



H2020 - INDUSTRIAL LEADERSHIP - Leadership in enabling and industrial technologies - Information and Communication Technologies (ICT)

ICT-11-2017 Collective Awareness Platforms for Sustainability and Social Innovation – Innovation Action (IA)



CHILD
RESCUE

Collective Awareness Platform for Missing Children Investigation and Rescue

D2.4 Profiling, Analytics and Privacy Methodological Foundations, Release I

Workpackage: WP2 – Grassroot Collective Intelligence in the Missing Children Investigation

Authors: S5, FRA-UAS, UBITECH, NTUA

Status: Final

Date: 29/12/2018

Version: 1.00

Classification: Public

Disclaimer:











The ChildRescue project is co-funded by the Horizon 2020 Programme of the European Union. This document reflects only authors' views. The EC is not liable for any use that may be done of the information contained therein.

ChildRescue Project Profile

Grant Agreement No.: 780938

Acronym:	ChildRescue
Title:	Collective Awareness Platform for Missing Children Investigation and Rescue
URL:	http://www.childrescue.eu
Start Date:	01/01/2018
Duration:	36 months

Partners

	National Technical University of Athens (NTUA), Decision Support Systems Laboratory, DSSLab <u>Co-ordinator</u>	Greece
	European Federation for Missing and Sexually Exploited Children AISBL - Missing Children Europe (MCE)	Belgium
	The Smile of the Child (SoC)	Greece
	Foundation for Missing and Sexually Exploited Children – (Child Focus)	Belgium
	Hellenic Red Cross (REDCROSS)	Greece
	Frankfurt University of Applied Sciences (FRA-UAS)	Germany
	SINGULARLOGIC ANONYMI ETAIREIA PLIROFORIAKON SYSTIMATON KAI EFARMOGON PLIROFORIKIS (SLG)	Greece
	Ubitech Limited (UBITECH)	Cyprus
	MADE Group (MADE)	Greece
	SUITE5 DATA INTELLIGENCE SOLUTIONS LIMITED (S5)	Cyprus

Document History

Version	Date	Author (Partner)	Remarks
0.10	14/11/2018	Minas Pertselakis, Anastasios Tsitsanis (S5)	Table of Contents
0.20	03/12/2018	Isabelle Brantl (FRA-UAS)	Contribution in sections 2.1, 3.1 and Annex II.1 and III by FRA-UAS
0.25	10/12/2018	Ariadni Michalitsi (NTUA)	Contribution in sections 2.1.3, and 2.1.4 by NTUA
0.30	10/12/2018	George Vafeiadis, Dimitris Ntalaperas, Danae Vergeti (UBITECH)	Contribution in sections 2.3, 3.2 and Annex II.3, V, and VI by UBITECH
0.40	12/12/2018	Minas Pertselakis, Konstantinos Tsatsakis (S5)	First draft with contribution integration
0.50	21/12/2018	Minas Pertselakis, Konstantinos Tsatsakis (S5)	Final draft
0.60	27/12/2018	Danae Vergeti (UBITECH)	Internal Review
0.70	28/12/2018	Nefeli Bountouni (SLG)	Internal Review
0.80	28/12/2018	Ariadni Michalitsi, Dimitris Varoutas (NTUA)	Quality Review
1.00	29/12/2018	Christos Ntanos (NTUA)	Final Review

Executive Summary

This report document contains the aggregated results of all Work Package 2 Tasks. In particular, it provides the definition and evaluation of the profiling, data analysing and privacy ensuring methodological aspects of ChildRescue.

An extensive in-depth analysis of the literature is presented first to describe the domains of interest, both from a theoretical and a practical point of view. The research analysis on the activity and behaviour profiling results in the identification of five distinct profiles for missing children cases. The appropriate indicators that compose each profile are extracted from literature examples, interviews with experts, as well as the recommendations from the pilot partners. The proposed profiling method is to be applied to new missing children cases, but also to past cases, which are considered to be the primary source of information for ChildRescue data analytics framework. In this framework, beside the case details and the child's characteristics, all events that occur during the investigation process, such as the feedback coming from citizens, as well as other sources of data, like social media activities and open data, are combined together into a multi-source analytical procedure. The outcome of this procedure has a threefold role: To generate an even more informative and complete profile, evaluate incoming feedback and predict Points of Interest or routes, which the missing child is going to visit or follow. For this to work, the algorithmic background on multi-source analytics is examined and several key-points are pointed out.

Since most of the data utilised are of personal nature, the report also investigates the current trends and technologies regarding privacy and data protection through the means of anonymisation and pseudo-anonymisation. The compliance of the ChildRescue data-driven methodology to the new GDPR guidelines is examined and verified through a number of suggested techniques.

