general, to one another; these are substantially different across events. A me only builds his/her personal contexts because of only knowing his/her subjective views of the world, but may not know other people's subjective views even for the same situation. For example, me_1 has dinner in a restaurant, in most cases, she/he cannot know that me_2 is the wife of me_3 in the restaurant or me_4 has a meeting with me_5 in a library. This is a reason why we need to declare the *Completeness with respect to users* for building C_R in Section 4.1.2, which enables C_R as a reference system for connecting and reasoning in personal contexts of different people.

Subjectivity arises due to the *diversity of people*, and the extremely partial knowledge that any me has towards the behavior of the other me. For example, me_1 may know that me_2 and me_3 are colleagues but does not know they are couples. The only way to know about such contextual information is to ask people. This forms the key intuition behind the concept of big-thick data. No big data will ever be able to provide the kind of information that thick data can provide.

- The second observation is that we have assumed that inside each event, namely, the locations, functions, and therefore entities are not time-varying. Their time variance has to be managed by splitting the event into two or more events, following the similar process of *time invariance* of C_R described in Section 4.1.2. For example, in traveling, the transportation vehicle changes from a train to a bus, which may be split into two events: taking a train and taking a bus.

- The third observation is about the *location* of an event. The same spatial region can be assumed with the role (function) of a spatial entity or of a location, depending on the specific event. For instance, if me is driving home, then home is the entity that me needs to reach, where the me entity and home entity are located in the same broader location, such as the city of Trento. However, when me is at home, home becomes the smallest location outside me. Notice that any object can serve as a location. The discriminating factor is the space granularity of the event. For example, a building can be perceived as an entity inside a city in a traveling event, yet simultaneously serve as a location where a person is studying; while a person (body) can be considered as an entity in a studying event, but also act as the location where COVID-19 operates, and so forth.

- The fourth observation is about the *entities* $e \in \{e\}_E$ and their associated functions, involved in an event. Any e may be concurrently involved in multiple events, usually from different me, typically with different functions. Some functions of the entities $e \in \{e\}_R$, belonging to part of the objective knowledge, are known by all me, whereas, some functions of the entities $e \notin \{e\}_R$ are known only to some me. For instance, if me_1 meets me_2 at a University, me_1 can know that the University is the study place for students of me_3 and me_4, and will not know that it is the workplace for me_2. The same principle applies to the functions modeling personal relations.

- The fifth observation is about the *actions* $\{a\}_{F_e}$ of entities used to carry out an event. Unlike the second observation above, actions change frequently in time with a duration $\Delta t \subseteq \Delta T$. This variability makes Human Activity Recognition (HAR) particularly challenging. However, HAR becomes much easier with big-thick data owing to the extra information provided by humans, especially regarding the specific functions that actions are supposed to carry out, see, e.g., [158].

- Except for the functions and actions characterizing the interactions of entities with the outside as above, the sixth observation is about the behavior of entities, particularly humans, which is significantly influenced by their internal states. For instance, a person's internals may include mood, personality, and stress. For C_P, we only consider those internals that influence how me changes her/his states. Many studies explore the influence of human internals on their behaviors. For example, student academic performance is affected by their personality traits [84], depression [70, 32, 116], stress [3]. The study in [29] discusses the relationships between the Big Five personality factors, including neuroticism, extraversion, conscientiousness, agreeableness, and openness, with depression and suicidal behavior. Big-thick data often includes such internal information. We say the internal states can be categorized as two types: one is about *internal functions and characteristics*. It is assumed to be stable over time, such as personality traits, procrastination syndrome, and depression. Another is about *internal actions*, which typically exhibits high temporal dynamics, such as mood and tiredness (see, e.g., [14]).

Chapter 5

Reference and Personal Context Model and Unification

According to the definitions of the reference context C_R and the personal context C_P in Chapter 4, this chapter aims to provide the formalization and unification of C_R and C_P. We utilize the $teleology$, $ontology$ and $teleontology$ theories to model C_R and C_P. This work facilitates organizing and integrating various heterogeneous data sources as Knowledge Graphs (KGs) [74, 78] following [52]. Due to the different ways we proposed for context unification, the generated big-thick data can be used to answer diverse enquiries about human behaviors and the world around people, as explored in Chapter 6.

Section 5.1 aims to formalize the C_R teleology model to represent a general C_R. In Section 5.1.1, we first establish an OpenStreetMap (OSM) Ontology (OSM provides extensive, widely-used geographical information) which serves as an illustration to help understand how real-world geographical data provides information about location entities and other spatial entities. In Section 5.1.2, we build the Spatial-temporal Context (C_S) teleontology, based on the OSM Ontology, representing a generic C_S as described in Section 4.1.1. As derived from the C_S teleontology, Section 5.1.3 formalize the C_R teleontology. Finally, in Section 5.1.4, we formalize the C_R teleology based on the C_R teleontology, through a standard process.

Section 5.2 aims to formalize the C_P teleology model to represent a general C_P. In Section 5.2.1 and Section 5.2.2, we demonstrate the construction of the teleontology and teleology of C_P, respectively, as well as discuss the differences with those of C_R.

Section 5.3 illustrates that the context teleologies can be populated with real-world data (e.g., the approach shown in [51, 10], and our previous work in [61, 55]), which generates a set of Entity Graphs (EGs) in the form of Knowledge Graphs (KGs), representing the real-world contexts. Based on this fact, we define various types of *unifications* for unifying the reference context and a set of personal contexts as *observation contexts*, which are the elements composing the life sequences.

5.1 Reference Context Model

As the definition of the reference context (C_R) in Equation (4.3), in order to build the C_R teleontology, it is crucial to first understand how real-world geographical data, like that from OpenStreetMap (OSM), provides information about location entities and other spatial entities. To exemplify this process, we examine a widely recognized and popular geographical data source, namely OSM. Our initial step involves establishing the *OSM ontology*.

5.1.1 OpenStreetMap (OSM) Ontology

The OSM data is a freely accessible and open spatio-temporal geographic database that is continuously updated through public collaboration by volunteers. OSM data is extensively used across various studies, including transportation and mobility [35, 93, 136], geographic information system (GIS) [80, 66], and social sciences [43, 104]. To enhance the readability of OSM data, we use the *OpenStreetMap Data in Layered GIS Format* file[1] (GIS file), it describes a mapping

[1]For further details at the link: http://download.geofabrik.de/osm-data-in-gis-formats-free.pdf.

from OSM data formats to the conventional GIS formats. From the part of information of the places in the GIS file, including geometrical features (with geometry shapes like point, line, or polygon), `Fclass` (indicating the category of places, e.g., `University` or `Hostel`), and etc., we capture the hierarchy structures of OSM places. These hierarchy structures serve as a foundation for dividing the OSM places in the OSM ontology. The detailed formalization process and the underlying principles of generating the OSM Ontology are delineated in Appendix A.

The OSM Ontology records 334 entity types (etypes) and their subsumptions, organized in a tree structure with a maximum layer of six, e.g., `OSM_Place < Point < Point_Of_Interest < Accommodation < Indoor < Hostel`. These etypes include a total of 15 data properties, such as `Name`, and `Population`. Detailed explanations of these etypes and data properties can be found in the *OSM Ontology.owl*[2] file using Protégé[3], which provides a visualized view for the OSM Ontology. For instance, a part of the etypes of the *OSM ontology.owl* is as depicted in Figure 5.1. The full overview of the *OSM Ontology.owl* is introduced in Appendix B.

[2]OSM Ontology.owl file is available on GitHub at https://github.com/xiaoyueli1994/OSM-for-Big-thick-Data/blob/main/OSM%20Ontology.owl.

[3]Protégé is a free, open-source ontology editor and framework for building intelligent systems, more information can be found at https://protege.stanford.edu/.

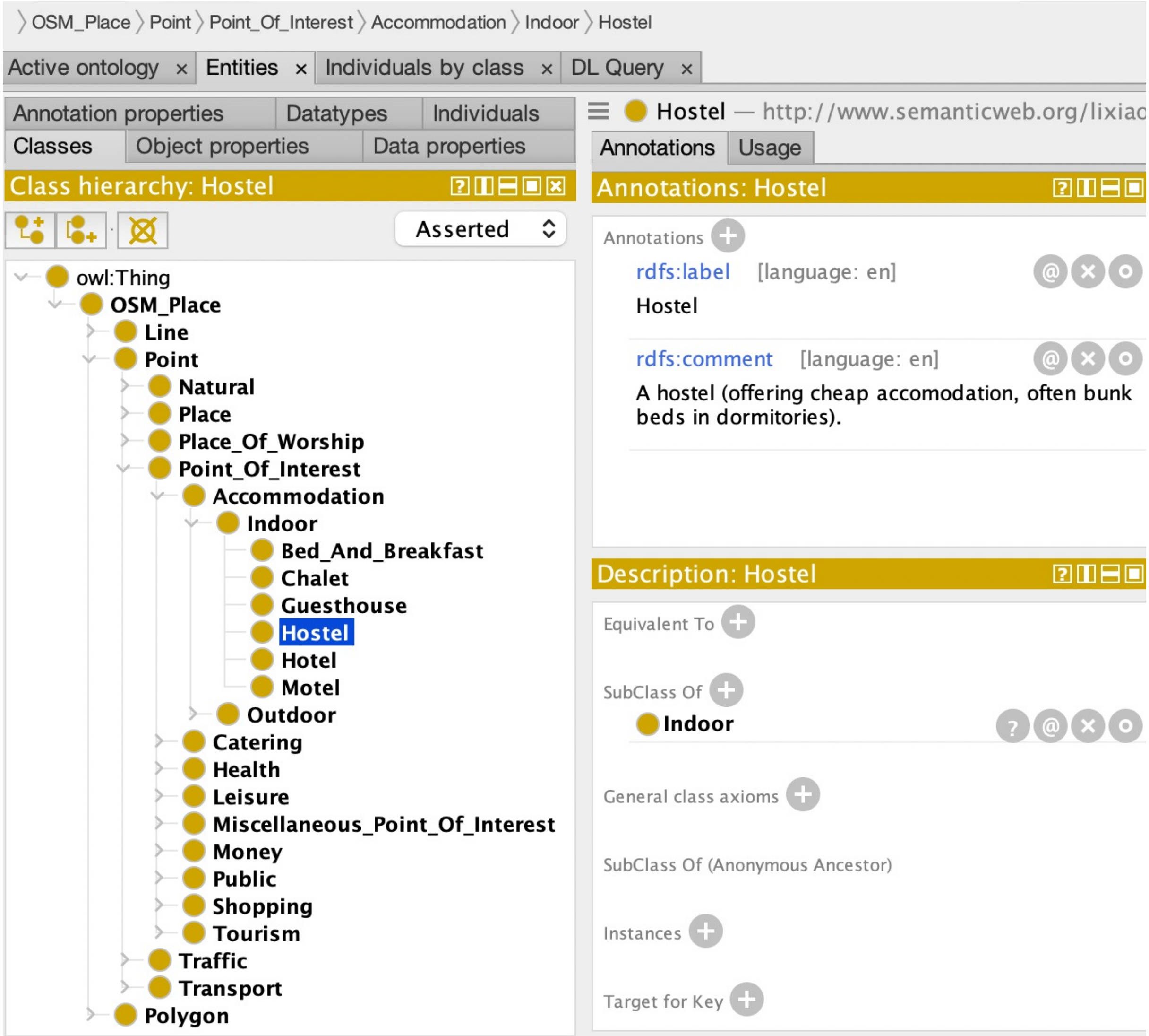

Figure 5.1: OSM ontology in protégé.

To better understand the OSM Ontology, we draw a fragment of the OSM Ontology in Figure 5.2, which provides a more partial view of OSM Ontology compared with Figure 2 (The OpenStreetMap Ontology) in our recently published paper [60]. In Figure 5.2, the upper part of the hierarchy, down to the level of `Point`, `Line` and `Polygon`, describes the three types of geometrical features of `OSM_Place`. In the real OSM datasets, such as from Geofabrik[1],

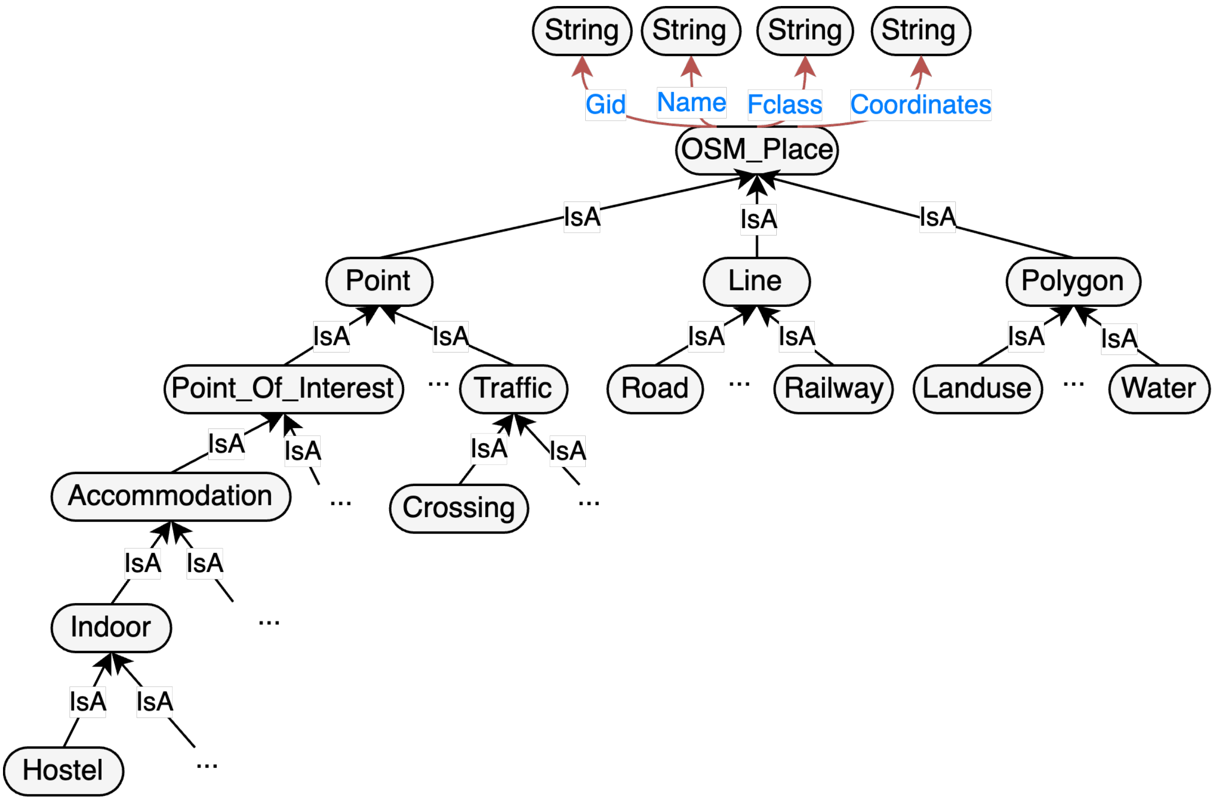

Figure 5.2: OSM ontology fragment.

the places as the point are annotated with the coordinates composed of a pair of latitude and longitude, while the places as the line or polygon are annotated with the coordinates composed of sets of pairs of latitudes and longitudes. Each geometrical feature, in turn, is further refined into sub-hierarchies, allowing the division of the three basic geometrical classes into more specific classes depending on their characteristics. Figure 5.2 shows the four most foundational data properties of `OSM_Place`.

- `Gid`: The place identifier.
- `Name`: The local name of places, such as an Italian name for a place in Italy or a Chinese name for a place in China.
- `Fclass`: The property used to denote the place category.

- `Coordinates`: The spatial coordinates of the place, such as the pair of $\langle latitude, longitude \rangle$.

While a general OSM Ontology has been formalized based on the GIS file in this thesis, it is important to note that this is not the only approach to constructing it. The formalization of the OSM Ontology should also depend on the data and knowledge sources utilized, as well as the specific research purposes. For example, in the real OSM data source downloaded from Geofabrik[1], the `Natural`, `Place`, `Place_Of_Worship`, `Point_Of_Interest`, `Traffic`, `Transport` places are stored both as point and polygon places. Hence, (differently from the OSM Ontology in this work) these types of places should be considered as the sub-hierarchies of both point and polygon places in the OSM Ontology built by the Geofabrik OSM dataset, such as `Place < Point < Natural< ...`, and `Place < Polygon< Natural< ...`. For another example, the OSM ontology in this work considers the geometrical features of places as the standard for categorizing the second layer in the OSM Ontology, namely `Place < Point < ...`, `Place < Line < ...` and `Place < Polygon < ...`. Different from it, the *Trentino Territory Lightweight Ontology* and the corresponding separated *Trentino OSM Data*[4], were developed by us and used in the Knowledge Graph Engineering (KGE) courses[5] at the University of Trento. These sources aim to provide basic knowledge and data sources for students to conduct various types of projects[6] with different purposes. In this case, the `Fclass` property provides the information for the second layer in the OSM Ontology, while geometrical features

[4]Details can be found at the link: https://unitn-knowledge-graph-engineering.github.io/KGE2023-website/material/slides/KGE-OSM_source.pdf.

[5]https://unitn-knowledge-graph-engineering.github.io/KGE2023-website/.

[6]Examples of project proposals can be found at the link: https://unitn-knowledge-graph-engineering.github.io/KGE2023-website/material/slides/KGE_Projects_Proposals.pdf. In addition, a project example of integrating the natural, facilities and transportation places in Trentino can be seen in a sample project documenthttps://github.com/xiaoyueli1994/KGE-example/blob/main/KGE-project-example-Trentino-OSM%20.pdf from us, providing a comprehensive and accurate representation of the region's infrastructure.

become the last consideration. For example, structuring the hierarchy includes `Place < Point_Of_Interest < Health < Hospital < Point_Hospital`.

Based on the OSM Ontology depicted in Figure 5.2, it is intuitive to recognize that the upper part (involving `OSM_Place`, `Point`, `Line` and `Polygon`) encodes the geometrical features of objects, thus, should be represented in C_S teleontology, as shown in Section 5.1.2. The remaining part describes the properties of entities, primarily functions (e.g., accommodation and road), thus should be represented in C_R teleontology as shown in Section 5.1.3. Finally, from the C_R teleontology, we generate the C_R teleology in Section 5.1.4.

5.1.2 Spatial-temporal Teleontology

Figure 5.3 depicts a general C_S teleontology, representing a generic C_S, as described in Equation (4.2). The C_S teleontology, as well as the C_R and C_P teleontologies and teleologies shown in later sections, can be extended or changed for specific purposes. The main components of the C_S teleontology are as follows.