Table of Contents

1	Introduction	11
1.1	Purpose & Scope	11
1.2	Structure of the Deliverable	11
1.3	Relation to other WPs & Tasks	12
2	Landscape Analysis	13
2.1	Behavioural and Activity Profiling Theories.....	13
2.1.1	System Theories	16
2.1.1.1	<i>Activity Theory.....</i>	16
2.1.1.2	<i>Collective Behaviour Theory.....</i>	17
2.1.1.3	<i>Social Network Analysis.....</i>	18
2.1.1.4	<i>Subcultural Theory.....</i>	19
2.1.2	Victimology Views on Creating Activity and Behaviour Profiles.....	20
2.1.3	Computational Methods and Tools for Activity and Behaviour Analysis.....	21
2.1.3.1	<i>Social Network Analysis & Sentiment Analysis.....</i>	22
2.1.3.2	<i>Descriptive Analytics</i>	24
2.1.3.3	<i>Machine Learning for modelling human behaviour.....</i>	26
2.1.3.4	<i>Prominent Computational tools</i>	27
2.1.4	Discussion and key-takeaways	36
2.2	Multi-Source Data Analytics	37
2.2.1	Computational Learning in Human Profiling	39
2.2.1.1	<i>Research Literature Study.....</i>	40
2.2.1.2	<i>Key points extracted from the Literature Analysis.....</i>	41
2.2.1.3	<i>Challenges and Perspectives.....</i>	44
2.2.2	Spatiotemporal Data Analysis.....	45
2.2.2.1	<i>Research Literature Study.....</i>	46
2.2.2.2	<i>Key points extracted from the Literature Analysis.....</i>	46
2.2.2.3	<i>Challenges and Perspectives.....</i>	49
2.2.3	Social Media Analytics	50
2.2.3.1	<i>Research Literature Study.....</i>	51
2.2.3.2	<i>Key points extracted from the Literature Analysis.....</i>	52
2.2.3.3	<i>Challenges and Perspectives.....</i>	55
2.2.4	Discussion and key-takeaways	56
2.3	Privacy and Anonymisation	57
2.3.1	Research Literature study	58

2.3.2	Techniques & Methodologies	59
2.3.2.1	<i>Data anonymisation models for dealing with attackers having access to multiple data sets</i>	60
2.3.2.2	<i>Data Anonymisation via aggregation</i>	64
2.3.2.3	<i>Location obfuscation</i>	65
2.3.2.4	<i>Spatial Data transformation methods</i>	66
2.3.2.5	<i>ARX</i>	67
2.3.2.6	<i>Encryption</i>	68
2.3.2.7	<i>Know Your Customer (KYC)</i>	71
2.3.3	GDPR Compliance	72
2.3.3.1	<i>Pseudonymisation and anonymisation in GDPR</i>	73
2.3.3.2	<i>Data breaches</i>	74
2.3.3.3	<i>Right of access and right to be forgotten</i>	75
2.3.3.4	<i>User Identification</i>	76
2.3.3.5	<i>Consent</i>	76
2.3.4	Discussion and key-takeaways	78

3 Methodological Approach..... 79

3.1 Setting up a Behavioural and Activity Profile 79

3.1.1	Forming Profiles based on Theories.....	80
3.1.2	Prioritising profile indicators.....	82
3.1.3	Discussion and Limitations.....	84

3.2 Ensuring Privacy and Anonymisation 85

	Preparation and Profiling	87
3.2.1	Pseudonymisation and Anonymisation	88
3.2.2	Pseudonyms	90
3.2.3	Encryption	92
3.2.4	Discussion and Limitations.....	93

3.3 Performing Predictive Analytics 94

3.3.1	Predictions based on Behavioural and Activity Profile Data.....	100
3.3.1.1	<i>Possible sources</i>	101
3.3.1.2	<i>Methods and algorithms</i>	101
3.3.1.3	<i>Profiling Algorithms for ChildRescue</i>	102
3.3.2	Evidence Analysis and Evaluation	103
3.3.2.1	<i>Possible sources</i>	104
3.3.2.2	<i>Methods and algorithms</i>	105
3.3.3	Real time Route/Destination Estimation	107

3.3.3.1	<i>Possible sources</i>	108
3.3.3.2	<i>Methods and algorithms</i>	109
3.3.3.3	<i>Prediction Algorithms for ChildRescue</i>	110
3.3.4	Discussion and Limitations.....	111
4	Conclusions & Next Steps	113
	Annex I: References	115
	Annex II: Analysis of Research Literature	127
II.1	Social Science theories applicable to cases of missing children	127
II.2	Creating Activity and Behaviour Profiles of Missing Children	129
II.3	Spatiotemporal Data Analysis	134
II.4	Social Media Analytics	140
II.5	Privacy and Anonymisation	148
	Annex III: Interviews	155
III.1	Insights from the Hellenic Amber Alert	155
III.1.1	Interviews with Hotline operator.....	155
III.1.2	Interviews with canine search unit member	159
III.2	Insights from the Hellenic Red Cross	161
III.2.1	Interviews with the Danish Red Cross	161
III.2.2	Interviews with the British Red Cross.....	163
III.3	Insights from expert interviews in Germany	164
III.3.1	State actors – German Youth Institutes.....	164
III.3.2	Specialised NGOs.....	166
III.3.3	Police (Germany/UK)	167
III.3.4	Current researchers on related areas (Paedophilia, homeless youth).....	169
	Annex IV: Past cases list of reference	171
	Annex V: Sample Consent Form	180
	Annex VI: Addressing the Ethics Requirements	184
	Annex VII: Incidental Findings Policy	195
	Annex VIII: Ethics Advisory Board report for D2.4	196
VIII.1	The ChildRescue Ethics Advisory Board D2.3 Teleconference	196
VIII.2	Meeting minutes	196
VIII.3	Conclusions	201

List of Figures

Figure 1-1 Relation of WP2 with other WPs.....	12
Figure 2-1 Plutchik's wheel of emotions [32].....	24
Figure 2-2 the 4 stages of Business Analytics.....	25
Figure 2-3 The 4 stages of Business Analytics, Gartner's Model.....	26
Figure 2-4 Main categories of Data Analytics techniques.....	38
Figure 2-5 General methodology for predictive analytics in social media	53
Figure 2-6 Example of data masking	60
Figure 2-7 Example of a k-anonymous set with k=2 and three equivalent classes	61
Figure 2-8 Example of diverse, yet semantically linked, sensitive information.....	63
Figure 2-9 Data Cube Example.....	65
Figure 2-10 Public Private Key pair generation	69
Figure 2-11 Encryption and Decryption of a message under the public key encryption scheme	69
Figure 2-12 Digital signature under the public key encryption scheme	70
Figure 2-13 Typical PKI Infrastructure	71
Figure 2-14 An example of a KYC Infrastructure Setup.	72
Figure 3-1 Different sources contributing to the indicators of the profiles	82
Figure 3-2 Architecture of the Pseudonymisation module.....	89
Figure 3-3 Pseudonym generation module architecture	92
Figure 3-4 Encryption for data fields.....	93
Figure 3-5 The proposed preliminary Data Model for ChildRescue	98
Figure 3-6 A simple data analytics flow.....	99
Figure 3-7 Behavioural Prediction process.....	100
Figure 3-8 Evidence Evaluation process	104
Figure 3-9 Route estimation process	108

List of Tables

Table 2-1 Categories and characteristics of Runaways [2]	14
Table 2-2 Categories of missing children cases and theories	15
Table 2-3 Computational tools towards Activity and Behaviour Analysis	28
Table 2-4 List of computational learning algorithms for human profiling	42
Table 2-5 Variables based on information in police files describing missing persons [42]	43
Table 2-6 Methods and Algorithms for spatiotemporal data analysis	47
Table 2-7 Most widely used data sources in spatiotemporal data analysis	48
Table 2-8 Methods and Algorithms for social media analytics	54
Table 2-9 Most widely used data types in social media analytics	55
Table 2-10 Data subjects and data recipient summary for ChildRescue	78
Table 3-1 Indicators and Theories applicable to cases	81
Table 3-2 Applicable information from FBI offender profiling for missing children cases	82
Table 3-3 Different indicators named in the interviews	83
Table 3-4 Most important profile indicators for ChildRescue	84
Table 3-5 Requirements mapping to technical requirements of the Privacy and Anonymisation Framework	86
Table 3-6 Use cases mapping to technical requirements of the Privacy and Anonymisation Framework	86
Table 3-7 ChildRescue processes mapping to technical requirements of the Privacy and Anonymisation Framework	87
Table 3-8 Pseudonymisation techniques offered by ChildRescue	90
Table 3-9 List of data fields included in the Profiling Template	95
Table 3-10 List of data fields for the Events Template	97
Table 3-11 Relation of ChildRescue investigation phases and Analytics types	99
Table 3-12 Data sources to be used for profiling	101
Table 3-13 Summary of algorithms for profile modelling and predictions	102
Table 3-14 Profiling approach	102
Table 3-15 Data sources to be used for evidence evaluation	104
Table 3-16 Algorithms related to Evaluation Steps	106
Table 3-17 Data sources that will be used as input for the recommended methods	108
Table 3-18 Methods and algorithms for route/destination estimation	109
Table 3-19 Summary of data collection and analysis limitations	112
Table Annex VI- 1 Ethics requirements list and how they will be addressed in ChildRescue	184
Table Annex VIII- 1 The ChildRescue Ethics Advisory Board (names and position)	196

Table Annex II- 1 Social Science applicable to missing children cases.....	127
Table Annex II- 2 Literature review for the creation of activity and behaviour profiles of missing children	130
Table Annex II- 3 Literature review in Spatiotemporal Data Analysis	135
Table Annex II- 4 Literature review in Social Media Analytics	141
Table Annex II- 5 Literature review in Privacy and Anonymisation.....	148
Table Annex III- 1 Evaluation of the significance of sources related to the investigation process by the hotline operator	159
Table Annex III- 2 Evaluation of the significance of sources related to the investigation process by the canine search unite member	160
Table Annex IV- 1 List of reference for past cases.....	171
Table Annex IV- 2 List of 6 full cases of missing children (2 per pilot partner).....	176

1 Introduction

1.1 Purpose & Scope

The present deliverable is released under the framework of Work Package 2 “Grassroot Collective Intelligence in the Missing Children Investigation”, and contributes the first iteration (release I) of the work conducted in Task 2.1 “Behavioural and Activity Profiling of Missing children”, Task 2.2 “Multi-source Analytics for Missing Children Investigation” and Task 2.3 “Stakeholders Privacy and End-to-end Information Pseudo-anonymisation”. In other words, the results reported in the respective deliverables D2.1, D2.2 and D2.3 are aggregated into an integrated ChildRescue methodology, which is presented in this document and will be independently assessed by the Ethics Advisory Board of ChildRescue.

That said, the purpose of the ChildRescue deliverable D2.4 “Profiling, Analytics and Privacy Methodological Foundations, Release I” is to provide the necessary theoretical foundations and state-of-the-art tools, in order to build, in accordance to the ChildRescue Description of Action, an efficient and scientifically sound methodological approach for a) timely and effectively profiling the missing children, b) analysing multi-layered data coming from multiple sources and c) applying anonymisation techniques on data, while respecting all privacy issues. Therefore, the scope of this deliverable is:

- To study in depth the underlying state-of-play concerning profiling, data anonymisation and data analytics in the context of ChildRescue.
- To identify and examine methods and algorithms for predictive analytics, aiming to enhance the profiling process, as well as to assist in real-time investigation by predicting possible locations or routes of interest.
- To elaborate on data anonymisation, pseudo-anonymisation and related techniques, which can cope with privacy issues and help ChildRescue enforce privacy and data protection, while complying to relevant European and national laws and regulations.