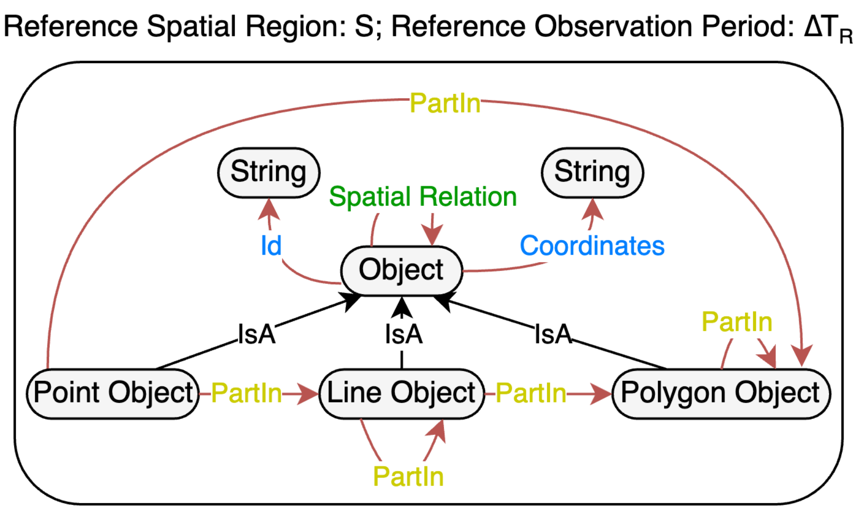

Figure 5.3: Spatial-temporal context teleontology.

- *Spatial-temporal boundary*. The box of the C_S teleontology represents the spatial and temporal boundaries of C_S, defined by the `Reference Spatial Region` S and the `Reference Observation Period` ΔT_R. In the actual implementation, the information about boxes is encoded in the metadata.

- *Object types*. There are four object types, as the nodes of `Object`, `Point Object`, `Line Object`, and `Polygon Object`. The `Object` represents all objects of C_S, corresponding to $\{O(t)\}_S$ in Equation (4.2). The `Point Object`, `Line Object` and `Polygon Object` represent objects with the geometrical shapes of point, line, and polygon.

- *Data properties*. The `Object` node is associated with two foundational data properties, `Id` and `Coordinates`, which linked to `String` nodes representing the *Data Types*.

- *Object properties*. Two types of object properties link these nodes, showing their relations. The `Spatial Relation` represents objects' position relative to each other in space. It can be further specialized into more refined relations, such as *near*, *right*, or *north*. The `PartIn` represents spatial containment relation, similarly to that described in Equation (4.1). It is applied recursively, allowing for the definition of locations at any level of spatial granularity, such as polygons or lines, where, ultimately, locations are populated by point objects. In practice, the values of `Spatial Relation` and `PartIn` can be obtained by calculating the distances between places based on their coordinates.

- *IsA relations*. They are subsumption relations rooted in the superclass of `Object` from the rest of objects, representing the inheritance of objects. The `Point Object`, `Line Object`, and `Polygon Object`, as the subclasses of `Object`, inherit all the data and object properties of

`Object.`

5.1.3 Reference Context Teleontology

Figure 5.4 illustrates the *reference context* C_R *teleontology*, which represents a general C_R. In the C_R teleontology, the nodes are entity types (etypes). The

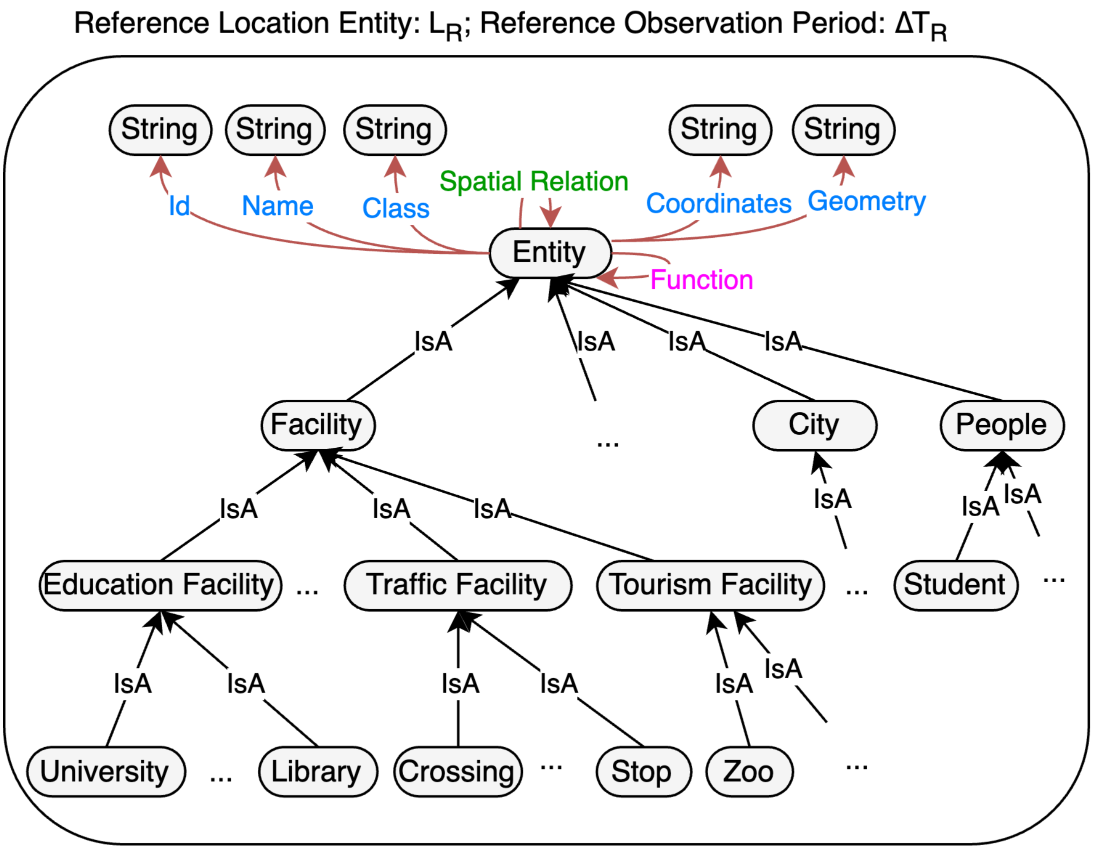

Figure 5.4: Reference context teleontology.

root node is the `Entity`, representing the generic entities in C_R, i.e., reference location entity L_R and a set of entities in C_R $\{e\}_R$ as shown in Equation (4.3). Similarly to the inheritance mechanism in the C_S teleontology, the lower etypes in the hierarchy inherit the data and object properties of the higher etypes. This means that each etype under the root etype `Entity` inherits the properties of the `Entity`. We explain some properties as follows, which are not shown in the OSM Ontology and C_S teleontology.

- The data property `Class` is our term (representing the `Fclass` in the OSM Ontology) denoting the category of a place entity. The `Class` of an etype can take as value either the etype itself or a more specific designation. For instance, we might specify that a particular `Facility` is a `Pub`.

- The data property `Geometry` specifies whether a place entity is a point, a line, or a polygon.

- The `Function` as an object property defines the functions among entities in C_R. It refers to the *function* concept in Section 3.2.1. It is important to note that, unlike the function in C_P, the function in C_R is characterized by objectivity, grounded in the objective reality of entities. For instance, a zoo serves a `RecreationPlaceof` function for a person.

We observe that a point or a line object can be modeled as an entity, whereas a polygon object can be modeled not only as an entity but also as a location (i.e., as the `Reference Location Entity` L_R of the surrounding box in the C_R teleontology), such as `Building`, `Forest`, or `Vineyard`. An entity possessing both a point/line and a polygon geometry can function as either an entity or a location.

5.1.4 Reference Context Teleology

From the C_R teleontology, we can generate C_R teleology (teleology is also referred to as the Entity Type Graph (ETG) [51]). The generation of teleology from teleontology is the process of selecting what is relevant to the specific purpose. Here, the purpose should also consider facilitating context unification between a reference context with the personal contexts (of one or more me).

A teleology can be obtained from the teleontology by the following purpose-driven steps.

i. Select a subset of the etypes of teleontology.

ii. For each etype, select a subset of its object and data properties.

iii. For each (selected) etype, distribute its (selected) properties to its (selected) subclasses.

iv. Delete the `IsA` relations.

Resulting from the C_R teleontology, we represent a fragment of C_R teleology as illustrated in Figure 5.5. In the C_R teleology, we specify the functions among etypes representing that `University` is the `StudyPlaceOf Student` and `Library` is the `ReadPlaceOf Student`. We display a fragment of C_R teleology as an example here to display what a C_P teleology looks like. As we said, depending on the different purposes, different C_R teleologies can be built as needed.

5.2 Personal Context Model

5.2.1 Personal Context Teleontology

Similarly to the construction of the C_R teleontology, we illustrate a fragment of the personal context (C_P) teleontology in Figure 5.6, which represents a general C_P according to Equation (4.7). Compared to the C_R teleontology in Figure 5.4, the C_P teleontology differs in the following key aspects.

- To distinguish from the C_R teleontology, we use the box without rounded corners to depict the C_P teleontology (as well as for the C_P teleology in Section 5.2.2). This allows us to define the spatial-temporal boundary of C_P by defining the `Personal Location Entity` L_R, `Personal`

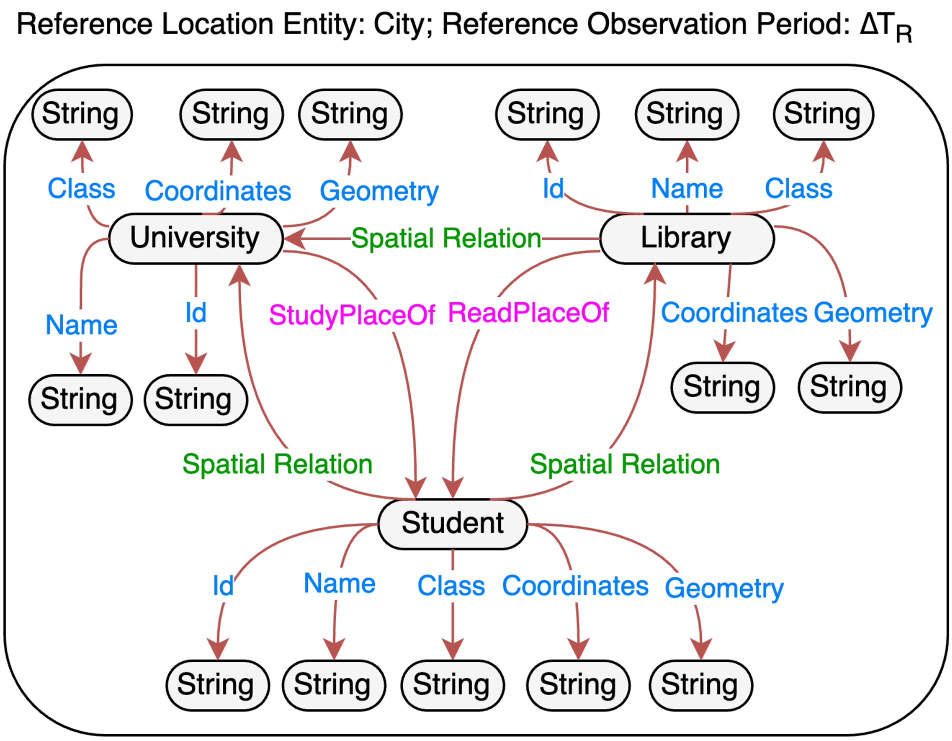

Figure 5.5: Reference context teleology.

`Event Entity` $E(L, t)$ and `Personal Observation Period` ΔT, as shown in Equation (4.5).

- In the C_P teleontology, an additional `me` etype is represented, which is the type of an entity that, from her/his subjective view, builds C_P. Apart from the `me` etype, we list different etypes from that of the C_R teleontology. This does not imply that the etypes of C_P teleontology and that of C_R teleontology are fundamentally different, rather, it is due to space limitations and the extensive number of etypes in the world. Hence, we list a subset of etypes that are important in C_P and C_R in their teleontologies, particularly, to explain the example of context unification in the later section.

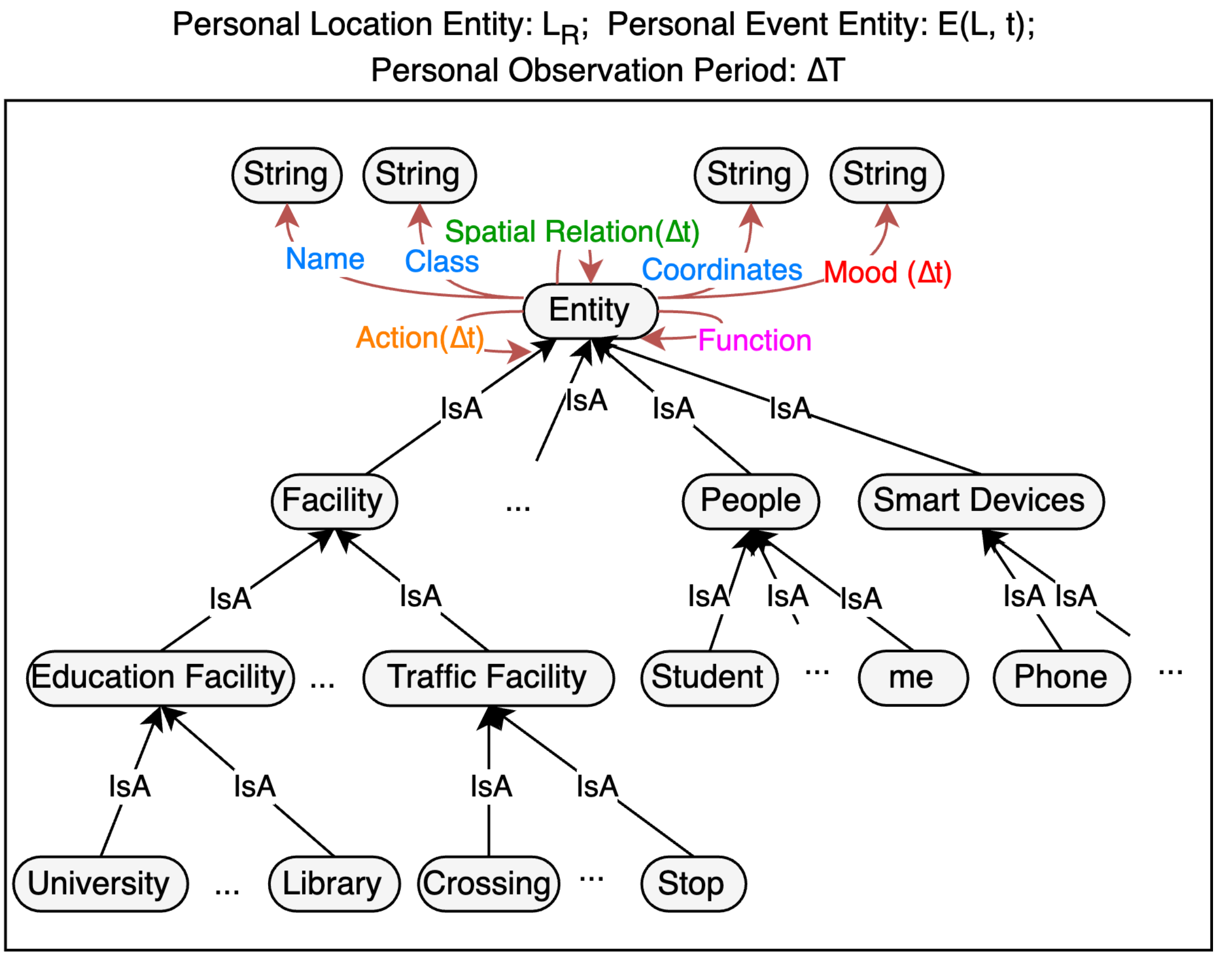

Figure 5.6: Personal context teleonology.

- The `Id` and `Geometry` data properties of the `Entity` are deleted from the C_P teleontology, because such properties of entities are challenging to ascertain from a subjective knowledge of people. Notably, we keep the `Coordinates` data property of `Entity` due to the existence of entities like phones and smartwatches, that can provide their coordinates via sensors, e.g., GPS. Additionally, the `Mood`(Δt) data property is added to the `Entity` in the C_P teleontology, which indicates one type of internal states (as discussed in Section 4.2.2), where Δt is included since mood is time-varying.

- The `Action(`Δt`)` object property is introduced in the C_P teleontology to represent the time-varying actions of entities. For the `Spatial Relation` object property, Δt is added because it is time-varying in C_P.

5.2.2 Personal Context Teleology

The C_P teleology is constructed following the same building process of C_R teleology (namely, from C_R teleonology to C_R teleology as described above in Section 5.1.4). We illustrate a fragment of C_P teleology in Figure 5.7. The C_P

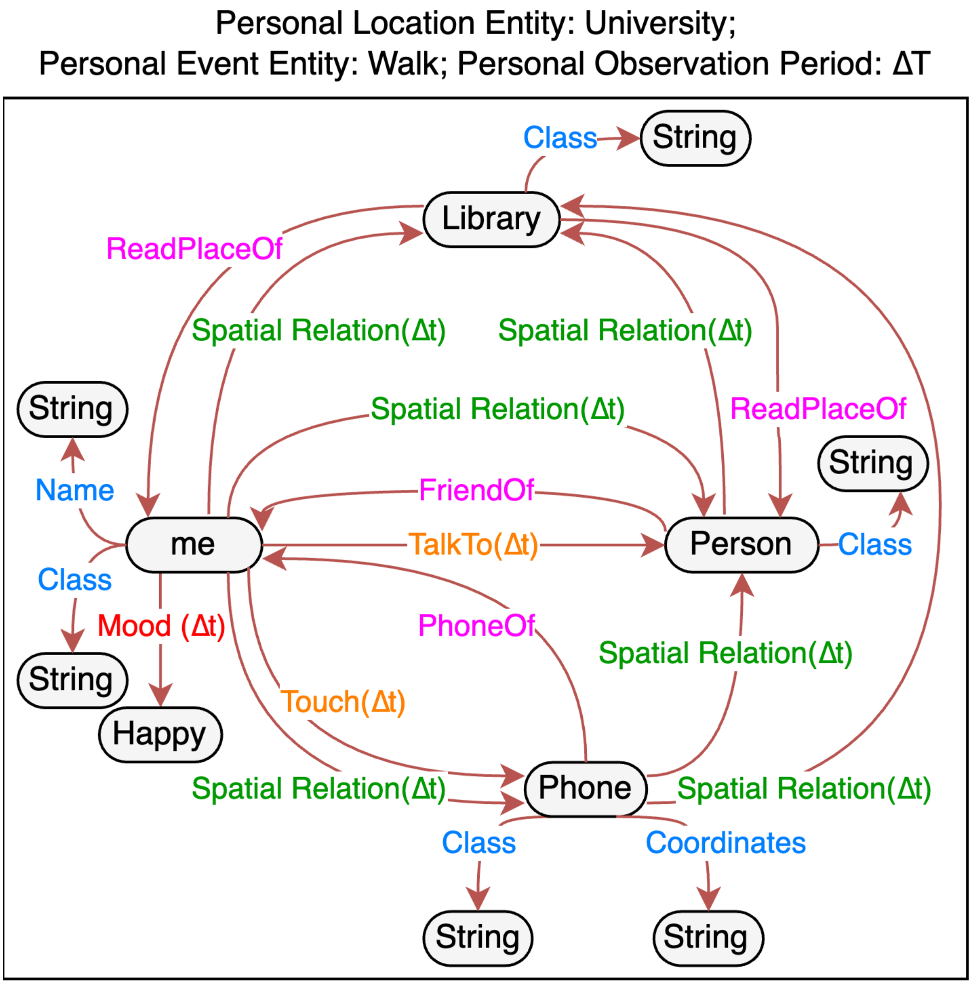

Figure 5.7: Personal context teleology.

teleology has some key differences with the C_R teleology, as follows.

- The `Personal Location Entity` in the C_P teleology can be either different (e.g., `University`) or the same (e.g., `City`) with the `Reference Location Entity` in the C_R teleology. In the former case, a possible comprehensive visualization of C_P depicts that me was first reading in the `Library` within `University`, then taking a class inside the `University`. Hence the personal location entity in C_P is perceived by *me* as the university. In this case, the university is as L_R and the library is as L in Equation (4.7). The latter case assumes that the involved event in C_P does not occur in a sub-location $L \subseteq L_R$, still in $\Delta T \subseteq \Delta T_R$.

- The C_P teleology shares certain elements with the C_R teleology, e.g., the `Library` etype and `Spatial Relation` (for example, a library is near a park). These shared elements should be relevant to the events of me in C_P. This allows positing me within C_R while she/he moves around. It is worth noting that, due to C_P evolving in time, hence what is shared from C_R is time-dependent.

- Although both C_R and C_P involve functions, the function in C_P defines the function of an entity with respect to me envolving in time. While the function in C_R defines the function of two entities based on publicly available objective information, these functions keep static in the reference observation period ΔT.

- Similarly to the differences we observed between the C_R teleontology and the C_P teleontology in Section 5.2.1, the C_P teleology contains actions (e.g., `TalkTo(me, Person)` and `Touch(me, Phone)`), and internal states (e.g., `Mood(me, Happy)`). All of them are in time, with the time duration Δt ($\Delta t \subseteq \Delta T$).

5.3 Context Unification

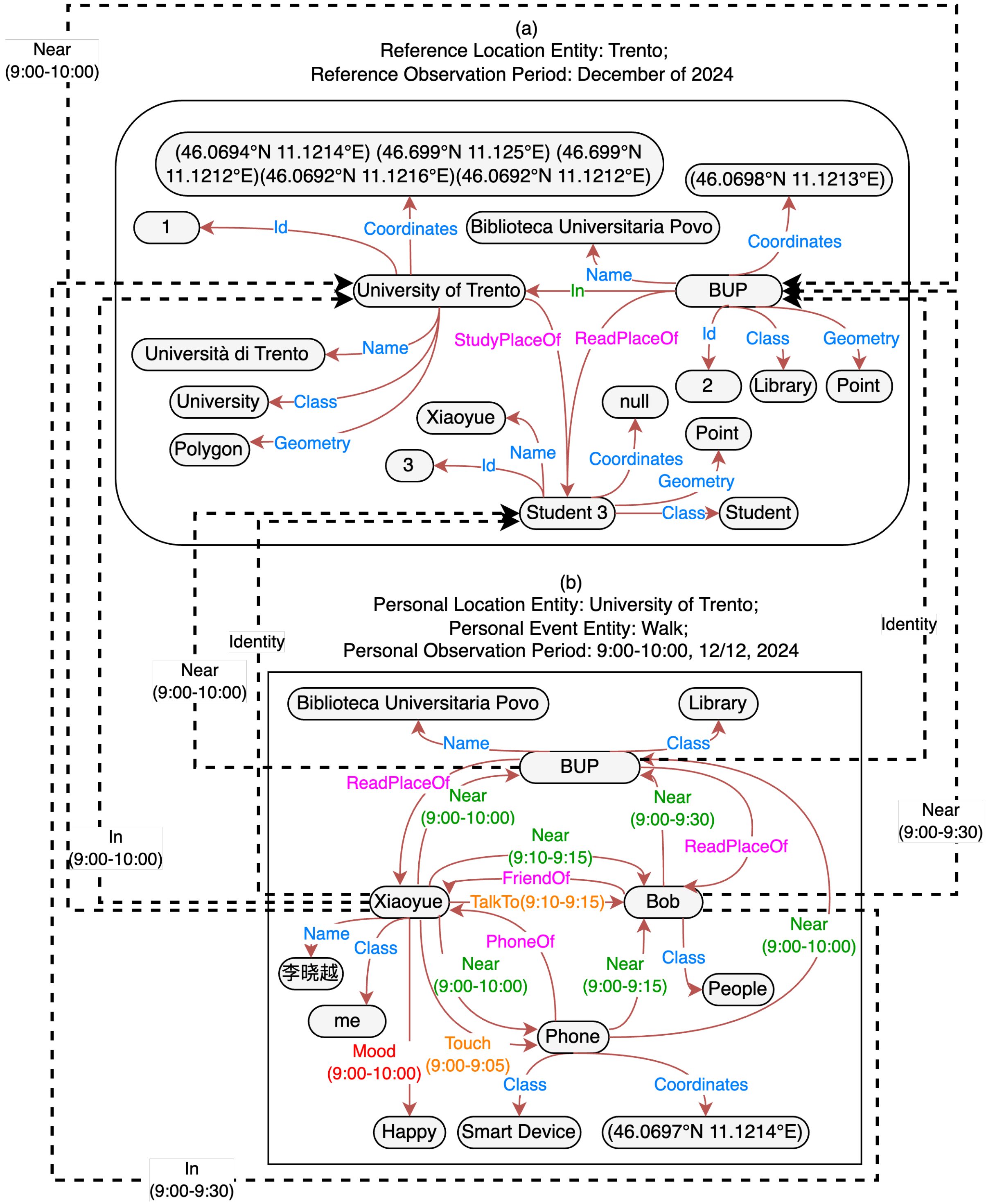

Figure 5.8: An EG example, unifying a reference context with a personal context.

By deleting the inheritance relations of etypes in teleontology, teleology more clearly represents the relations and properties among the etypes. The C_R and C_P teleologies (i.e., Entity Type Graphs) can be populated with real-world data streams, thus generating a set of knowledge graphs, that we call Entity Graph (EGs). The methodology for this population process is extensively discussed in [51, 10], as well as in our previous work [61, 55]. Roughly speaking, EGs are KGs constructed from a teleology by expanding each etype node into the specific entities of that etype. For example, Figure 5.8.(a) and Figure 5.8.(b) show a C_R EG and a C_P EG, respectively, where the `University` etype is instantiated by the `University of Trento` entity and `me` etype is instantiated by the `Xiaoyue` entity, etc. Additionally, all other elements within a teleology, e.g., etyeps' properties, functions, etc., should also be instantiated, such as `Name` of `University of Trento` is `Università di Trento`. In the case study presented in Chapter 7, the C_R teleology is populated using OSM data within Trentino, which generates a C_R EG that represents the spatial entities in Trentino. The C_P teleology is populated with the SmartUnit2 dataset, comprising daily life data from students at the University of Trento in the Trentino region, which generates sequences of C_P EGs for each student (as me).

Given one reference context and a set of personal contexts, the next step is to *unify* them, which generates an *observation context* C. As illustrated in Figure 5.9, an observation context is generated by unifying a reference context (C_R) with a set of personal contexts ($C_{P_1}, \ldots, C_{P_n}$, where $n \geq 0$). This can be

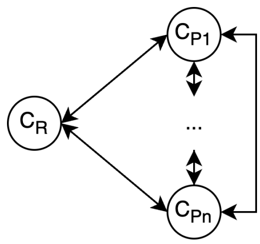

Figure 5.9: An observation context.

formally represented as:

$$C = C_R \uplus \{C_P\} \tag{5.1}$$

where $\uplus$ is the *context unification operator* and $\{C_P\}$ represents the set, one or more, of personal contexts under consideration. The process of context unification involves two primary steps:

i. Implement the unification between a C_R with each C_P of $\{C_P\}$ one by one. This step aims to objectify the entities of C_P by relating with that of C_R. This enables obtaining new properties for entities. For example, as shown in Figure 5.8, identifying the `BUP` (library) in C_R with the `BUP` (library) in C_P, allows us to determine the proximity position *me* entity at the time even in the absence of phone coordinates.

ii. Implement the pairwise unification between any two C_P after they are unified with C_R. Notice that this step means that we perform the unification of two personal contexts always with respect to the reference context.

Figure 5.8 shows the possible unification of one C_R EG in Figure 5.8(a) and one C_P EG example in Figure 5.8(b), which are linking by unification relations represented as dashed arrows. Context unification employs three distinct types of unification methods, as shown in the following.

- *Entity type and property unification (EPU)*. This unification is by identifying or relating entities by their etypes and properties. This is a typical problem of *ontology/schema alignment* (we follow the approach described in [59, 133]). For example, Figure 5.8 identifies the `BUP` in the C_R EG and the C_P EG as the same entity, with the consideration that they belong to the same etype of `Library` and have the same name of `Biblioteca Universitaria Povo`.

- *Spatio-temporal unification (STU)*. This unification is by identifying or relating entities by exploiting their spatial-temporal coordinates. There are two possible types of results. The first is identifying two entities as being the same entity, e.g., two person instances belonging to two different C_P are the same. For this to be true, the coordinates of the two entities must be the modulo approximations at all times. The second is identifying spatial relations at a certain time between entities, For instance, the `Xiaoyue` (is provided coordinates by her phone) is `in University of Trento` at the period of `9:00-10:00, 12/12, 2024` as shown in Figure 5.8.

- *Entity unification (EU)*. This unification is by identifying specific properties of entities, distinct from spatio-temporal properties, but (entity) specific properties [51]. The properties used for this type of unification often present clear and unambiguous information, such as name, identifier, ID card number and phone number.

As from Equation (3.1), the life sequence is a sequence of observation contexts. Given the above, we can see that these observation contexts are instantiated by data streams as a sequence of EGs. The life sequence of a person represents the evolution in time of the life of the person, with each EG in the life sequence representing the state of affairs at a certain time and for a certain time interval. It is worth noticing that, from the life sequence, we can easily acquire a subset of it, as a set of EGs (not necessarily adjacent in time) with a specific purpose (e.g., satisfying a certain property). For example, we can have the EGs where a person was moving around in Trento, or a person studied at the University in the last year. This is a very powerful representational mechanism that can be used, as we discussed in our previous paper [96], to represent various *habits* (not necessarily adjacent) occurring recursively with a certain frequency. Furthermore, when putting the life sequences of multiple persons together, we can acquire the habits of a group of these persons, such as students, or patients.

Chapter 6

Observation

From the above chapters, we define life sequences as composed of unified observation contexts unifying the reference context with personal contexts, i.e., sequences of spatio-temporal Entity Graphs (EGs), as the big-thick data, representing people's daily lives. This chapter introduces the observations for the generated big-thick data.

Section 6.1 discusses two types of possible enquiries on the big-thick data. Section 6.1.1 introduces the *observation enquiry*, which is distinguished by different interested *observation purpose*, enquirying from the C_R, C_P, or both of them. Section 6.1.2 shows the *habit enquiry*, which is distinguished by different types of activities involved in the definition of habit as in Equation (3.2).

Due to no commonly committed evaluation for the big-thick data, Section 6.2 proposes some assessments for validating the quality of big-thick data. Section 6.2.1 introduces the *data size* and *data connectivity* assessments. Section 6.2.2 introduces the *purpose feasibility* metric for enquiries. Given any big-thick data, if it has purpose feasibility for an enquiry, which can further be *enquired* (by *machine learning* or *statistical analysis*) providing answers for the enquiry.

6.1 Enquiry

6.1.1 Observation Enquiry

Context unification process in Section 5.3 allows for flexible selection of the reference context, the personal contexts, and the diverse elements from them to unify, e.g., based on different purposes or unification methods. The result of this process is sequences of spatio-temporal EGs, representing observation contexts of people that describe daily scenarios in their life sequences, which can be used for widely *enquires*. Here, we say *enquiry* (synonymous with inquiry) as a request for information. The target object of an enquiry can be a person or a group of people. By distinguishing any specific possible *observation purpose* (as a function of what one is interested in observing) on observation contexts, we classify observation enquires into four main groups as follows:

- *Reference enquiry (R-enquiry).* R-enquiry is to enquire details of the C_R part of the observation contexts. R-enquiries are posed *only* to the C_R, independently of the dynamics that may occur inside it. Such enquiries typically arise when someone arrives in a new place, for instance, while traveling, to find her/his way or destination. Examples of R-enquiry questions that could be posed to the observation context include: 'Which university is the BUP Library in?', 'Where is the nearest bus stop to the BUP Library?', or 'Which is the nearest restaurant to the BUP Library?'. The answers to these euquries are associated with the entities and their spatial relations in C_R.

- *Personal enquiry (P-enquiry).* P-enquiry is to enquire from a single C_P or streams of C_P of the same me of the observation contexts. It is not relevant to the information from C_R, which is not part of C_P. Such enquiries allow for knowing the details about what people have done in a certain period of time, including also their subjective view of what happened. Examples of P-enquiry questions that could be posed include: ‘Where was *me* in a certain period?’, ‘What did me do there?’, ‘Who was *me* together with?’ or ‘What was me mood?’. The answers to these euquries are associated with location, event, functions of entities, and mood from a subjective view of *me* in C_P, such as ‘Study place’, ‘Friend’, ‘Walking’ and ‘Happy’.

- *Personal-reference enquiry (PR-enquiry).* PR-enquiry is posed to the observation context C, focusing on C_R and is about its state as influenced by the activities of one or more personal contexts. It aims to explore how the state of C_R dynamically evolves as a function of the entities that populate C_R within a certain period of time. Examples of PR-enquiry questions that could be posed include: ‘How many people were in the BUP library last Monday?’ or ‘How many libraries *me* have visited in Trento during the last month?’.

- *Reference-personal enquiry (RP-enquiry).* RP-enquiry is posed to the observation context C, focusing on the C_P of one or more *me*, to explore the environment and its impact on *me* while *me* moving inside C_R. Examples of RP-enquiry questions that could be posed include: ‘What was *me* mood when she/he was in the BUP library last month?’ or ‘Who was *me* together with when in the BUP Library last month?’.

Roughly speaking, R-enquiries and P-enquiries correspond to the typical issues addressed using big data and thick data, respectively. PR-enquiries and RP-enquiries are new types of enquiries, facilitated by big-thick data. These two types of enquiries merge objective facts with human subjective perception. Specifically, PR-enquiries can examine how specific elements of objective knowledge can be enriched by the subjective knowledge provided by multiple people, as a function of their behaviour and subjective perspective. RP-enquiries, on the other hand, enable us to analyze the subjective behaviors of people through the objective facts provided by a third party (as a reference system to linking human behaviors), potentially offering cross-individual inter-subjective views of the world.

6.1.2 Habit Enquiry

Habits are defined as a set of subsequences from life sequences where specific activities (i.e., $\{a_i, \ldots, a_n\}$ in Equation 3.2) are selected and annotated by their frequency of occurrence. These subsequences are the sequences of observation contexts. According to the types of possible activities occurring in observation contexts, as illustrated in Figure 3.2 and the internals of *me* described in Section 4.2.2, we classify the *habit enquiries* as follows:

- *Spatial habit enquiry*. This type of enquiry focuses on the activities related to location entities, which can be derived from both reference and personal contexts within the life sequence of *me*. Spatial habits highlight the places frequently visited by a person. Examples are: 'Where was *me* often last month?' Possible answers include: *me* was often in the BUP library or at home last month.

- *Temporal habit enquiry*. This type of enquiry focuses on the activities related to event entities, typically sourced from personal contexts within the life sequence of *me*. Temporal habits indicate the events regularly attended by a person. Examples are: 'What was *me* often doing last year?' Answers might be: *me* was often studying, watching YouTube, and shopping.

- *Social habit enquiry*. This type of enquiry focuses on the activities about being with person entities, generally obtained from personal contexts within the life sequence of *me*. Social habits illustrate the the persons that a person often is together with. Examples are: 'Who was *me* always with last month?' Answers could be: *me* was always with friends, parents, and classmates.

- *Material habit enquiry*. This type of enquiry focuses on the activities associated with thing entities, such as a phone and a table, typically acquired from personal contexts within the life sequence of *me*. Material habits represent the things that a person often are together with. Examples are: 'What thing was *me* frequently with last week?'. Answers may include: *me* was frequently with her/his phone, computer and glasses.

- *Action habit enquiry*. This type of enquiry focuses on the activities about the actions performed by *me*, like touching her/his chin or picking fingers, usually derived from personal contexts within the life sequence of *me*. Action habits denote the repetitive actions of a person. For instance: 'What actions did *me* regularly perform last year?' A possible answer is that *me* was consistently touching her/his chin.

- *Internal habit enquiry*. This type of enquiry focuses on the activities related to the internal states of *me*, such as mood and stress, just acquired from personal contexts within the life sequence of *me*. Internal habits are always influenced by external contexts, like feeling positive in restaurants or experiencing less stress at home. Examples include: 'What mood did *me* frequently have last week?' Answers might be: *me* was always happy or experienced high stress (e.g., during exam week).

6.2 Evaluation

6.2.1 Data Size and Connectivity

Data size: Data size refers to the measure of space consumed by data, typically quantified in units such as Bytes (B), Kilobytes (KB), Megabytes (MB), Gigabytes (GB), Terabytes (TB), and so on. Data size is important for evaluating data quality. The larger data often provides more information, assuming the data quality is maintained at a high standard. For example, machine learning models trained on larger data generally exhibit better predictive performance. However, larger data also present distinct challenges, such as an increased possibility of uncertain data and inconsistent data, as well as the requirement of greater computational power and storage capacity, which impact data processing and analysis [83, 42].