D2.4 will be publicly available after delivery, through the ChildRescue site and other project communication channels, as determined in the dissemination plan and the data management plan of WP5. By the end of the second year of the project [M24], the proposed methodology will be further refined and optimised, relying on the feedback from the end-users, as well as the technical specifications of the ChildRescue platform and will be delivered in D2.5 “Profiling, Analytics and Privacy Methodological Foundations, Release II”.

1.2 Structure of the Deliverable

The deliverable is structured as follows:

In Section 1, an introductory description of the document is provided, communicating its purpose, its structure and its relation to other tasks and work packages.

In Section 2, the results of the landscape analysis conducted on the fields of human profiling, multi-source analytics and privacy are presented. Key points, comparison of different methods, challenges, as well as possible applications and perspectives are discussed in this section.

Following the literature overview, Section 3 focuses on constructing the foundations of a sound methodological approach for ChildRescue. The proposed methodology consists of three major pillars

that cover the objectives of the Work Package. The key-takeaways, along with possible challenges and risks, are discussed at the end of the section.

Section 4 concludes the document by summarising the most important findings of this deliverable and sketching out the future steps.

The final pages of the deliverable contain a number of annexes, so as to provide more detailed information and relevant tables in an organised fashion without cluttering the main document.

1.3 Relation to other WPs & Tasks

As already noted, the reporting document at hand is an aggregation of the results of the first iteration of Tasks 2.1, 2.2 and 2.3. All of these tasks are directly affected by the results of Task 1.1 "User Requirements" and Task 1.3 "ChildRescue Integrated Methodology, Release I". Furthermore, Task 2.3 is closely related to Task 1.2 "Regulatory Framework for Data Protection, Privacy and Ethical Issues" since both handle privacy and ethical issues but from a different angle.

The proposed methodological approach is the first step to develop the appropriate implementation specifications required by WP3 - "ChildRescue Platform Architecture Definition and Implementation". More specifically, the theoretical approach described in D2.4 will be translated into particular components of the ChildRescue architecture of T3.1 "ChildRescue Architecture and Platform Design", which will be then implemented in T3.2 and T3.3. As a next step, the technical verification in T3.4 will provide the necessary feedback on the performance of the suggested methods and algorithms.

The overall methodology will also be tested by the pilots through the evaluation framework of WP4. After completion of the first platform implementation iteration and the pilot evaluation, the algorithms and methodology will be revised accordingly. These updates will be part of D2.5 – "Profiling, Analytics and Privacy Methodological Foundations, Release II" [M24], which concludes the iterations and covers the final methodological approach of WP2.

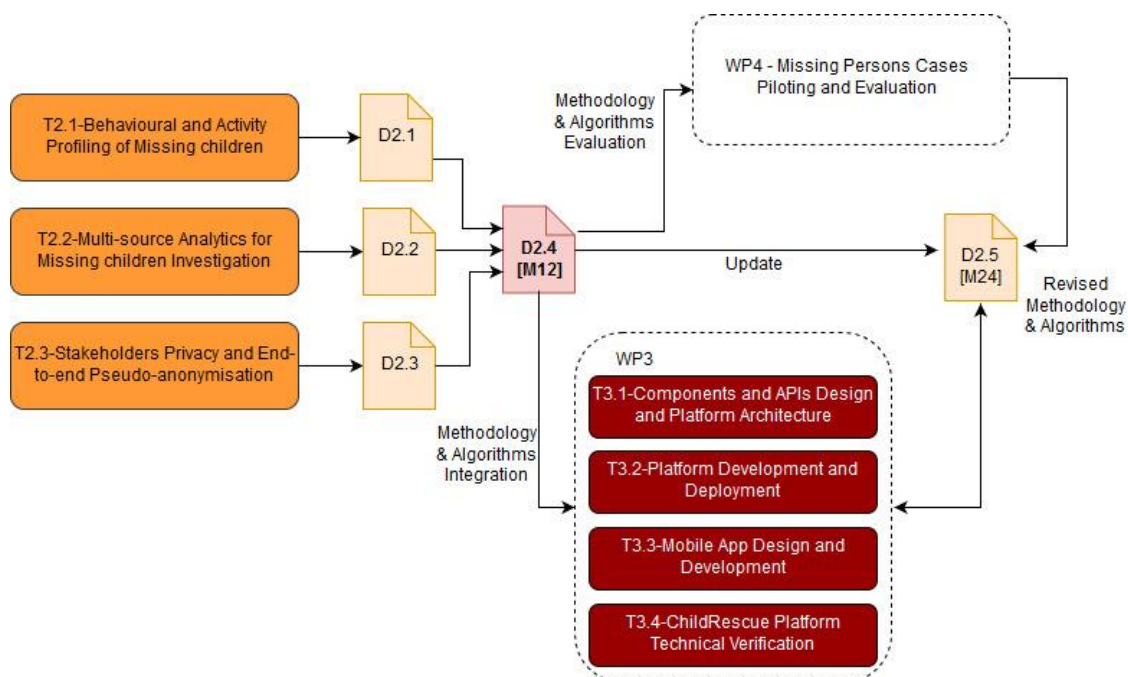


Figure 1-1 Relation of WP2 with other WPs

2 Landscape Analysis

Landscape analysis is the process of studying particular areas of research and analysing chosen key aspects in order to identify useful methods or discover important trends and evaluate key challenges related to the project at hand. In other words, this analysis offers a better understanding of a broader context in which the project is operating, thus helping one design a more appropriate and effective methodological approach, which is the purpose of this deliverable.

In the context of ChildRescue, the landscape analysis is divided into three parts, which are inextricably connected. First, a theoretical approach on behavioural and activity profiling is provided by analysing related theories from a social science point of view. This is accompanied by a, more practical, presentation of computational tools for activity and behaviour analysis. Delving more in the field of computer science, a thorough investigation on applied algorithmic methods and techniques is conducted which covers the most significant aspects of data analytics for missing persons. The last part is devoted to privacy and data protection issues, and how they can be handled with modern means of anonymisation and other techniques under the new GDPR guidelines.

2.1 Behavioural and Activity Profiling Theories

When establishing the essential information for cases of missing children in a scientific based manner to create reliable activity and behaviour profiles of children, it is vital to consider the state-of-the-art research and opinions of experts in the field while grounding these findings in appropriately chosen theories. In order to appropriately discuss the phenomenon of missing children in the face of different legal definitions in different European countries, it is vital to adhere to the differentiation of cases that was proposed by Missing Children Europe (MCE), which distinguishes between runaways, abductions by third persons, international parental abductions and missing unaccompanied migrant minors [1]. Additionally, there are the cases with an unknown outcome and categorisation, which cannot be determined, the so-called "Lost, injured or otherwise missing children" [1], who have gotten lost (e.g. little children at the seaside in summer) or hurt themselves and cannot be found immediately (e.g. accidents during sports activities, at youth camps, etc.), as well as children whose reason for disappearing has not yet been determined.

For the purpose of building an activity and behaviour profile of a child, these cases are extremely challenging as their distinguishing factor is their lack of information, which unfortunately makes a characterisation of the missing child and an approach that takes the circumstances of their disappearance into consideration impossible. Thus, the other named four categories will be at the heart of the theoretical classification in order to react appropriately to each case, but the category of lost, injured or otherwise missing children shall be considered along their side. Due to the vastly different circumstances of missing children in the different categories, the different theories shall be applied to the different categories as the varying degree of agency and behavioural choices at the disposal of the child should be reflected in the theoretical approach to their explanation. Within each of the categories, further differences might appear which can lead to a different outcome for the missing child and alter their predicted place of interest.

Runaways, for instance, were further divided into five different categories based on the circumstance of the going-away situation by Payne (1995) (cf. Table 2-1).

Table 2-1 Categories and characteristics of Runaways [2]

Category of 'Runaways'	Characteristics
Runaway	Leave willingly, spontaneous, loss of (self) control, mostly reported
Throwaway	Forced out of home by parents, leave unwillingly, mostly unreported
Pushaway	Forced out of home by circumstances, leave due to lack of alternative, sometimes unreported
Fallaway	Contact lost, possibly leave willingly or due to natural disaster/humanitarian crisis, sometimes unreported
Takeaway	Forcefully taken away, leave unwillingly, mostly reported

Runaways, who decide to disappear relatively spontaneous based on a loss of control, either by themselves or the legal guardians and a corresponding subjective feeling of being overwhelmed with immediate social pressure. Throwaways or rejected missing children, who are being forced out of their home by their parents or other legal guardians and thus do not leave by their own choice [2]. Most of those throwaway-cases will not be reported as a missing person case however, leaving the children unfortunately out of the reach of the ChildRescue project. Pushaways or people forced to go missing, in the American literature also dubbed 'push-outs' refers to cases in which the social network of the missing person forced them to leave due to unbearable circumstances. While there is some commonality with the category of throwaways, the difference is that pushaways are not necessarily asked to leave, but can be pushed out solely by the circumstances of their home, such as sexual or physical abuse [2]. Fallaways, otherwise known as people who have lost contact, are mostly adults, but can also include children in situations of natural disasters, humanitarian crises or so-called 'drifters'. They are defined as relationships that fall away due to either the constant movement of a person or other outside-factors, like a loss of contact after a disaster or death. The last category summarises cases in which the behaviour is not actually the active running away of a person, but rather missing children person cases in which the people were taken by force, hence called takeaways or people forced out of contact [2]. The last category clearly overlaps with the set of abductions by strangers or parents drawn up by Missing Children Europe (MCE), but can be a helpful distinction in regard to the apparent intent of people. In order to draft the methodology of ChildRescue and find suitable theories to accompany the research, a close examination of different reasons for going missing – and thus the categories explained above – are important as the circumstances of going missing can determine the action taken by the missing child after the initial instance and the resulting danger of being alone. Spontaneous instances of going missing, as in cases of runaways, for example, can sometimes be linked to short-term absence from the home and thus carry a lower risk than throwaway or takeaway situations. However, this is not always the case and should not be assumed due to the categorisation of the case as a runaway.