Data Connectivity: Data connectivity refers to the links between different data sources. It measures the capability of integrating and using data from various sources together for comprehensive analysis and more effective decision-making [113]. In the context unification process, not always all of the C_P EGs can be unified with the C_R EG. In the big-thick data, the data connectivity typically depends on the numbers of the unified C_R and C_P EGs, as an example shown in Table 7.4. It is posited that when assessing data quality, the big-thick data with high data connectivity indicates that this data carries more purpose-relevant information, which is more complete and consistent in representing human daily lives.

6.2.2 Enquiry Purpose Feasibility

Purpose Feasibility: *Purpose feasibility* is a novel boolean metric we propose to indicate whether the data has the ability to support a certain purpose. The purpose feasibility (to an enquiry) demonstrates the ability of the big-thick data to provide an answer to this enquiry. The data that is purpose-feasible for a greater number of enquiries indicates that the data can provide more comprehensive information representing human daily lives.

Observation Enquiry Predictions: When the big-thick data has purpose feasibility for a set of observation enquiries, these enquiries can be *enquired* with predictions by machine learning methods. As a core of Artificial Intelligence (AI) and data science, machine learning focuses on the development of algorithms and statistical models to enable computers to learn experiences and make decisions automatically from data [81, 102]. Machine learning facilitates evidence-based predictions on such as human behaviors [149, 71] and human activity recognition (HAR) [99], by allowing machines to learn about humans from data and thus enabling high-quality and high-value human-machine interactions [15]. We used several supervised machine learning algorithms to make predictions for various observation enquires in Chapter 7. These algorithms learn from selected feature properties (one or more variables of data) to predict the selected target property (a variable of data involved in the observation enquires). One observation is that the big-thick data, as a result of the purpose-driven context unification, theoretically provides more purpose-related data (such as the unification relations between C_R EG and C_P EGs) for the selection of feature properties, thus potentially enhancing the performance of machine learning predictions. This is evident from the prediction experiments of various observation enquiries presented in Table 7.6.

Habit Statistical Information: Supposing the big-thick data has purpose feasibility for a set of habit enquiries. These habit enquiries can be answered by statistical analysis based on big-thick data, allowing us to visualize and learn the habits of humans. To discover the habits from big-thick data, the first step is to set the frequency of defining habits as shown in Section 3.1.2. Various levels of flexibility can be used to define the frequency. For example, as shown in Table 7.8 in the case study, the frequency for spatial and social habits of students is set as 25%, being in a location (e.g., home or 'Bar al Drago') or together with a person (e.g., parents or friends) is viewed as a habit of a student only when the contexts involved the location or the person occupies no less than the 25% of the overall recorded contexts in their her/his life sequence. Unless the temporal, social, material, action, and internal habit enquiries mainly pose enquires towards only personal contexts, the spatial habit enquiry allows for posing enquires towards reference context as well as personal context, because both of the two types of context can provide location entities. This is to say that spatial habit enquiry can be a type of R-enquiry.

Chapter 7

SU2OSM Big-thick Data Case Study

The *purpose* of this case study is to *study the daily lives of students at the University of Trento across the physical places within the Trentino region.* The related *big-thick data project*[3] has been published in the DataScientia community. This chapter implements the methodology proposed in the above chapters, where the SU2OSM big-thick dataset is generated by unifying the reference context (C_R) Entity Graph (EG) constructed from the Trentino OpenStreetMap (OSM) dataset with the streams of personal context (C_P) EGs constructed from the SmartUnitn2 (SU2) dataset. This demonstrates the efficiency of the proposed methodology in the generation of the big-thick data. We observed and evaluated the generated SU2OSM dataset using the *enquiries* and the evaluation methods introduced in Chapter 6, which highlight the generated big-thick data as a type of high-quality context-aware data used to develop the meaningful lifelong human-in-the-loop human-machine interaction.

Section 7.1 aims to introduce the datasets we used. In Section 7.1.1, we introduce the SU2 dataset, which consists of time diaries and sensor data collected from the students at the University of Trento in four weeks. In Section 7.1.2, we introduce the OSM dataset and how we extract the OSM data of the places within the Trentino region, that is the OSM-Trentino dataset.

Section 7.2 focuses on generating the SU2OSM big-thick data. In Section 7.2.1, we generate a C_R EG by populating the C_R teleology with the OSM-Trentino dataset, which represents the geographic entities and location in Trentino. In Section 7.2.2, we generate the sequences of C_P EGs by populating the C_P teleology with the SU2 dataset, which represents the daily lives of students. In Section 7.2.3, following the three distinct types of unification methods mentioned in Section 5.3, we unify the OSM-Trentino C_R EG with the SU2 C_P EGs into a set of observation contexts, referred to as the SU2OSM dataset. Section 7.2.4 shows the life sequences of students based on the generated SU2OSM dataset.

Section 7.3 aims to evaluate and observe the generated SU2OSM dataset. In Section 7.3.1, by summarizing the statistics information of the OSM-Trento, SU2, and SU2OSM datasets, we demonstrate the SU2OSM big-thick data exhibits higher quality, characterized by the smallest data size but greater connectivity. Section 7.3.2 proposes four observation enquiries, as the illustrative examples of the reference enquiry (R-enquiry), personal enquiry (P-enquiry), personal-reference enquiry (PR-enquiry) and reference-personal enquiry (RP-enquiry) discussed in Section 6.1. These enquiries are then addressed using various machine learning algorithms to make predictions. Section 7.3.3 specifies the spatial and social habit enquiries and then shows the statistical results of these habit enquiries from the SU2OSM dataset.

7.1 Datasets

7.1.1 SmartUnitn2

The SmartUnitn2 (SU2) dataset[4] consists of smartphone sensor data, time diaries and surveys collected from one hundred and fifty-eight students at the University of Trento (in the Trentino region) over a period of four weeks, specifically from 05-08 22:02:19 to 06-06 21:51:22 (the year is removed for privacy reasons). The experimental protocols, participant recruitment strategies and data cleaning methods have been comprehensively described in [56]. The SU2 dataset has been suitably anonymized and abided by the General Data Protection Regulation (GDPR). It has been utilized in numerous case studies, see, e.g., [157, 158, 61]. Here, we introduce the key data that we used in this case study as follows.

In the SU2 dataset, a set of *time diaries* were collected by asking users questions using the iLog app [156, 155] running on students' smartphones. These questions include three Harmonized European Time Use Surveys (HETUS) questions[1], including: `Where`: 'Where are you?'; `What`: 'What are you doing?'; `WithWhom`: 'With whom are you?'; Additionally, one question on `Mood` was asked: 'What is your mood?'. All the questions were asked every thirty minutes during the first two weeks and every hour during the last two weeks. The variables of time diaries we used in this case study include:

- *userid*: The participant's identifier, valuing integers from 0 to 157.
- *questiontimestamp*: The time when a set of questions are sent by the server, with the form of 'mm-dd hh:mm:ss', e.g., '5/14 21:02'. This time is viewed as the time when what involved user answers happens.
- *where*: The user answer of the `Where` question representing the location, which can be 17 possible answers, e.g., 'Home', 'Relatives Home', 'Classroom / Study hall'.
- *what*: The user answer of the `What` question representing the event, which can be 23 types of answers, e.g., 'Sleeping', 'Eating' and 'Studying'.
- *withWhom*: The user answer of the `WithWhom` representing the role of the interacted people, which can be 9 possible answers, e.g., 'Friend(s)', 'Classmate(s)'.

[1]More information about the HETUS standard can be found at https://ec.europa.eu/eurostat/web/time-use-surveys.

- *mood*: The user answer of the `Mood` question, representing user mood. Mood values integers from 1 to 5. The higher value means a more positive mood.

The SU2 dataset also includes a wide range of sensor data. Here, we used only the GPS in this case study. GPS data collected users' coordinates using smartphones with an estimated frequency of once a minute. The variables of GPS data we used include:

- *userid*: The participant identifier. The same value of the userid in GPS data as that of the userid in the time diaries denoting the same participant.
- *timestamp*: The time when smartphone GPS/network provides coordinates, which with the form of *mm-dd hh:mm:ss*.
- *latitude*: A geographic latitude coordinate.
- *longitude*: A geographic longitude coordinate.
- *altitude*: A geographic altitude coordinate.

By comparing the *userid*, *questiontimestamp* in time diaries data, with the *userid* and *timestamp* in GPS data, we can link the two types of data together. It can be used to locate users with their coordinates.

7.1.2 OSM-Trentino Dataset

Numerous services are available to provide OSM datasets[2]. These OSM datasets provide geographic information data about multiple types of places over the world or in confined regions, such as countries or specific areas like the north-east of Italy. However, when a purpose is involved within a smaller region, such as cities or towns, it is necessary for researchers to independently extract the

[2]Examples of such services include: Planet OSM at https://planet.openstreetmap.org/, BBBike at https://download.bbbike.org/osm/, Geofabrik[1] and OSM Today at https://osmtoday.com/.

required OSM dataset. Given that the current OSM data sources do not provide the OSM data specific to Trentino. In this case study, to know the places that students travel within Trentino, we used the OSM data provided by Geofabrik[1] to extract a dataset specific to the Trentino region, referred to as the OSM-Trentino dataset. The extraction process is as follows.

Table 7.1: Northeast Italy OSM dataset and OSM-Trentino dataset.

Tables	Place Geometry Types	Place Number in Northeast Italy OSM Dataset	Place Number in OSM-Trentino Dataset
places	point	35,595	6,891
pois	point	208,545	44,033
pofw	point	1,280	193
natural	point	342,129	25,032
traffic	point	157,892	14,559
transport	point	34,398	6,025
railways	line	17,261	1,605
waterways	line	102,093	27,078
roads	line	1,331,470	177,161
buildings_a	polygon	4,592,840	290,957
landuse_a	polygon	551,670	70,628
water_a	polygon	31,142	4,360
places_a	polygon	614	152
pois_a	polygon	96,361	12,463
pofw_a	polygon	12,016	1,444
natural_a	polygon	12,27	402
traffic_a	polygon	60,662	6,693
transport_a	polygon	951	149
Total Number	-	7,578,146	689,825

i. Geofabrik offers various data formats, such as SHP and PBF. For our purpose, we used the OSM data in the SHP format. We downloaded the *nord-est-latest-free.shp.zip*[6] on the 9th of June of 2023, which contains OSM data for the places in the northeast of Italy, encompassing Trentino. Notably, we specify the download date as the OSM dataset is updated daily.

ii. The *nord-est-latest-free.shp.zip* file cannot be used directly, hence we imported it into a PostgreSQL database, and subsequently exported 18 tables in the TXT format. The names of these tables are listed in the first column of Table 7.1, where *pois* and *pofw* are the abbreviations for *points of interest* and *places of worship*, respectively. If the table name includes *_a*, it indicates the places in this table are recorded with the geometry of the polygon shape. Each table contains information for a specific type of place. For instance, the *waterways* table includes places such as *canal* (an artificial waterway) and *stream* (a smaller river or stream). We selected the TXT format because the coordinates of OSM places are too extensive for some formats like CSV to store all characters. During the exporting process, we used the *st_astext* function to convert the *geom* data property (a geometry input parameter, such as '010100000077A22424D2162840BCDDDCA9C A4D4740') into a well-understood text representation, denoted as *coordinates* (such as 'POINT(12.044572 46.6077473)').

iii. To extract the OSM data specific to the Trentino region from the north-east Italy OSM dataset, we constrained the coordinates of the involved places to lie within the maximum and minimal latitudes and longitudes of the Trentino region. We defined three coordinate boundaries for Trentino, as shown in Table 7.2. Each boundary forms a box on a map limiting the maximum and minimum latitudes and longitudes of the Trentino region. (Of course, users may define as many boxes as needed, with additional boxes allowing for a more precise regional delineation.) Specifically, for each place with the point geometry, if its longitude or latitude is outside the boundaries, we excluded the point place data from the OSM-Trentino dataset. For each place with the line or polygon geometry, including multiple pairs of longitude and latitude coordinates (e.g., 'MULTI-

LINESTRING((10.5756812 46.6741329,10.5736334 46.6718544))'), we examined each pair of coordinates. If any longitude or latitude of a coordinate pair fell outside the boundaries, we removed that pair. If a line or polygon place had no coordinate left after this process, we excluded the line or polygon place data from the OSM-Trentino dataset.

Table 7.2: Coordinate boundaries of Trentino.

Trentino Boundaries	Minimum Latitudes	Maximum Latitudes	Minimum Longitudes	Maximum Longitudes
Boundary 1	46.671853	46.971215	10.381315	11.206114
Boundary 2	45.783669	47.014468	11.206114	11.962911
Boundary 3	46.531540	47.094276	11.962911	12.478587

After the extraction process, we acquired the OSM-Trentino dataset. Table 7.1 lists the numbers of places in the northeast Italy OSM dataset and that of the OSM-Trentino dataset. The related codes[3] have been uploaded to GitHub, and the OSM-Trentino dataset[5] is available. The OSM-Trentino dataset represents real places, which are characterized by a set of variables. The complete variables of OSM-Trentino can be found in Appendix C. We only introduce some key variables in this section. All the places in the OSM-Trentino dataset have the following variables.

- *gid*: Place Identifier, being unique in each table.
- *name*: Place local name, e.g., 'Post St.Martin'.
- *fclass*: Place category name, e.g., 'library' and 'post_office'.
- *coordinates*: Place coordinates including *latitude* and *longitude*, such as 'POINT(11.350667 46.498294)', 'MULTILINESTRING((10.5756812 46.6741329, 0.5736334 46.6718544))' and 'MULTIPOLYGON(((11.225754

[3]See, https://github.com/xiaoyueli1994/OSM-for-Big-thick-Data/blob/main/0.shpToTxt.rtf and https://github.com/xiaoyueli1994/OSM-for-Big-thick-Data/blob/main/1.extract_local_osm_place.py.

46.5216903,11.2258764 46.5222162,11.227201 46.5220703,11.2270786 46.5215443,11.225754 46.5216903)))'.

In addition, *building* places have a special variable named *type*, which shows the type of a building if specified in OSM (otherwise empty). For example, for a building named 'Chiesa della Natività di Maria', its type is 'church'.

7.2 SU2OSM Dataset Generation

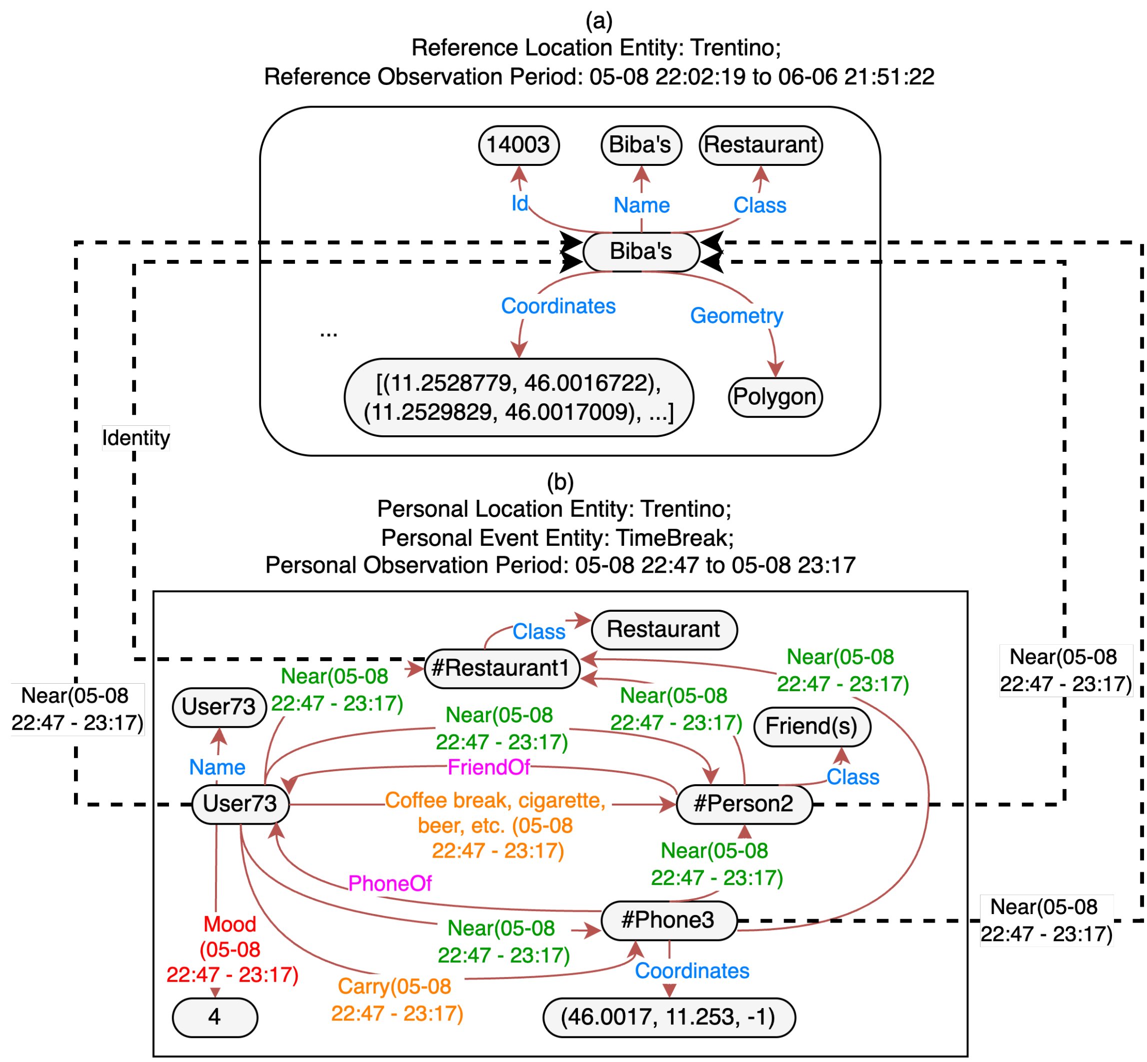

Figure 7.1: An observation context EG from SU2OSM dataset.

7.2.1 Reference Context EG from OSM-Trentino Dataset

A C_R EG is constructed using the *OSM-Trentino* dataset based on the C_R teleology as shown in Figure 5.5 in Section 5.1.4. A fragment of the C_R EG is shown in Figure 7.1(a), where the `Reference Location Entity` is `Trentino`, and the `Reference Observation Period` (ΔT_R) corresponds to the period of the SU2 data collection experiment, specifically `05-08 22:02:19 to 06-06 21:51:22`.