In cases of non-parental abductions, different distinctions can be made, either by motive, distinguishing between premeditated kidnappings for sexual or other types of exploitation vs. spontaneous kidnappings due to opportunity and kidnappings of foetus or babies satisfying maternal feelings, or dividing by offender-victim relationship into abductions by a family member, acquaintance or stranger [3]. It is noteworthy that the abductions by strangers are statistically least likely to appear and only accounted for 24% of the 1,214 cases analysed in the study [3]. In general, child abductions are the least likely scenario in missing children cases [4]. Another possible distinction between cases is whether the child was found dead or alive, which is most prominently being utilised in the profiling of offender types in the US [4]. Fatal outcomes can be linked to demographic characteristics of the victims in the sense of a victimological analysis. Beasley et al.'s study (2009) has shown a prevalence of fatal outcomes with older children going missing, as well as being Caucasian in the US [4]. Warren et al. (2016) additionally found a higher risk of fatal outcomes for female kidnapping victims in the US [3]. Parental abductions, if classified as a missing persons case, which depends on national legislation with Germany for instance not regarding this as a missing persons case unless the whereabouts are utterly unknown [1], also have different motivations and can be triggered by different circumstances, such as a custody lawsuit by one parent of an international couple in divorce or the desire to force a child into marriage in another country. Depending on the circumstances of the abduction, the physical and mental danger for the child can vary.

Due to the stark differences in the circumstances and consequent behaviour of missing children in those different categories, it seems vital to mirror those in the choice of different theories that might have the highest potential to explain certain forms of missing children cases. Thus, the selected theories shall be applied to the different cases and are not always appropriate for all scenarios as alluded to in D1.3 (cf. Table 2-2). Runaway shall, in this case, include all five forms of runaway behaviour classified by Payne (1995) [2].

Table 2-2 Categories of missing children cases and theories

Categories of missing children cases and theories	
<i>Missing children case type</i>	<i>Theories (examples)</i>
Runaways	Subcultural Theory, Activity Theory, Social Network Analysis
Third-person abductions	Social Network Theory, Victimology
Parental Abductions	Activity Theory, Social Network Analysis
Missing unaccompanied migrant minors	Subcultural Theory, Activity Theory, Collective Behaviour Theory
Lost, injured or otherwise missing child	Not applicable due to lack of information on the case

Focussing on an efficient and quick recovery of the child ideally prerequisites knowledge on the type of case, which can be received through interviews with parents and friends of the missing child,

which are already a part of the best practices in the field. After this first determination is made, the behaviour can be explained and predicted with a higher precision if the appropriate theories are utilised to advise the quest for the most essential information needed for those cases. The identification of indicators that can follow after a correct initial categorisation of the case, allow for a timely response and recovery of the child.

2.1.1 System Theories

System theories analyse the environment of an individual to explain their behaviour and thus identify correlations between environment and person. In missing children cases, this is especially vital when the child has not been taken by another person and thus has retained a certain degree of agency, so can be classified in Payne's concept as either a runaway, throwaway or pushaway or a case of a missing unaccompanied migrant minor. In those cases, the environment of the child that has gone missing can hold important information about the location of the child as minors sometimes join pre-existing networks of other runaway youths. In the following, different theories that follow the paradigm of system theories are discussed in regard to their applicability to ChildRescue. In cases of parental abductions or kidnappings by strangers, where the children have limited, if any, agency, the system theories don't demonstrate a high degree of explanatory power as the behaviour of the child is controlled by someone else. However, in some cases, the social network analysis can still be a useful tool, especially if the adult in control is a central person in the network or is in touch with other key persons of the child. For a comparison of all system theories described in the following, see Annex 4II.1.

2.1.1.1 Activity Theory

Activity Theory (AT) focusses the analysis of human behaviour on the correlation between individual and their natural environment. So, it "takes into account cultural factors and developmental aspects of human mental life" [5]. Activity Theory was developed in Russia in the 1920s and 30s and tried to link human consciousness and their behaviour by looking at the way the consciousness of people is shaped by what they do and in change shapes their handling of other, unrelated situations [6]. The goal of AT then is to understand the mental capabilities of actors and to analyse the cultural and technical aspects of human actions. However, this demands an understanding of the environment that influences the shaping of the conscious and thus the following, unconnected actions as man-made as the influential tools that are man-made too so that no single individual lives outside of the influence of others [6]. In the context of missing children cases, this suggests that pushaways are heavily influenced in their decision to leave by their conscious which has been shaped through their environment and runaways as well as throwaways and unaccompanied migrant minors will be influenced in their choice of (re-)location by the tools available to their conscious. One of the main motives in the decision-making process of youth who have run away from home was their limited access to service providers or other key individuals in the past [7], thus alternative tools might not be available to those adolescents. Furthermore, various studies on the personality traits of runaway youth in comparison to non-runaways were conducted and found that runaway children are more comparable to 'maladjusted' than 'normal' children, and are also non-compliant with social control instances, generally avoidant of difficult situations and typically grow up in family environments with little cohesion [8]. Thus, the consciousness which has been shaped by the family environment seems to favour building personality traits that increase the likeliness of running away. Hence, it is really not

a certain characteristic, but rather a learned way of behaving that influences that activity of the individual.

“While it is necessary to be very cautious about imputing personality types to categories of people, it may be that people with particular ways of dealing with social situations and life problems are more likely to react by running away. Feelings of lack of commitment, being out of control, being likely to respond by action rather than inaction, all seem likely possibilities. This set of characteristics will not apply to those who have been ‘takenaway’” [2]. Consequently, the environment should be regarded as a major influence on points of interest of these children and activity theory can be utilised in explaining the running away behaviour by allowing for an explanation of contradictory behaviour: “AT also tries to trace the causes for problems in an activity system by exploring the contradicting/problematic relations between the elements in an activity system and the influence of these contradictions on the results of activity (outcome)” [8].

Furthermore, in cases of parental abductions, both the parent and the (willing) child can be influenced by their perceived set of behaviour alternatives or the lack thereof. A perceived threat to the wellbeing of the child coupled with a mistrust of the state actors and the justice system to respond appropriately can be one of the causes for a parental abduction. In these cases, earlier reports of domestic violence may occur. However, it should be noted that also in cases without incidents of domestic violence or threats to the child’s wellbeing, a fear of losing custody over the child can lead to a parental abduction, which is more likely if the parent fails to recognise other tools or has lived in an environment of mistrust towards state authority.

2.1.1.2 Collective Behaviour Theory

Collective behaviour analysis focusses on the behaviour of groups of more or less organised people, as found in crowds, subcultures or mass events. However, the level of organisation can vary depending on the sort of crowd, which create their own patterns of behaviour. “Action is first; but the effect of action is to create an action pattern. This action pattern, as may be observed in the crowd, is frequently extremely fragile and ephemeral, and may exist without any clearly defined organisation. Permanence of the action pattern, however, is dependent upon the existence of structure, upon a division of labour, and upon some degree of specialisation in the individuals who compose the group” [9]. At the core of collective behaviour theory is thus not the action of a single individual, but rather a group acting as an entity, much like the Occupy Wall Street movement, which was loosely organised through social media accounts and created specific action patterns [10].

In missing children cases, collective behaviour theory is relevant when it comes to behaviour exhibited after the original instance of going missing, especially in cases where children are homeless and live in loosely organised collectives on the street, which can occasionally lead to a division of labour. However, as the formation of such a collective takes time, these cases prerequisite a longer period of absence from home. Minors living on the street usually do not fall under Payne’s category of runaways as their decision is made gradually by establishing their network or collective and are not spontaneous, but are often not reported missing when staying away permanently as the parents fear sanctions or do not care [11]. Thus, in the context of ChildRescue the collective behaviour analysis has limited relevance and can only be applied to ‘newcomers’ in those collectives who have been reported missing.

2.1.1.3 *Social Network Analysis*

"A social network can be defined as an array of social relations among social actors (such as individuals, groups, associations, institutions, nations, and even blogs), or as a set of nodes linked by a social relationship or tie [12]. A relation is symbolised as a link or flow between these units. The number of possible relations is potentially infinite and the term 'relation' can have many different meanings: acquaintance, kinship, family, friendship, commercial exchange, physical connection, presence on a web-page in the form of a link to another page, and so on. SNA [Social Network Analysis] is the systematic investigation of patterned relations among actors at multiple levels of analysis. The multi-level perspective of conceptualisation" [13]. The focus of the social network analysis is thus not on the single individual but rather on the relations to other people in their network and the resulting ties that influence the behaviour of the person more than their personal characteristics [13]. Due to the ever-changing and relational nature of social reality, decisions of individuals are influenced by a set of different factors that can be hard to locate. Going missing, which is usually defined from the standpoint of those left behind, is a good example for this, as it can be defined as not being where the social network expects one to be and to thus fail to comply with the 'normal' social reality of their network [2]. In practical terms, this means that the structure of the social network defines their reactions to a child going missing: a report will only be filed when said child is expected to come or be home and the speed in which the report is filed also depends on how explicit these expectations are. In the case of missing children, a child who is supposedly out to play with friends for a few hours will not be regarded as missing until it has failed to return after the time it was expected to show back up. Hence, being missed prerequisites functioning social relations, as the network will be the deciding factor that shapes the behavioural expectations of the children. The relations in the social network can furthermore very directly influence the child going missing, especially in cases of throwaways or pushaways, where the social network and the resulting environment of the child are the deciding factors. A social network analysis of the family structures can be fruitful in cases of missing children as there are often pre-existing relationship issues in the families that were identified as characteristic for runaways [2]. The instance of running away is then often triggered by an acute episode of familiar conflict, like a fight. Additionally, in cases of runaways, existing social networks can act as pull-factors which spontaneously sway the decision to leave the guardian's control. In contrast, children might also run towards their families instead of from them, if they are in the care of a foster home or institution and want to access their families, but are restricted from doing so [7].

In cases of unaccompanied migrant minors, social network analysis should both focus on relationships with other adolescents in transit as well as online social networks as studies have shown that new technologies, especially social media, are utilised by refugees to plan routes, seek shelters and legal or medical advice, but can also be used by different actors to track the behaviour and location of people and possibly threaten them [14]. However, even with the use of new technologies, human relations are still essential as 49% of respondents in Borkert et al.'s study (2018) relied on friends and 23% on other refugees to help search for information. Despite relying on facebook and whatsapp for information, the most important source of information on routes to travel remains people [14]. Social networks thus highly impact on the actions of missing people as they can initiate the instance of going away or also potentially prevent it. Furthermore, as especially youth connect with peers via

social media and 99% of adolescents in Frith's study have reported to use social media on a weekly basis, an analysis of their online network structures can be helpful in identifying key persons and thus determining potential points of interests for youth, who move on their own account as "[s]ocial media is now a part of the way in which young people interact with each other and build relationships" [15]. While social media data has already been proven helpful in assessing risky health behaviour [16], it could certainly also be adopted for the analysis of the actions of missing persons.