The C_R EG contains a set of place entities, which are planned from the OSM places with the point and polygon geometries, a total of 483,981 OSM places, recorded in the OSM-Trentino dataset. OSM line places are not included in the OSM-Trentino C_R EG for this case study, as they are not deemed relevant to the current purpose. Because positing a person in a point place (e.g., a restaurant or supermarket) or a polygon place (e.g., a forest or hamlet) provides more precise contextual information, whereas positing a person in a line place (e.g., a street) usually results in ambiguous contextual information. We notice that representing OSM line place entities in a C_R EG leads to more complex spatial relations, such as a line place (e.g., a street, a river) $intersecting$ a polygon place (e.g., a park). The representation of such dedicated spatial relationships is part of our future work in refining the reference context.

In the OSM-Trentino C_R EG, each entity is decorated with data properties derived from the OSM place variables. For instance, the `Biba's` entity in Figure 7.1(a) has the `Class` of `Restaurant` and the `Geometry` of `Polygon`, etc. The `Coordinates` of OSM places allow computing the spatial relations between them in C_R, such as `Near`. However, due to the extensive number of place entities, not all of them are unified with personal context (C_P) EGs, it is a selection to only calculate the spatial relations of the place entities unified with C_P EGs.

7.2.2 Personal Context EGs from SmartUnitn2

The C_P EGs are constructed using the SU2 dataset, wherein each participant is regarded as me in her/his C_P EGs. The SU2 data is formalized into sequences of C_P EGs based on the C_P teleology as depicted in 5.7 in Section 5.2.2 in an obvious way, but with a few adaptations that take into account the specificity of the SU2 dataset.

During the overall four-week data collection period, the 158 participants were asked a total of 168,095 question batteries. Each question battery comprised the questions listed in Section 7.1.1 (and more). Each participant received between 1,063 and 1,067 question batteries, where each battery formalized one C_P EG of this participant (as me). Consequently, this results in a maximum of 168,095 C_P EGs for the 158 participants. However, given that participants did not always respond to all the question batteries, finally, a total of 104,414 C_P EGs were formalized for the 158 participants. For each me, the length of life sequence includes the EGs with the length in the range of $371 - 875$.

Figure 7.1(b) presents an example of a C_P EG. We consider the following key points to populate C_P.

- In a C_P EG, the `Personal Location Entity` is `Trentino`, and each C_P EG is associated with a single `Personal Event Entity` $E(L)$ and the `Personal Observation Period` ΔT. These are associated with the answer of the `What` question and the *questiontimestamp* of the question battery, respectively. We assume that ΔT is the interval of time between two question batteries, specifically, half an hour during the first two weeks and one hour during the second two weeks, centered on the time of the question battery. Figure 7.1(b) illustrates a C_P EG example, where a `TimeBreak` event involving the participant `User73`.

- Because HETUS questions are designed for the general intended use (mainly within the Social Sciences), these questions are designed to allow for collecting generic answers, which represent the Entity Types (etypes) in C_P. From these generic answers, we need to allocate the entities for C_P EGs. The answers to the `What` question can be directly encoded as actions, such as `Coffee break, cigarette, beer, etc.(05-08 22:47-23:17)` as shown in Figure 7.1(b). However, the encoding of the answers to the `WithWhom` and `Where` questions is a little more elaborated. For example, when a user `User73` answered `Friend(s)` and `Restaurant` to the `WithWhom` and `Where` questions, we can know that `User73` as me is with a person entity and located in a restaurant entity, whose detailed information (e.g., names) are unknown. As shown in the C_P EG example in Figure 7.1(b), we translate the `WithWhom` answer as a function of `FriendOf(#Person2, User73)`, where the anonymous person entity `#Person2` has the `Class` of `Friend(s)`. We have a set of spatial relations in time Δt, namely `Near(User73,#Person2,` Δt`)`, `Near(User73, #Restaurant1,` Δt`)` as well as `Near(#Person2, #Restaurant1,` Δt`)`, where an anonymous place entity `#Restaurant1` has the `Class` of `Restaurant`. The time parameter Δt encodes when the time-variant property holds.

- The questions were sent by participants' phones. Hence, we assume that, in any C_P EG, each participant was associated with a phone entity. As shown in the C_P EG example in Figure 7.1(b), we have a function of `PhoneOf(#Phone3, User73)`, where the anonymous phone entity is denoted as `#Phone3`; an action of `Carry(User73, #Phone3,` Δt`)`; and the spatial relations, namely `Near(User73, #Phone3,` Δt`)`, `Near(#Phone3, #Restaurant1,` Δt`)` as well as `Near(#Phone3, #Person2,` Δt`)`. It is worth noticing that the phones always have coordinates, which can provide the spatio-temporal positions of the entities near phones, such as me. We computed the coordinates of a phone in C_P

by considering the coordinates from the SU2 GPS data within a time window of 10 minutes around the question time. We applied the DBSCAN clustering algorithm [129] to identify the largest cluster of associated coordinates. The mean of the coordinates in the largest cluster was taken as the phone coordinates. In total, in 26,827 C_P EGs, the phone entities were associated with coordinates.

7.2.3 EG Unification

Given that the OSM-Trentino C_R EG focuses on geographic entities and locations, the sequences of SU2 C_P EGs represent the daily lives of students. To understand the physical places where students live, we unify the OSM-Trentino EG with the SU2 EGs of 158 me into the observation EGs, which we refer to as SU2OSM dataset. Figure 7.1 illustrates an observation context EG obtained from the unification of the C_R and a C_p EG, where the dashed arrows indicate the results of the unification process. The unification process involves implementing the three distinct types of unifications mentioned in Section 5.3, as follows.

- *Entity type and property unification* (EPU) is performed manually. This task is relatively straightforward, given the limited scope of the etypes and properties of the C_R and C_P teleologies. Considering C_R and C_P teleontologies, this task is largely within the reach of the algorithm described in [133]. For example, the place etype in the C_R teleology should correspond to the place etypes in the C_P teleology.
- *Spatio-temporal unification* (STU) is conducted to establish the spatial relations between the entities inside each timed C_P EG, and place entities in C_R EG. The focus is on the spatial relation `Near`. The phone entities in C_P EGs have coordinates. Thus, this task is initially achieved, using their coordinates, by calculating the distances between the phone entities and

the nearest OSM-Trentino place entities. We assume that two such entities are near each other if their distance is less than 50 meters. For instance, in Figure 7.1, `Near(#Phone3, Biba's)` is the result of computing the distance to `Biba's`, the closest OSM-Trentino place to `#Phone3`, which is approximately 8.2 meters. Given that in any C_P EG, the entities near the phone entity also be near the place entities that are near the phone entity, which allows us to derive which *me*, place and person entities (near the phone entity) in a C_P EG are close to the place entities in the C_R EG. For instance, in Figure 7.1, we have `Near(User73, Biba's)` and `Near(#Person2, Biba's)`.

- Entity Unification (EU) is to establish which specific C_R place entity is the one where me is (i.e., place entity in C_P EGs). The idea is to select among the place entities in C_R that are close to me, the one with the same `Class` value with that of the place where me is. This allows us to *identify* a generic C_P place entity, e.g., `#Restaurant1` in Figure 7.1(b), as corresponding to a specific C_R entity, e.g., the `Biba's` in Figure 7.1(a). In case of multiple correct place entities near the phone, we select the closest one.

Following the EG unification process described above, a total of 147 out of the $483,981$ OSM-Trentino place entities were unified with the C_P EGs, many of which matched more than once during the four weeks, thus highlighting the students' habits about visiting places. A total of 1955 C_P EGs (that is, 1.87% of the total number of timed C_P EGs) were unified with the C_R EG. We computed $1955 \times 4 = 7820$ relations linking the C_R EG to the C_P EGs, where $1955 \times 3 = 5865$ are `Near` spatial relations, the rest 1955 relations can be either `Near` or `Identity` relations depending on the identified OSM-Trentino place entities are near or identity with the place entities in C_P EGs. This unification would enable the unification of entities across C_P EGs, allowing us, for instance, to establish when two different me entities were in the same location (e.g., Biba's restaurant) at the same time (e.g., in the morning or on Monday), or to measure how much time a me spent in the same place during certain periods.

7.2.4 Life Sequences of SU2OSM Dataset

Table 7.3: Life sequences of SU2OSM dataset.

Length of Life Sequence	Number of Users
[371-434]	4
(434-497]	12
(497-560]	17
(560-623]	29
(623-686]	28
(686-749]	27
(749-812]	21
(812-875]	20

Table 7.3 provides a comprehensive view of the distribution of users based on the length of their life sequences in the SU2OSM dataset. The length of a life sequence represents the number of EGs in a user's life sequence. The number of users indicates the count of users whose life sequences fall within the specified range of the length of life sequences. The SU2OSM dataset consists of 158 users, distributed across 8 groups according to the length of their life sequences. The group with the shortest life sequences (length from 371 to 434) has the fewest users (4 users). Most users fall into a common range of life sequence lengths, i.e., from 560 to 749. This reveals the number of answers that most users are willing to provide during the experiment period. Additionally, we observe that user 113 and user 122 have the minimum and maximum lengths of their life sequences, which include 371 and 875 EGs, respectively.

7.3 Evaluation

7.3.1 Dataset Size and Connectivity

Table 7.4: Statistics of OSM-Trentino, SU2 and SU2OSM datasets.

Dataset	Data size	CR EG	CP EG Streams	Unified EG Streams
OSM-Trentino	156,5 MB	1	0	0
SU2	441,9MB	0	104,414	0
SU2OSM	19,2 MB	1	104,414	1955

Table 7.4 presents the statistics of the OSM-Trento, SU2 and SU2OSM datasets, including the data size and the number of formalized EGs. The SU2OSM dataset has a size of approximately 19,2MB, which constitutes around 3.21% of the cumulative total size of the original OSM-Trentino and SU2 datasets, collectively amounting to 598,4MB. Moreover, the SU2OSM dataset carries more purpose-relevant information than the OSM-Trentino plus SU2 datasets together. This results from the fact that only a very small minority of the entities in the OSM-Trentino dataset and the coordinate records in the SU2 GPS data are relevant to the current purpose. Additionally, the 1955 unified EGs generated from the SU2OSM dataset demonstrate the SU2OSM dataset's higher data connectivity which carries more purpose-relevant information. This statistics information of datasets provides evidence of the scalability and effectiveness of the proposed approach in the generation of big-thick data.

7.3.2 Predictions for Observation Enquiries

We specify four distinct observation enquiries, namely E1, E2, E3, and E4, as shown in Table 7.5. These are illustrative examples of the reference enquiry (R-enquiry), personal enquiry (P-enquiry), personal-reference enquiry (PR-enquiry), and reference-personal enquiry (RP-enquiry) introduced in Section 6.1.1. The *purpose feasibility* of a dataset (see column 4 in Table 7.5) is a boolean metric

Table 7.5: Observation enquiry purpose feasibility.

Observation Enquiry	Enquiry Type	Dataset	Purpose Feasibility
E1. Does a place be classified as a residence?	R-enquiry	OSM-Trentino	✓
		SU2	×
		SU2OSM	✓
E2. Is a student at a living place?	P-enquiry	OSM-Trentino	×
		SU2	✓
		SU2OSM	✓
E3. Is a student in a bank?	PR-enquiry	OSM-Trentino	×
		SU2	×
		SU2OSM	✓
E4. Does a student have a positive mood in different categories of places?	RP-enquiry	OSM-Trentino	×
		SU2	×
		SU2OSM	✓

Note. ✓ and × indicate whether E1, E2, E3, and E4 succeed or fail, respectively, in satisfying the purpose feasibility metric of the selected dataset.

indicating whether a dataset has the ability to support a certain enquiry. This metric depends on whether the dataset can provide the related properties (as characterized by the target property and feature properties in Table 7.6) necessary to answer the enquiry.

As illustrated in Table 7.5, the OSM-Trentino dataset can only answer E1 (a R-enquiry) because it exclusively populates a C_R EG, thus providing only the C_R properties required for E1. Similarly, the SU2 dataset can only answer the E2 (a P-enquiry) because it exclusively populates streams of C_P EGs, thus providing only the C_P properties needed for E2. However, the merged SU2OSM dataset can answer all the proposed observation enquires as it unifies a C_R EG with C_P EGs, thereby offering both C_R and C_P properties for all observation enquiries. This result demonstrates the capability of the generated big-thick data to address a broader range of enquiries about human daily lives.

As shown in Table 7.6, we conducted binary classification (details, see notes in the table) prediction experiments for the proposed observation enquiries based on their purpose-feasible datasets. Each experiment identified a *target property*

Table 7.6: Observation enquiry prediction experiments.

Observation Enquiry	**Dataset**	**Prediction Experiment**						
		Target Property	**Feature Properties**	**Best Performance Algorithm**	**Accuracy**	**Recall**	**F1 Score**	**AUC**
E1	OSM-Trentino	type	name, class	Random Forest	78.23%	0.529	0.491	0.529
	SU2OSM		day_of_week, time_of_day, name, class	Random Forest	82.51%	0.714	0.728	0.825
E2	SU2	where	what, mood, withWhom	Decision Tree	87.43%	0.850	0.855	0.930
	SU2OSM		name, class, what, mood withWhom,	Random Forest	94.42%	0.846	0.878	0.949
E3	SU2OSM	class	what, where, mood, withWhom	Decision Tree	90.91%	0.731	0.736	0.923
E4	SU2OSM	mood	fclass, type	Logistic Regression	82.46%	0.5	0.452	0.656

Note 1 (E1): the *target property* indicates if the property *type* has values 'apartments' or 'house' or 'residential'. Based on *questiontimestamp2* property, the *day_of_week* feature is labeled with the weekdays (from 'Monday' to 'Sunday'), and the *time_of_day* feature is labeled with 'morning', 'afternoon', 'evening' or 'night'.
Note 2 (E2): the *target property* indicates if the property *where* has values of 'Home', 'Relatives Home' or 'House (friends, others)'.
Note 3 (E3): the *target property* indicates if the property *class* has the value of 'bank'.
Note 4 (E4). the *target property* indicates if the property *mood* values '4' or '5'.

and a set of *feature properties* from the feasible datasets. The prediction experiments tested the Logistic Regression, Decision Tree, and Random Forest algorithms. The prediction experiments follow the *5-fold cross-validation* approach [86], and using some standard evaluation metrics including Accuracy, Recall, F1 Score, Area Under the Curve(AUC)[4]. Table 7.6 shows the results of

[4]Accuracy refers to the ability of the prediction model to accurately predict results, as the percentage of correct predictions out of the total predictions. Recall refers to the ability of the prediction model to find the positive cases, as the proportion of actual positive cases. F1 Score combines precision and recall, especially useful when

prediction experiments for the observation enquiries with the best-performance algorithm. From the results, the E1 can be answered both using OSM-Trentino and SU2OSM datasets, but the metrics from the SU2OSM dataset are better than those of OSM-Trentino dataset, despite the SU2OSM dataset being much smaller (as shown in Table 7.4). The E2 can be answered both using the SU2 and SU2OSM datasets, but, again, the metrics from the SU2OSM dataset are, in general, better than those of the SU2 dataset. The E3 and E4 can be answered only by the SU2OSM dataset. Hence, we can see the merged SU2OSM dataset exhibits better prediction performance in general than the SU2 and OSM-Trentino datasets on the observation enquires on students' daily lives.

7.3.3 Statistical Information for Habit Enquiries

As illustrated in Table 7.7, we exemplify the spatial and social habit enquiries, where E5 and E6 are two spatial habit enquiries, and E7 is a social habit enquiry. Considering that any habit must involve me entities in C_P EGs, hence,

Table 7.7: Habit enquiry purpose feasibility.

Habit Enquiry	Enquiry Type	Dataset	Purpose Feasibility
E5. Where were you often during the SU2 period?	Spatial habit enquiry	OSM-Trentino	×
		SU2	✓
		SU2OSM	✓
E6. What physical places did you often visit during the SU2 period?	Spatial habit enquiry	OSM-Trentino	×
		SU2	×
		SU2OSM	✓
E7. Who were you often with during the SU2 period?	Socail habit enquiry	OSM-Trentino	×
		SU2	✓
		SU2OSM	✓

Table 7.7 indicates that the OSM-Trentino dataset does not have purpose fea-

there is an uneven class distribution. Area Under the Curve (AUC) refers to the ability of the prediction model to distinguish between classes. The higher values of the metrics mean better predictions are performed.

sibility for all the proposed enquiries because it cannot provide the C_P EGs involving me. Although both E5 and E6 are spatial habit enquiries, their location entities involved the activities defining habits, such as being at *Home* or being in *Bar al Drago*, are from C_P EGs and C_R EG, respectively. Hence, E5 is purpose-feasible based on the SU2 and SU2OSM datasets, which provide location entities from personal contexts. While E6 is purpose-feasible only on the SU2OSM dataset, which provides the location entities from C_R EG and me entity from C_P EGs. E7 has similar conditions to E5 but requires person entities, such as *Parent*. This table illustrates that the integrated SU2OSM dataset facilitates a more comprehensive understanding of human habits.

Table 7.8 presents the discovered habits of students corresponding to E5, E6, and E7. The feasible datasets for these enquiries, previously presented in Table 7.7, are not reiterated here. This is due to the fact that, unlike in machine learning processes where different selections of feature properties lead to varied prediction results, the calculation of habits involves the same variables. Notably, these variables remain unchanged between the SU2 and SU2OSM datasets. Hence, it is not needed to list the involved datasets in Table 7.8.

For E5, Table 7.8 indicates that the most popular habit among students is being at home, with 126 students reporting this. Considering the universal need for sleep at home, secondary habits such as being at the relative's home (42 students) and in the classroom/laboratory (6 students) are also significant. Regarding E6 about the specific places that students were, Table 7.8 shows each identified location is associated with only one student, which indicates a wide range of personalized, individual-specific spatial habits. For E7, being alone is the most common habit, with 142 users indicating this. This could reflect a preference for solitude or a lack of social interaction opportunities among students. The next most common social habits involve close relationships, such as being with friends (33 users), relatives (14 users), and roommates (14 users). Understanding spatial and social habits offers valuable insights into students' daily routines and behaviors, thus providing information for designing targeted support services or interventions to help them.

Table 7.8: Spatial and social habits of students ($f = 25\%$).

Habit Enquiry	Discovered Habits	Number of Users
E5	Being in 'Classroom/Laboratory'	6
	Being in 'Classroom/Study hall'	4
	Being in 'Home'	126
	Being in 'House (friends, others)'	4
	Being in 'Other Library'	1
	Being in 'Other place'	2
	Being in 'Relatives Home'	42
	Being in 'UNITN Library'	1
E6	Being in 'Caldonazzo'	1
	Being in 'Merler'	1
	Being in 'Cassa Rurale Alta Valsugana'	1
	Being in 'Biba's'	1
	Being in 'Ristorante Al Pescatore'	1
	Being in 'Lungolago'	1
	Being in 'Centro di raccolta zonale'	1
	Being in 'Macelleria Eccher'	1
	Being in 'Giornali e Tabacchi Defrancesco Angela'	1
	Being in 'Bar al Drago'	1
	Being in 'Bar Italia'	1
	Being in 'Palazzo Norcen Dal Covolo'	1
	Being in 'Biblioteca Comunale O. Gasperini'	1
	Being in 'Chiesa dei Santi Pietro e Paolo'	1
E7	Being 'Alone'	142
	Being with 'Classmate(s)'	10
	Being with 'Friend(s)'	33
	Being with 'Other'	1
	Being with 'Partner'	13
	Being with 'Relative(s)'	14
	Being with 'Roommate(s)'	14

Note ($f = 25\%$). The f is the frequency threshold for defining the habits of students. It implies that being in a particular location entity or together with a specific person entity is considered a habit only when such contexts constitute no less than 25% of the total recorded contexts within a person's life sequence.

Chapter 8

Conclusion

8.1 Summary

To generate big-thick data, we represent and model the *life sequences* of people, which are the sequences of their *observation contexts*. Each observation context is composed of two principal components, namely, *reference context* and *personal context*.

In Chapter 3, we first formalize the life sequence of a person as a sequence of observation contexts experienced over a certain period of time, representing her/his real-world life. We demonstrate that habits can be built by selecting the subsequences from the life sequence. Then, we introduce the general notion of *observation context*, defined by *location* and *time*, as the world evolves in space and time, which is populated by *entities* with the components of *objects*, *functions* and *actions*. We show two key partial views of observation context. One is the *objective view* representing the observation context to establish the actual case as it is. Acknowledging that different people may hold different subjective views of the same scenario due to their knowledge, experiences, and so on, another *subjective view* is proposed to represent observation context as an internal and subjective representation of the world from the view of people.

In Chapter 4, according to the objective view of the observation context, we define the notion of reference context (C_R), which is the representation of a scenario from an objective third-party view. Given that people within C_R possess different subjective representations of the world, even when observing the same scenario. For example, a professor perceives a university as a workplace, while it is a study place for a student. Hence, we define the notion of personal context (C_P) from a subjective view of a person. The C_R and C_P are complementary. C_P enriches the representations of C_R. For instance, a building can be viewed as a shopping place or a workplace. In turn, C_R can fix to some extent the partial and potentially incorrect information in C_P, e.g., due to human wrong labeling [157, 15]. The notions of C_R and C_P are defined under the consideration of the notion of a general observation context we introduced in Chapter 3, as well as a specific purpose.

In Chapter 5, we aim to unify C_P and C_R to generate the *observation context* (C). An OpenStreetMap (OSM) Hierarchy is first constructed to illustrate (spatial) objects and entities in the world. Drawing from this and the C_R definition in Chapter 4, we encode C_R using the C_R teleontology and C_R teleology as the Knowledge Graph (KG) schemas. Similarly, based on the definition of C_P in Chapter 4, C_P is modeled using the C_P teleontology and C_R teleology. Teleontologies model the purpose-specific interrelationships of Entity Types (etypes), inheritances, and properties (including functions and actions) within the scope of knowledge graph engineering. Teleologies are generated from teleontologies, mainly by selecting etypes under a specific purpose and excluding inheritance relations, which clearly encode the elements with relations for any context in space and time. The C_R and C_P teleologies can be populated with real-world data, thus generating C_R and C_P Entity Graphs (EGs), as Knowledge Graphs (KGs). Based on the C_R and C_P teleologies, we specify unification methods to unify C_R and C_P EGs into observation context (C) EGs, which are the big-thick data we aim to generate. We proposed three distinct types of unification methods, namely *entity type and property unification*, *spatio-temporal unification*, *entity unification*. These unification methods are flexible and purpose-driven processes.

In Chapter 6, considering a specific interested *observation purpose*, we distinguish four types of observation enquiries of observation contexts, namely *reference enquiry (R-enquiry)*, *personal enquiry (P-Enquiry)*, *personal-reference enquiry (PR-enquiry)* and *reference-personal Enquiry (RP-enquiry)*. These types of enquiries are posed with different focuses on the aspects of observation contexts. For instance, the R-enquiry is posed to the reference context of the observation context. These four types of enquiries demonstrate the various categories of enquiries on the generated big-thick data. In addition, we distinguish seven types of habit enquiries based on the types of possible activities defining habits occurring in observation contexts, namely *spatial habit enquiry*, *temporal habit enquiry*, *social habit enquiry*, *material habit enquiry*, *action habit enquiry*, and *internal habit enquiry*. Different habit enquiries demonstrate the different habits discovered from human daily lives. Furthermore, given the absence of a standard evaluation for big-thick data, we proposed evaluation metrics including *data size*, *data connectivity*, and *enquiry purpose feasibility* to assess the quality of big-thick data.

In Chapter 7, we implement our method in a SU2-OSM case study, where we start from two datasets. The one is the SmartUnitn2 (SU2) dataset consisting of the time diaries and sensor data about the everyday life of one hundred and fifty-eight students at the University of Trentino over four weeks. Another is the OSM-Trentino dataset, which was extracted from the OpenStreetMap (OSM) data for the northeast Italy region. The big-thick data, referred to SU2-OSM dataset, populating students' life sequences, which are the sequences of observation contexts, generated by unifying the reference context EG built from the OSM-Trentino dataset with a set of personal context EGs built from the SU2 dataset. We used the proposed evaluation metrics to assess the quality of the SU2-OSM dataset, compared to the original SU2 and OSM-Trentino datasets. The results demonstrate that the SU2-OSM dataset occupies the smallest data volume but with the most intensive connectivity providing more purpose-related information. The proposed observation and habit enquiry examples posed on the datasets show that the SU2-OSM dataset can answer a broader range of enquiries and enhance enquiry prediction performance.

8.2 Benefits

This work generates a new type of data, namely *big-thick data*, which is extensive *big data* complemented with highly contextualized *thick data*. This type of data is not only machine-generated, such as the sensor or web data, but is also *purposely* provided by humans, who are the exclusive source of high-quality context-aware data. This level of data is self-explanatory and meaningful data, represents human life in the real world and thus makes machines understand humans better, which can be used to analyze and study the key elements necessary for the development of meaningful lifelong human-in-the-loop human-machine interactions.

In addition, our work generates big-thick data as the life sequences of humans, which can track the human daily lives over time. The element in life sequences, namely, observation context, is the result of the flexible integration of reference context built by an objective view and personal contexts built by-subjective views, we call it *context unification*. In the unification process, the subjective views of people are merged with the objective facts, where these subjective views enrich and annotate objective facts, in turn, the objective facts verify subjective views. Our context unification method can integrate different data sources as big-thick data, and it is a flexible process involved as a function of the specific *purpose*. Thus, multiple possible enquires can be achieved from the generated big-thick data even it is the integrated results from the same datasets.

8.3 Limitations and Future Work

One limitation of the work is that, as part of the evaluation of big-thick data, we only present the types of possible enquires. This does not imply that our capabilities are limited to this scope. It is necessary and possible to develop a general method to list all the possible enquires from big-thick data, guiding researchers to better explore and understand multidimensional information from big-thick data. Given that any enquiry is a question sentence, a potential approach is to refer to the process of *Analisi logica tabella sui complementi* [1], which declares the elements (complements or expansions) that can be added to the minimum sentence, thereby enriching and/or completing its content.

Another limitation is that we only performed predictions for observation enquires using machine learning methods. In our future work, we will conduct statistical analysis of the observation enquires on big-thick data, in addition to machine learning. Statistical analysis provides a foundational framework for understanding and interpreting the relationships between different variables within data. By applying techniques such as regression analysis [138, 147], hypothesis testing [123, 40], and multivariate analysis [103], we can uncover patterns, correlations, and causal relationships that may not be immediately apparent through machine learning alone. For instance, regression analysis can model the relationship between students' behaviors and various contextual factors, such as location and time. Hypothesis testing allows us to determine the statistical significance of observed patterns, providing a rigorous basis for drawing conclusions. Multivariate analysis allows for the examination of multiple variables at the same time, facilitating an understanding of the complex interplay between different factors influencing human behavior. By integrating statistical analysis with machine learning, we can achieve more comprehensive and reliable results of enquires from big-thick data.

[1] https://facolta.unica.it/studiumanistici/files/2015/09/Corso-di-riallineamento-Lingua-italiana-Profssa-Cossu-Tabella-complementi.pdf

In addition, in our future work, we will employ our methodology to conduct more diverse case studies targeting more diverse groups, integrating additional extensive and comprehensive big data and thick data, such as General Transit Feed Specification (GTFS) data, to generate more big-thick datasets. This will explore the generalizability of the methodology. In the future, enriching the evaluation metrics for our methodology will also be considered. For example, the evaluations of observation enquiry predictions will involve more novel algorithms, not constricting in machine learning, which would strengthen the validity of the methodology on intelligently representing and studying the daily lives of humans.

Bibliography

[1] Gregory D Abowd, Anind K Dey, Peter J Brown, Nigel Davies, Mark Smith, and Pete Steggles. Towards a better understanding of context and context-awareness. In *Handheld and Ubiquitous Computing: First International Symposium, HUC'99 Karlsruhe, Germany, September 27–29, 1999 Proceedings 1*, pages 304–307. Springer, 1999.

[2] Khalid Abualsaud, Tarek M Elfouly, Tamer Khattab, Elias Yaacoub, Loay Sabry Ismail, Mohamed Hossam Ahmed, and Mohsen Guizani. A survey on mobile crowd-sensing and its applications in the iot era. *Ieee access*, 7:3855–3881, 2018.

[3] Serap Akgun and Joseph Ciarrochi. Learned resourcefulness moderates the relationship between academic stress and academic performance. *Educational Psychology*, 23(3):287–294, 2003.

[4] Yuen Yuen Ang. Integrating big data and thick data to transform public services delivery. *IBM Center for The Business of Government*, 2019.

[5] Bon Adriel Aseniero, Charles Perin, Wesley Willett, Anthony Tang, and Sheelagh Carpendale. Activity river: Visualizing planned and logged personal activities for reflection. In *Proceedings of the International Conference on Advanced Visual Interfaces*, pages 1–9, 2020.

[6] Wageesha Bangamuarachchi, Anju Chamantha, Lakmal Meegahapola, Haeeun Kim, Salvador Ruiz-Correa, Indika Perera, and Daniel

Gatica-Perez. Inferring mood-while-eating with smartphone sensing and community-based model personalization. *arXiv preprint arXiv:2306.00723*, 2023.

[7] Claudio Bettini, Gabriele Civitarese, and Riccardo Presotto. Caviar: Context-driven active and incremental activity recognition. *Knowledge-Based Systems*, 196:105816, 2020.

[8] G etc. Blank. Online research methods and social theory. *The SAGE Handbook of Online Research Methods*, page 537–549, 2008.

[9] Anders Blok and Morten Axel Pedersen. Complementary social science? quali-quantitative experiments in a big data world. *Big Data & Society*, 1(2):2053951714543908, 2014.

[10] Simone Bocca, Mauro Dragoni, and Fausto Giunchiglia. iTelos - case studies in building domain specific knowledge graphs. In *ISIC-22 International Semantic Intelligence Conference*, 2022.

[11] Simone Bocca, Mauro Dragoni, and Fausto Giunchiglia. itelos—case studies in building domain-specific knowledge graphs. In *Semantic Intelligence: Select Proceedings of ISIC 2022*, pages 277–295. Springer, 2023.

[12] Tom Boellstorff. Making big data, in theory. *First Monday*, 18(10), 2013.

[13] Andrey Bogomolov, Bruno Lepri, Michela Ferron, Fabio Pianesi, and Alex Pentland. Daily stress recognition from mobile phone data, weather conditions and individual traits. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 477–486, 2014.

[14] Andrea Bontempelli, Marcelo Rodas Britez, Xiaoyue Li, Haonan Zhao, Luca Erculiani, Stefano Teso, Andrea Passerini, and Fausto

Giunchiglia. Lifelong personal context recognition. In *HHAI-22 Workshop on Human-Centered Design of Symbiotic Hybrid Intelligence*, 2022. https://arxiv.org/abs/2205.10123.

[15] Andrea Bontempelli, Stefano Teso, Fausto Giunchiglia, and Andrea Passerini. Learning in the wild with incremental skeptical gaussian processes. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence (IJCAI)*, pages 2886–2892, 2021.

[16] Tobias Bornakke and Brian L Due. Big-thick blending: A method for mixing analytical insights from big and thick data sources. *Big Data & Society*, 5(1):2053951718765026, 2018.

[17] Damien Bouchabou, Sao Mai Nguyen, Christophe Lohr, Benoit LeDuc, and Ioannis Kanellos. A survey of human activity recognition in smart homes based on iot sensors algorithms: Taxonomies, challenges, and opportunities with deep learning. *Sensors*, 21(18):6037, 2021.

[18] Danah Boyd and Kate Crawford. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*, 15(5):662–679, 2012.

[19] Gerhard Brewka and Thomas Eiter. Equilibria in heterogeneous nonmonotonic multi-context systems. In *AAAI*, volume 7, pages 385–390, 2007.

[20] Gerhard Brewka, Thomas Eiter, and Michael Fink. Nonmonotonic multi-context systems: A flexible approach for integrating heterogeneous knowledge sources. In *Logic Programming, Knowledge Representation, and Nonmonotonic Reasoning: Essays Dedicated to Michael Gelfond on the Occasion of His 65th Birthday*, pages 233–258. Springer, 2011.

[21] Gerhard Brewka, Thomas Eiter, Michael Fink, and Antonius Weinzierl. Managed multi-context systems. In *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, volume 22, page 786, 2011.

[22] Gerhard Brewka, Stefan Ellmauthaler, Ricardo Gonçalves, Matthias Knorr, João Leite, and Jörg Pührer. Reactive multi-context systems: Heterogeneous reasoning in dynamic environments. *Artificial Intelligence*, 256:68–104, 2018.

[23] Gerhard Brewka, Stefan Ellmauthaler, and Jörg Pührer. Multi-context systems for reactive reasoning in dynamic environments. In *ECAI*, pages 159–164, 2014.

[24] Jerome Bruner. The narrative construction of reality. *Critical inquiry*, 18(1):1–21, 1991.

[25] Erhan Bulbul, Aydin Cetin, and Ibrahim Alper Dogru. Human activity recognition using smartphones. In *2018 2nd international symposium on multidisciplinary studies and innovative technologies (ismsit)*, pages 1–6. IEEE, 2018.

[26] Matteo Busso et al. The ilog methodology for fostering valid and reliable big thick data. 2024.

[27] Matteo Busso, Xiaoyue Li, et al. A context model for collecting diversity-aware data. In *CEUR WORKSHOP PROCEEDINGS*, volume 3456, pages 214–219. RWTH Aachen, 2023.

[28] Andrea Calì, Diego Calvanese, Giuseppe De Giacomo, and Maurizio Lenzerini. Data integration under integrity constraints. In *Advanced Information Systems Engineering: 14th International Conference, CAiSE 2002 Toronto, Canada, May 27–31, 2002 Proceedings 14*, pages 262–279. Springer, 2002.

[29] Junting Chen and Hao Huang. The influence of big five personality traits on depression and suicidal behavior. In *The Association Between Depression and Suicidal Behavior*. IntechOpen, 2024.

[30] John Curran. Big data or 'big ethnographic data'? positioning big data within the ethnographic space. In *Ethnographic praxis in industry conference proceedings*, volume 2013, pages 62–73. Wiley Online Library, 2013.

[31] Tushar Kant Das and P Mohan Kumar. Big data analytics: A framework for unstructured data analysis. *International Journal of Engineering Science & Technology*, 5(1):153, 2013.

[32] Virginia M Deroma, John B Leach, and J Patrick Leverett. The relationship between depression and college academic performance. *College Student Journal*, 43(2):325–335, 2009.

[33] Anind K Dey. Understanding and using context. *Personal and ubiquitous computing*, 5:4–7, 2001.

[34] Anind K Dey, Gregory D Abowd, and Daniel Salber. A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications. *Human–Computer Interaction*, 16(2-4):97–166, 2001.

[35] Ali Enes Dingil, Joerg Schweizer, Federico Rupi, and Zaneta Stasiskiene. Transport indicator analysis and comparison of 151 urban areas, based on open source data. *European Transport Research Review*, 10:1–9, 2018.

[36] Trinh Minh Tri Do and Daniel Gatica-Perez. The places of our lives: Visiting patterns and automatic labeling from longitudinal smartphone data. *IEEE Transactions on Mobile Computing*, 13(3):638–648, 2013.

[37] Alessandro Duranti and Charles Goodwin. *Rethinking context: Language as an interactive phenomenon*. Number 11. Cambridge University Press, 1992.

[38] Nathan Eagle, Alex Pentland, and David Lazer. Inferring friendship network structure by using mobile phone data. *Proceedings of the national academy of sciences*, 106(36):15274–15278, 2009.

[39] Nathan Eagle and Alex Sandy Pentland. Reality mining: sensing complex social systems. *Personal and ubiquitous computing*, 10(4):255–268, 2006.

[40] Frank Emmert-Streib and Matthias Dehmer. Understanding statistical hypothesis testing: The logic of statistical inference. *Machine Learning and Knowledge Extraction*, 1(3):945–962, 2019.

[41] Jérôme Euzenat. An api for ontology alignment. In *International Semantic Web Conference*, pages 698–712. Springer, 2004.

[42] Jianqing Fan, Fang Han, and Han Liu. Challenges of big data analysis. *National science review*, 1(2):293–314, 2014.

[43] Daniel Feldmeyer, Claude Meisch, Holger Sauter, and Joern Birkmann. Using openstreetmap data and machine learning to generate socio-economic indicators. *ISPRS International Journal of Geo-Information*, 9(9):498, 2020.

[44] Heather Ford. Big data and small: Collaborations between ethnographers and data scientists. *Big Data & Society*, 1(2):2053951714544337, 2014.

[45] Aldo Gangemi, Nicola Guarino, Claudio Masolo, Alessandro Oltramari, and Luc Schneider. Sweetening ontologies with dolce. In *International conference on knowledge engineering and knowledge management*, pages 166–181. Springer, 2002.

[46] Clifford Geertz. Thick description: Toward an interpretive theory of culture. In *The cultural geography reader*, pages 41–51. Routledge, 2008.

[47] Clifford Geertz. *The interpretation of cultures*. Basic books, 2017.

[48] Chiara Ghidini and Fausto Giunchiglia. Local models semantics, or contextual reasoning= locality+ compatibility. *Artificial intelligence*, 127(2):221–259, 2001.

[49] F Girardin. Insights from network data analysis that yield field observations. *Ethnography Matters. Retrieved November*, 20:2019, 2013.

[50] Nicolò Alessandro Girardini, Simone Centellegher, Andrea Passerini, Ivano Bison, Fausto Giunchiglia, and Bruno Lepri. Adaptation of student behavioural routines during covid-19: a multimodal approach. *EPJ Data Science*, 12(1):55, 2023.

[51] F Giunchiglia, A Zamboni, M Bagchi, S Bocca, et al. Stratified data integration. In *CEUR WORKSHOP PROCEEDINGS*, volume 2873, pages 1–15. CEUR-WS, 2021.

[52] Fausto Giunchiglia, Enrico Bignotti, and Mattia Zeni. Personal context modelling and annotation. In *2017 IEEE international conference on pervasive computing and communications workshops (PerCom workshops)*, pages 117–122. IEEE, 2017.

[53] Fausto Giunchiglia, Simone Bocca, Mattia Fumagalli, Mayukh Bagchi, and Alessio Zamboni. itelos–purpose driven knowledge graph generation. *arXiv preprint arXiv:2105.09418*, 2021.

[54] Fausto Giunchiglia, Simone Bocca, Mattia Fumagalli, Mayukh Bagchi, and Alessio Zamboni. Popularity driven data integration. In *Iberoamerican Knowledge Graphs and Semantic Web Conference*, pages 277–284. Springer, 2022.

[55] Fausto Giunchiglia, Marcelo Rodas Britez, Andrea Bontempelli, and Xiaoyue Li. Streaming and learning the personal context. In *Twelfth International Workshop Modelling and Reasoning in Context*, page 19, 2021.

[56] Fausto Giunchiglia, Matteo Busso, Mattia Zeni, Ivano Bison, et al. A survey on students' daily routines and academic performance at the university of trento. 2022.

[57] Fausto Giunchiglia et al. Contextual reasoning. *Epistemologia, special issue on I Linguaggi e le Macchine*, 16:345–364, 1993.

[58] Fausto Giunchiglia and Mattia Fumagalli. Teleologies: Objects, actions and functions. In *ER- International Conference on Conceptual Modeling (ER2017)*, pages 520–534. ICCM, 2017.

[59] Fausto Giunchiglia and Mattia Fumagalli. Entity type recognition–dealing with the diversity of knowledge. In *Proc. iInt conference on principles of knowledge representation and reasoning (KRR 2020)*, volume 17, pages 414–423, 2020.

[60] Fausto Giunchiglia and Xiaoyue Li. Big-thick data generation via reference and personal context unification. In *ECAI 2024*, pages 1975–1984. IOS Press, 2024.

[61] Fausto Giunchiglia, Xiaoyue Li, Matteo Busso, and Marcelo Rodas-Britez. A context model for personal data streams. In *Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data*. Springer, 2022.

[62] Fausto Giunchiglia and Luciano Serafini. Multilanguage hierarchical logics, or: how we can do without modal logics. *Artificial intelligence*, 65(1):29–70, 1994.

[63] Fausto Giunchiglia, Mattia Zeni, and Enrico Big. Personal context recognition via reliable human-machine collaboration. In *2018 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, pages 379–384. IEEE, 2018.

[64] Fausto Giunchiglia, Mattia Zeni, Elisa Gobbi, Enrico Bignotti, and Ivano Bison. Mobile social media usage and academic performance. *Computers in Human Behavior*, 82:177–185, 2018.

[65] Shravya G Gowda, Shyleshwari M Shetty, MS Darshini, and D Rajani. Analysis of human activity detection using machine learning approaches. *SN Computer Science*, 4(2):177, 2023.

[66] A Yair Grinberger, Marco Minghini, Levente Juhász, Godwin Yeboah, and Peter Mooney. Osm science—the academic study of the openstreetmap project, data, contributors, community, and applications, 2022.

[67] Nicola Guarino and Giancarlo Guizzardi. Relationships and events: towards a general theory of reification and truthmaking. In *25th International Conference AIxIA, Genova, Italy, November 29–December 1, 2016, Proceedings XV*, pages 237–249. Springer, 2016.

[68] Nicola Guarino and Christopher A Welty. An overview of ontoclean. *Handbook on ontologies*, pages 201–220, 2009.

[69] Juan Guerrero-Ibáñez, Sherali Zeadally, and Juan Contreras-Castillo. Sensor technologies for intelligent transportation systems. *Sensors*, 18(4):1212, 2018.

[70] Mary E Haines, Margaret P Norris, and Deborah A Kashy. The effects of depressed mood on academic performance in college students. *Journal of College Student Development*, 1996.

[71] Gabriella M Harari, Sandrine R Müller, Clemens Stachl, Rui Wang, Weichen Wang, Markus Bühner, Peter J Rentfrow, Andrew T Campbell, and Samuel D Gosling. Sensing sociability: Individual differences in young adults' conversation, calling, texting, and app use behaviors in daily life. *Journal of personality and social psychology*, 119(1):204, 2020.

[72] Daniel Hintze, Philipp Hintze, Rainhard D Findling, and René Mayrhofer. A large-scale, long-term analysis of mobile device usage characteristics. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(2):1–21, 2017.

[73] Marvin Hofer, Daniel Obraczka, Alieh Saeedi, Hanna Köpcke, and Erhard Rahm. Construction of knowledge graphs: State and challenges. *decision-making*, 10:11, 2023.

[74] Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard De Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, et al. Knowledge graphs. *ACM Computing Surveys (Csur)*, 54(4):1–37, 2021.

[75] Xiao Huang, Jingyuan Zhang, Dingcheng Li, and Ping Li. Knowledge graph embedding based question answering. In *Proceedings of the twelfth ACM international conference on web search and data mining*, pages 105–113, 2019.

[76] Vito Janko and Mitja Luštrek. A general framework for making context-recognition systems more energy efficient. *Sensors*, 21(3):766, 2021.

[77] Natasha Jaques, Sara Taylor, Asaph Azaria, Asma Ghandeharioun, Akane Sano, and Rosalind Picard. Predicting students' happiness from physiology, phone, mobility, and behavioral data. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 222–228. IEEE, 2015.

[78] Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and S Yu Philip. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE Trans.neural networks and learning systems*, 33(2):494–514, 2021.

[79] Long Jiang, Bin Gao, Jun Gu, Yuanpeng Chen, Zhao Gao, Xiaole Ma, Keith M Kendrick, and Wai Lok Woo. Wearable long-term social sensing for mental wellbeing. *IEEE Sensors Journal*, 19(19):8532–8542, 2018.

[80] Jamal Jokar Arsanjani, Alexander Zipf, Peter Mooney, and Marco Helbich. An introduction to openstreetmap in geographic information science: Experiences, research, and applications. *OpenStreetMap in GIScience: Experiences, research, and applications*, pages 1–15, 2015.

[81] Michael I Jordan and Tom M Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015.

[82] Imène Jraidi and Claude Frasson. Student's uncertainty modeling through a multimodal sensor-based approach. *Journal of Educational Technology & Society*, 16(1):219–230, 2013.

[83] Karthik Kambatla, Giorgos Kollias, Vipin Kumar, and Ananth Grama. Trends in big data analytics. *Journal of parallel and distributed computing*, 74(7):2561–2573, 2014.

[84] Valentin Kassarnig, Enys Mones, Andreas Bjerre-Nielsen, Piotr Sapiezynski, David Dreyer Lassen, and Sune Lehmann. Academic performance and behavioral patterns. *EPJ Data Science*, 7:1–16, 2018.

[85] Rob Kitchin. Big data, new epistemologies and paradigm shifts. *Big data & society*, 1(1):2053951714528481, 2014.

[86] Ron Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Montreal, Canada, 1995.

[87] Roman Kontchakov, Mariano Rodriguez-Muro, and Michael Zakharyaschev. Ontology-based data access with databases: A short course. In *Reasoning Web International Summer School*, pages 194–229. Springer, 2013.

[88] Adrianna Kozierkiewicz and Marcin Pietranik. A formal framework for the ontology evolution. In *Intelligent Information and Database Systems: 11th Asian Conference, ACIIDS 2019, Yogyakarta, Indonesia, April 8–11, 2019, Proceedings, Part I 11*, pages 16–27. Springer, 2019.

[89] Robin Kraft, Winfried Schlee, Michael Stach, Manfred Reichert, Berthold Langguth, Harald Baumeister, Thomas Probst, Ronny Hannemann, and Rüdiger Pryss. Combining mobile crowdsensing and ecological momentary assessments in the healthcare domain. *Frontiers in neuroscience*, 14:496147, 2020.

[90] Davide Lanti, Martin Ignacio Rezk, Guohui Xiao, and Diego Calvanese. The npd benchmark: Reality check for obda systems. In *Advances in database technology-EDBT 2015: 18th International Conference on Extending Database Technology, Brussels, Belgium, March 23-27, 2015, proceedings*, pages 617–628. University of Konstanz, University Library, 2015.

[91] Davide Lanti, Guohui Xiao, and Diego Calvanese. Vig: Data scaling for obda benchmarks. *Semantic Web*, 10(2):413–433, 2019.

[92] Reed Larson and Mihaly Csikszentmihalyi. The experience sampling method. In *Flow and the foundations of positive psychology*, pages 21–34. Springer, 2014.

[93] Thinh Vinh Le, Huan Thien Tran, and Duy L Le. Dynamic traffic optimization in smart cities (dtos): Integrating openstreetmap, iot, and fog computing. *SN Computer Science*, 5(7):823, 2024.

[94] Maurizio Lenzerini. Data integration: A theoretical perspective. In *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 233–246, 2002.

[95] Qingqing Li, Jorge Peña Queralta, Tuan Nguyen Gia, Zhuo Zou, and Tomi Westerlund. Multi-sensor fusion for navigation and mapping in autonomous vehicles: Accurate localization in urban environments. *Unmanned Systems*, 8(03):229–237, 2020.

[96] Xiaoyue Li, Marcelo Rodas-Britez, Matteo Busso, and Fausto Giunchiglia. Representing habits as streams of situational contexts. In *International Conference on Advanced Information Systems Engineering*, pages 86–92. Springer, 2022.

[97] Zeju Li, Linya Cheng, Chunhong Zhang, Xinning Zhu, and Hui Zhao. Multi-source education knowledge graph construction and fusion for college curricula. In *2023 IEEE International Conference on Advanced Learning Technologies (ICALT)*, pages 359–363. IEEE, 2023.

[98] Robert LiKamWa, Yunxin Liu, Nicholas D Lane, and Lin Zhong. Moodscope: Building a mood sensor from smartphone usage patterns. In *Proceeding of the 11th annual international conference on Mobile systems, applications, and services*, pages 389–402, 2013.

[99] Aleksej Logacjov, Kerstin Bach, Atle Kongsvold, Hilde Bremseth Bårdstu, and Paul Jarle Mork. Harth: a human activity recognition dataset for machine learning. *Sensors*, 21(23):7853, 2021.

[100] Chuangtao Ma. Knowledge-enriched schema mapping: A preliminary case study of e-medsolution system. *Eötvös Loránd University, Tech. Rep.*, 2020.

[101] Chuangtao Ma, Bálint Molnár, Ádám Tarcsi, and András Benczúr. Knowledge enriched schema matching framework for heterogeneous data integration. In *2022 IEEE 2nd Conference on Information Technology and Data Science (CITDS)*, pages 183–188. IEEE, 2022.

[102] Batta Mahesh. Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet]*, 9(1):381–386, 2020.

[103] Kanti V Mardia, John T Kent, and Charles C Taylor. *Multivariate analysis*, volume 88. John Wiley & Sons, 2024.

[104] Maren Mayer, Daniel W Heck, and Franz-Benjamin Mocnik. Using openstreetmap as a data source in psychology and the social sciences. 2022.

[105] Hannah McCarthy, Henry WW Potts, and Abigail Fisher. Physical activity behavior before, during, and after covid-19 restrictions: Longitudinal smartphone-tracking study of adults in the united kingdom. *Journal of medical Internet research*, 23(2):e23701, 2021.

[106] John McCarthy. Generality in artificial intelligence. *Communications of the ACM*, 30(12):1030–1035, 1987.

[107] John McCarthy. From here to human-level ai. *Artificial Intelligence*, 171(18):1174–1182, 2007.

[108] John McCarthy and Patrick J Hayes. Some philosophical problems from the standpoint of artificial intelligence. In *Readings in artificial intelligence*, pages 431–450. Elsevier, 1981.

[109] Abhinav Mehrotra, Jo Vermeulen, Veljko Pejovic, and Mirco Musolesi. Ask, but don't interrupt: the case for interruptibility-aware mobile experience sampling. In *Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers*, pages 723–732, 2015.

[110] Christian Meurisch, Artur Gogel, Benedikt Schmidt, Timo Nolle, Frederik Janssen, Immanuel Schweizer, and Max Mühlhäuser. Capturing daily student life by recognizing complex activities using smartphones. In *Proceedings of the 14th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*, pages 156–165, 2017.

[111] Christian Meurisch, Benedikt Schmidt, Michael Scholz, Immanuel Schweizer, and Max Mühlhäuser. Labels: Quantified self app for human activity sensing. In *Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers*, pages 1413–1422, 2015.

[112] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[113] CARLO. SCANNAPIECO BATINI (MONICA.). *DATA AND INFORMATION QUALITY: Dimensions, Principles and Techniques*. SPRINGER, 2018.

[114] Raul Montoliu, Jan Blom, and Daniel Gatica-Perez. Discovering places of interest in everyday life from smartphone data. *Multimedia tools and applications*, 62(1):179–207, 2013.

[115] George Oguntala, Raed Abd-Alhameed, Stephen Jones, James Noras, Mohammad Patwary, and Jonathan Rodriguez. Indoor location identification technologies for real-time iot-based applications: An inclusive survey. *Computer Science Review*, 30:55–79, 2018.

[116] Matthew Owens, Jim Stevenson, Julie A Hadwin, and Roger Norgate. Anxiety and depression in academic performance: An exploration of the mediating factors of worry and working memory. *School psychology international*, 33(4):433–449, 2012.

[117] Ciyuan Peng, Feng Xia, Mehdi Naseriparsa, and Francesco Osborne. Knowledge graphs: Opportunities and challenges. *Artificial Intelligence Review*, 56(11):13071–13102, 2023.

[118] Evangelos Pournaras, Atif Nabi Ghulam, Renato Kunz, and Regula Hänggli. Crowd sensing and living lab outdoor experimentation made easy. *IEEE Pervasive Computing*, 21(1):18–27, 2021.

[119] Davy Preuveneers and Yolande Berbers. Mobile phones assisting with health self-care: a diabetes case study. In *Proceedings of the 10th international conference on Human computer interaction with mobile devices and services*, pages 177–186, 2008.

[120] Sonam Priyanka, Sandip Dutta, and Soubhik Chakraborty. Human tracking system based on gps and iot (internet of things). In *Proceeding of the International Conference on Computer Networks, Big Data and IoT (ICCBI-2018)*, pages 514–521. Springer, 2020.

[121] Shen Qiang, Stefano Teso, Wanyi Zhang, Hao Xu, Fausto Giunchiglia, et al. Multi-modal subjective context modelling and recognition. In *24th European Conference on Artificial Intelligence*. ECAI, 2020.

[122] Mohammad Arief Faizal Rachman and Gusti Ayu Putri Saptawati. Database integration based on combination schema matching approach (case study: Multi-database of district health information system). In *2017 2nd International conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, pages 430–435. IEEE, 2017.

[123] Adrian E Raftery, WR Gilks, S Richardson, and D Spiegelhalter. Hypothesis testing and model. *Markov chain Monte Carlo in practice*, 1:165–87, 1995.

[124] Elangovan Ramanujam, Thinagaran Perumal, and S Padmavathi. Human activity recognition with smartphone and wearable sensors using deep learning techniques: A review. *IEEE Sensors Journal*, 21(12):13029–13040, 2021.

[125] Joan C Rogers. Habits: Do we practice what we preach? *The Occupational Therapy Journal of Research*, 20(1_suppl):119S–122S, 2000.

[126] Dominik Rüegger, Mirjam Stieger, Marcia Nißen, Mathias Allemand, Elgar Fleisch, and Tobias Kowatsch. How are personality states associated with smartphone data? *European Journal of Personality*, 34(5):687–713, 2020.

[127] Seref Sagiroglu and Duygu Sinanc. Big data: A review. In *2013 international conference on collaboration technologies and systems (CTS)*, pages 42–47. IEEE, 2013.

[128] Bill N Schilit and Marvin M Theimer. Disseminating active map information to mobile hosts. *IEEE network*, 8(5):22–32, 1994.

[129] Erich Schubert, Jörg Sander, Martin Ester, Hans Peter Kriegel, and Xiaowei Xu. Dbscan revisited, revisited: why and how you should (still)

use dbscan. *ACM Transactions on Database Systems (TODS)*, 42(3):1–21, 2017.

[130] Mahsa Sheikh, Meha Qassem, and Panicos A Kyriacou. Wearable, environmental, and smartphone-based passive sensing for mental health monitoring. *Frontiers in Digital Health*, 3:662811, 2021.

[131] Qiang Shen, Stefano Teso, Fausto Giunchiglia, and Hao Xu. To transfer or not to transfer and why? meta-transfer learning for explainable and controllable cross-individual activity recognition. *Electronics*, 12(10):2275, 2023.

[132] Daqian Shi et al. Knowledge graph extension by entity type recognition. 2024.

[133] Daqian Shi, Xiaoyue Li, and Fausto Giunchiglia. Kae: A property-based method for knowledge graph alignment and extension. *Journal of Web Semantics*, 82:100832, 2024.

[134] Saul Shiffman, Arthur A Stone, and Michael R Hufford. Ecological momentary assessment. *Annu. Rev. Clin. Psychol.*, 4:1–32, 2008.

[135] Wesllen Sousa Lima, Eduardo Souto, Khalil El-Khatib, Roozbeh Jalali, and Joao Gama. Human activity recognition using inertial sensors in a smartphone: An overview. *Sensors*, 19(14):3213, 2019.

[136] Leo Strobel and Marco Pruckner. Omod: An open-source tool for creating disaggregated mobility demand based on openstreetmap. *Computers, Environment and Urban Systems*, 106:102029, 2023.

[137] Jessie Sun, Mijke Rhemtulla, and Simine Vazire. Eavesdropping on missing data: What are university students doing when they miss experience sampling reports? *Personality and Social Psychology Bulletin*, 47(11):1535–1549, 2021.

[138] Alan O Sykes. An introduction to regression analysis. 1993.

[139] Sanju Tiwari, Fatima N Al-Aswadi, and Devottam Gaurav. Recent trends in knowledge graphs: theory and practice. *Soft Computing*, 25:8337–8355, 2021.

[140] N Umapathi, Saiteja Sabbani, and S Poovarasan. Person location tracking using global positioning system and esp8266 with internet of things. In *International Conference on Futuristic Communication and Network Technologies*, pages 211–217. Springer, 2020.

[141] Yonatan Vaizman, Katherine Ellis, and Gert Lanckriet. Recognizing detailed human context in the wild from smartphones and smartwatches. *IEEE pervasive computing*, 16(4):62–74, 2017.

[142] Yonatan Vaizman, Nadir Weibel, and Gert Lanckriet. Context recognition in-the-wild: Unified model for multi-modal sensors and multi-label classification. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(4):1–22, 2018.

[143] Niels van Berkel, Jorge Goncalves, Simo Hosio, Zhanna Sarsenbayeva, Eduardo Velloso, and Vassilis Kostakos. Overcoming compliance bias in self-report studies: A cross-study analysis. *International Journal of Human-Computer Studies*, 134:1–12, 2020.

[144] Niels van Berkel, Jorge Goncalves, Peter Koval, Simo Hosio, Tilman Dingler, Denzil Ferreira, and Vassilis Kostakos. Context-informed scheduling and analysis: improving accuracy of mobile self-reports. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2019.

[145] Panos Vassiliadis. A survey of extract–transform–load technology. *International Journal of Data Warehousing and Mining (IJDWM)*, 5(3):1–27, 2009.

[146] Panos Vassiliadis, Alkis Simitsis, and Spiros Skiadopoulos. Conceptual modeling for etl processes. In *Proceedings of the 5th ACM international workshop on Data Warehousing and OLAP*, pages 14–21, 2002.

[147] Ilona WM Verburg, Nicolette F de Keizer, Evert de Jonge, and Niels Peek. Comparison of regression methods for modeling intensive care length of stay. *PloS one*, 9(10):e109684, 2014.

[148] Daniel T Wagner, Andrew Rice, and Alastair R Beresford. Device analyzer: Understanding smartphone usage. In *Mobile and Ubiquitous Systems: Computing, Networking, and Services: 10th International Conference, MOBIQUITOUS 2013, Revised Selected Papers 10*, pages 195–208. Springer, 2014.

[149] Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T Campbell. Studentlife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In *Proceedings of ACM - UBICOMP*, pages 3–14, 2014.

[150] Rui Wang, Gabriella Harari, Peilin Hao, Xia Zhou, and Andrew T Campbell. Smartgpa: how smartphones can assess and predict academic performance of college students. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*, pages 295–306, 2015.

[151] Rui Wang, Weichen Wang, Alex DaSilva, Jeremy F Huckins, William M Kelley, Todd F Heatherton, and Andrew T Campbell. Tracking depres-

sion dynamics in college students using mobile phone and wearable sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(1):1–26, 2018.

[152] Xuhai Xu, Xin Liu, Han Zhang, Weichen Wang, Subigya Nepal, Yasaman Sefidgar, Woosuk Seo, Kevin S Kuehn, Jeremy F Huckins, Margaret E Morris, et al. Globem: Cross-dataset generalization of longitudinal human behavior modeling. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(4):1–34, 2023.

[153] De Jong Yeong, Gustavo Velasco-Hernandez, John Barry, and Joseph Walsh. Sensor and sensor fusion technology in autonomous vehicles: A review. *Sensors*, 21(6):2140, 2021.

[154] Lin Yue, Haonan Zhao, Yiqin Yang, Dongyuan Tian, Xiaowei Zhao, and Minghao Yin. A mimic learning method for disease risk prediction with incomplete initial data. In *International Conference on Database Systems for Advanced Applications*, pages 392–396. Springer, 2019.

[155] Mattia Zeni, Ivano Bison, Britta Gauckler, Fernando Reis, and Fausto Giunchiglia. Improving time use measurement with personal big collection - the experience of the european big data hackathon 2019. *Journal of Official Statistics*, 2020.

[156] Mattia Zeni, Ilya Zaihrayeu, and Fausto Giunchiglia. Multi-device activity logging. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, pages 299–302, 2014.

[157] Mattia Zeni, Wanyi Zhang, Enrico Bignotti, Andrea Passerini, and Fausto Giunchiglia. Fixing mislabeling by human annotators leveraging conflict resolution and prior knowledge. *Proc. of ACM IMWUT*, 3(1):32, 2019.

[158] Wanyi Zhang, Qiang Shen, Stefano Teso, Bruno Lepri, Andrea Passerini, Ivano Bison, and Fausto Giunchiglia. Putting human behavior predictability in context. *EPJ Data Science*, 10(1):42, 2021.

[159] Haonan Zhao, Ivan Kayongo, Leonardo Malcotti, and Fausto Giunchiglia. Human-AI collaborative big-thick data collection. *arXiv preprint arXiv:2404.17602*, 2024.

Appendix A

OpenStreetMap (OSM) Ontology Generation

The OSM Ontology is generated based on the OSM data descriptions in the document of *OpenStreetMap Data in Layered GIS Format*[1] (GIS file).

Main process: The process of formalizing the OSM Ontology from top to bottom layers involves the following main steps.

1. According to the OSM places' geometry property (values point, line or polygon), the root `OSM_Place` Entity Type (etype) has three subclasses: `Point`, `Line`, and `Polygon`.

2. According to the `Layer` information as of the GIS file, `Point`, `Line` and `Polygon` have their subclasses. Specifically, `Point` has subclasses of `Natural`, `Place`, `Place_Of_Worship`, `Point_Of_Interest`, `Traffic`, and `Transport`. `Line` has subclasses of `Railway`, `Road` and `Waterway`. `Polygon` has subclasses of `Building`, `Landuse`, and `Water`.

3. The further `Layer` information is used to represent the hierarchy of the subclasses as listed in step 2. For example, `Point_Of_Interest` has subclasses including `Accommodation`, `Catering`, `Health`, `Leisure`,

Miscellaneous_Point_Of_Interest, Money, Public, Shopping, and Tourism.

4. The fclass information is used to define the further subclasses. For example, Money has subclasses of Atm and bank.

Etype naming principle: The OSM Ontology is generated in accordance with the following principles:

- All names of etypes start with the upper letter.
- We use the underscore symbol (_) to replace the black space in the etype names, e.g., Point_Of_Interest.
- If two etypes have the same name, they are distinguished by adding their superclass before their name, e.g., Waterway_River and Water_River.
- If a superclass has only one subclass, the superclass is deleted and only retains the subclass. For example, Water_Traffic only has one subclass Ferry_Terminal, Ferry_Terminal are kept in the OSM Ontology while Water_Traffic is dropped.
- All etypes are marked with comments, which are based on the GIS file, such as the comment of Hostel is *a hostel (offering cheap accommodation, often bunk beds in dormitories).*

Appendix B

OpenStreetMap (OSM) Ontology Overview

We list the entity types (etypes) in the OSM Ontology as below, where the subclass is annoated as the tab-indented text below its superclass. For example, `Line` is the subclass of `OSM_Place`, `Railway` is the subclass of `Line`.

OSM_Place
- Line
  - Railway
    - Cable_Car
    - Chair_Lift
    - Drag_Lift
    - Funicular
    - Gondola
    - Goods
    - Light_Rail
    - Miniature
    - Monorail
    - Narrow_Gauge
    - Other_Lift
    - Rack

- Rail
- Subway
- Tram

Road

- Highway_Link
  - Motorway_Link
  - Primary_Link
  - Secondary_Link
  - Tertiary_Link
  - Trunk_Link
- Major_Road
  - Motorway
  - Primary
  - Secondary
  - Tertiary
  - Trunk
- Minor_Road
  - Busway
  - Living_Street
  - Pedestrian
  - Minor_Road_Residential
  - Unclassified
- Path_Unsuitable_For_Cars
  - Bridleway
  - Cycleway
  - Footway
  - Path
  - Step
- Unknown

        Very_Small_Road
            Very_Small_Road_Service
            Track
            Track_Grade1
            Track_Grade2
            Track_Grade3
            Track_Grade4
            Track_Grade5
    Waterway
        Canal
        Drain
        Waterway_River
        Stream
Point
    Natural
        Beach
        Cave_Entrance
        Cliff
        Natural_Glacier
        Mine
        Peak
        Spring
        Tree
        Volcano
    Place
        City
        County
        Dwelling
        Farm

Hamlet
Island
Locality
National_Capital
Region
Suburb
Town
Village

Place_Of_Worship
- Buddhist
- Christian
- Christian_Anglican
- Christian_Baptist
- Christian_Catholic
- Christian_Evangelical
- Christian_Lutheran
- Christian_Methodist
- Christian_Mormon
- Christian_Orthodox
- Christian_Protestant
- Hindu
- Jewish
- Muslim
- Muslim_Shia
- Muslim_Sunni
- Shintoist
- Sikh
- Taoist

Point_Of_Interest

- Accommodation
  - Indoor
    - Bed_And_Breakfast
    - Chalet
    - Guesthouse
    - Hostel
    - Hotel
    - Motel
  - Outdoor
    - Alpine_Hut
    - Camp_Site
    - Caravan_Site
    - Shelter
- Catering
  - Bar
  - Biergarten
  - Cafe
  - Fast_Food
  - Food_Court
  - Pub
  - Restaurant
- Health
  - Clinic
  - Dentist
  - Doctor
  - Hospital
  - Pharmacy
  - Veterinary
- Leisure

- Cinema
- Dog_Park
- Nightclub
- Leisure_Park
- Playground
- Sport
  - Golf_Course
  - Ice_Rink
  - Pitch
  - Sports_Centre
  - Stadium
  - Swimming_Pool
  - Tennis_Court
- Theatre

Miscellaneous_Point_Of_Interest

- Bench
- Camera_Surveillance
- Drinking_Water
- Emergency_Access
- Emergency_Phone
- Fire_Hydrant
- Fountain
- Hunting_Stand
- Lighthouse
- Toilet
- Tower
- Tower_Comms
- Tower_Observation
- Waste_Basket

- Wastewater_Plant
- Water_Mill
- Water_Tower
- Water_Well
- Water_Works
- Windmill

Money

- Atm
- Bank

Public

- Arts_Centre
- Community_Centre
- Courthouse
- Education
  - College
  - Kindergarten
  - Public_Building
  - School
  - University
- Embassy
- Fire_Station
- Graveyard
- Library
- Market_Place
- Nursing_Home
- Police
- Post_Box
- Post_Office
- Prison

Recycling
Recycling_Clothes
Recycling_Glass
Recycling_Metal
Recycling_Paper
Telephone
Town_Hall
Shopping
Bakery
Beauty_Shop
Beverage
Bicycle_Rental
Bicycle_Shop
Bookshop
Butcher
Car_Dealership
Car_Rental
Car_Repair
Car_Sharing
Car_Wash
Chemist
Clothes
Computer_Shop
Convenience
Department_Store
Doityourself
Florist
Furniture_Shop
Garden_Centre

General
Gift_Shop
Greengrocer
Hairdresser
Jeweller
Kiosk
Laundry
Mall
Mobile_Phone_Shop
Newsagent
Optician
Outdoor_Shop
Shoe_Shop
Sports_Shop
Stationery
Supermarket
Toy_Shop
Travel_Agent
Vending_Cigarette
Vending_Machine
Vending_Parking
Video_Shop
Tourism
Destination
Archaeological
Art
Attraction
Battlefield
Castle

Fort

Memorial

Monument

Museum

Picnic_Site

Ruin

Theme_Park

Viewpoint

Wayside_Cross

Wayside_Shrine

Zoo

Information

Tourist_Board

Tourist_Guidepost

Tourist_Info

Tourist_Map

Traffic

Crossing

Ford

Fuel_And_Parking

Fuel

Parking

Parking_Bicycle

Parking_Multistorey

Parking_Site

Parking_Underground

Fuel_And_Parking_Service

Mini_Roundabout

Motorway_Junction

Speed_Camera

Stop

Street_Lamp

Traffic_Signal

Turning_Circle

Water_Traffic

Dam

Lock_Gate

Marina

Pier

Slipway

Waterfall

Weir

Transport

Aerialway_Station

Air_Traffic

Airfield

Airport

Apron

Helipad

Bus_Station

Bus_Stop

Ferry_Terminal

Railway_Halt

Railway_Station

Taxi_Rank

Tram_Stop

Polygon

Building

- Landuse
  - Allotment
  - Cemetery
  - Commercial
  - Farmland
  - Farmyard
  - Forest
  - Grass
  - Heath
  - Industrial
  - Landuse_Residential
  - Landuse_Park
  - Meadow
  - Military
  - National_Park
  - Nature_Reserve
  - Orchard
  - Quarry
  - Recreation_Ground
  - Retail
  - Scrub
  - Vineyard
- Water
  - Dock
  - Reservoir
  - Water_Glacier
  - Water_River
  - Unspecified_Water
  - Wetland

The data properties of etypes in the OSM Ontology are shown in Figure B.1. Details of etypes and data properties in the OSM Ontology can be seen in the *OSM Ontology.owl*[2] opened by Protégé.

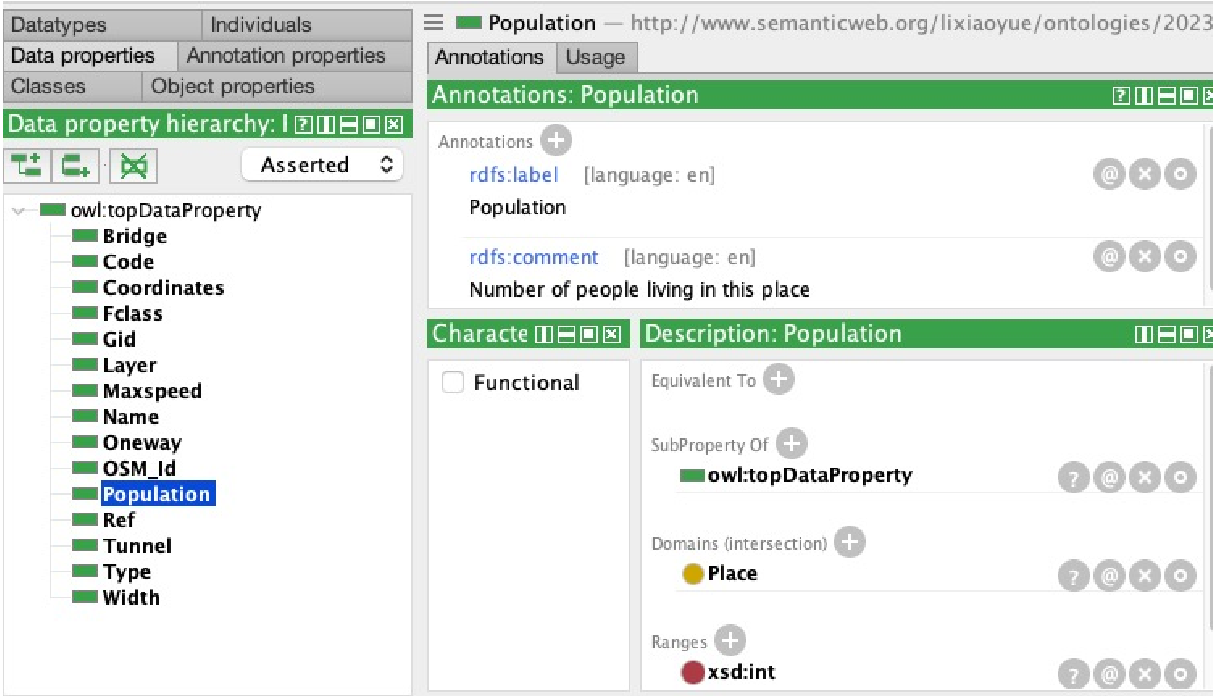

Figure B.1: Data properties of etypes in OSM ontology in protégé.

Appendix C

Variables in OSM-Trentino Dataset

The details of variables of places in the OSM-Trentino dataset can be referred to in the file of *OpenStreetMap Data in Layered GIS*[1]. All the places in the OSM-Trentino dataset have the following variables.

- *gid*: Place Identifier, being unique in each table.
- *code*: A code composed of four digits (between 1000 and 9999) defining the feature class. The first one or two digits define the layer, the last two or three digits the class inside a layer.
- *name*: Place local name, e.g., 'Post St.Martin'. If the name in OSM dataset contains obviously wrong data, such as 'fixme' or 'none', it will be empty.
- *fclass*: Place category name, e.g., 'library' and 'post_office'. This does not add any extra information that is not already in the *code* field, but it is better readable.
- *coordinates*: Place coordinates including *latitudes* and *longitudes*.
- *osm_id*: OSM Id taken from the Id of this feature (node_id, way_id, or relation_id) in the OSM database. This is a key variable in the OSM dataset with the PBF format, but not used in this thesis.

In addition, the *road* places have additional variables as below, where the *bridge layer* and *tunnel* are also the variables of *railway* places.

- *bridge*: It represents if a road is on a bridge. The 'T' value means true, 'F' value means false.
- *layer*: It represents the relative layering of a road, valuing the integers from -5 to 5.
- *tunnel*: It represents if a road is in a tunnel. The 'T' value means true, 'F' value means false.
- *maxspeed*: It represents the max allowed speed (km/h) on a road.
- *oneway*: It represents if a road is a oneway road. The 'F' value means that only driving in direction of the line string is allowed; the 'T' value means that only the opposite direction is allowed; the 'B' value (default value) means that both directions are ok.
- *ref*: It represents the reference number of this road, valuing such as 'A 5', 'L 605'.

Furthermore, the *place* places have the variable of *population*, which shows the number of people living in a place. For example, the 'Caldonazzo' (village) place's population is '2766'. The *building* places have the variable of *type*, which shows the type of a building if specified in OSM, otherwise empty. The *waterway* places have the variable of *width*, which shows the width of a waterway in meters.