In cases of parental abductions, the social network of both child and abducting parent can be enlightening in the creation of an activity and behaviour profile, as it will shed light on potential key persons, who will be approached for help and can thus be regarded as places of interest for the missing child. Additionally, in cases of third-person abductions, a social network analysis of both offline and online contacts can be vital in cases where the child has been 'groomed' by the perpetrator prior to the abduction as they will be a vital contact person for the child.

2.1.1.4 Subcultural Theory

Subcultural theory, much like collective behaviour and network analysis, focusses on the individual only in the context of their system and aims to explain the behaviour of people by considering the specific set of norms and rules that apply in their subculture [17]. "Theoretical explanations of subcultures contain two main elements or ingredients. The first is an attempt to demonstrate that the distinctive content of the subculture answers to the distinctive needs or interests of its members. It involves identifying those needs or interests and showing how the subculture is peculiarly fitted to satisfy them.... The other ingredient of explanation lies in the conditions of interaction – that is, in the availability and access to one another of people who have in common similar life conditions and, hence, similar needs or interests, so that they may freely associate with one another and elaborate common cultures. This availability is not just a matter of propinquity; it is also subject to social control" [18]. However, it should be noted that the membership to a subculture is not exclusive, but that individuals rather belong to a series of different social groups, which compete for the degree of participation of the individual. The degree to which the subculture influences the behaviour of the individual depends on the amount of dependence from that subculture [18]. Especially in families with low social cohesion, alternative subcultures such as peer groups that are competing for the participation of the individual can thus have a stronger effect on the behaviour of the child as there are little alternatives and the opportunity to gain resources such as affection, welfare and status will rather be found in other groups.

In relation to missing children cases, subcultural theory is a useful tool in addition to the social network analysis in understanding decision-making processes that lead to runaway behaviour. As discussed above, in some cases the decision to leave home is not made spontaneously, but rather develops gradually through contact with the subculture of street youth [11], which leads to the development of functioning networks in that subculture and an increasing subscription to their deviant values and norms, which makes leaving home seem like a desirable alternative. Subcultural theory can thus be used to explain runaway behaviour, but also be applied to missing unaccompanied migrant minors as they usually heavily rely on the experiences and the norms of peers and gate keepers within the subculture unaccompanied minor migrants. Subcultural theory is less applicable in situations where the children did not choose to leave or where to go after getting thrown out.

2.1.2 Victimology Views on Creating Activity and Behaviour Profiles

Victimology as a discipline looks at characteristics of victims and should be viewed as a dynamic approach that connects social sciences with legal categories [19]. Victimology focuses on the frequency and reasons of victimisation of people, especially in regard to the offender-victim dynamic and socio-political reactions to victimisations. In regard to missing children, the issue is twofold: On the one hand there are clear-cut cases of victimisation through stranger kidnappings against the will of the children. While the system theories mostly serve to explain instances of missing children in which the child has a high degree of agency albeit limited behavioural choices, the cases of low agency, especially third-person abductions, cannot fully be explained by solely focussing on the social system of the victim as most of the activity is influenced by the perpetrator rather than the victim. However, focussing on factors of vulnerability for becoming a victim can shed light on the probability that an incident of a missing child falls under the category of third-person abduction. While there are certain risk factors that can suggest a third-person abduction as the most likely scenario, such as the young age of the child and a healthy family dynamic, other possibilities should also be entertained.

Furthermore, there are numerous applications of victimisation that happen after the instance of going missing, when children were not actually taken, but have to navigate in the social situation of having left the home. "There is an undeniable connection between missing children and the issue of child exploitation. The threat of exposure to high-risk activities increases significantly the longer a child is missing. Children who go missing, run away, or are abducted are often exposed to or suffer from:

- Sexual exploitation, trafficking in persons, and prostitution;
- Illegal/unsafe employment;
- Involvement in criminal activity, both as a victim and as a perpetrator;
- Deterioration of physical and emotional health;
- Lack of education;
- Substance misuse;
- Risk of physical and sexual assault; or
- In some circumstances, death" [20](ICMEC 2016, IV).

While a lack of education and a deterioration of health are not instances of classical victimisation, the risk for young people who are missing to be criminally victimised is increased as the systems of guardianship are diminished or possibly even non-existent. Additionally, if children are being exploited as labourers or in prostitution, this might offer crucial insights into their location or places of interest and thus increase the likelihood of finding the missing children. Additional risk factors that make children vulnerable for exploitation are a lack of oversight through the guardian, either at home or in care, which makes throwaways, pushaways and unaccompanied migrant minors especially vulnerable.

Although it is highly desirable to locate the missing child before they are being victimised, being familiar with patterns of exploitation at the home location of the child might thus be of advantage too. While missing children are generally at a higher risk of further victimisation, this risk is increased in situations of long exposure to being left to fend for themselves, such as runaway, throwaway or pushaway cases as well as third-person abductions for the purpose of sexual or physical abuse. Especially cases that fall under the broader category of runaways are at high risk of further victimisation as they often live on the streets. "Homeless adolescents in all subgroups, are uniquely subject to victimisation as a result of the poverty, violence, and drug abuse they encounter on the

streets [21]. They experience numerous deprivations that may affect them for a lifetime. Many lose touch with any sort of environment that once offered a sense of security, identity, and belonging. Leaving home precociously may weaken primary supportive ties to caretakers. Many decide to reject the values and norms of mainstream society, which may in turn be detrimental to their mental health [21]. Thus, analysing the vulnerability of the missing child, especially in cases of runaways and unaccompanied minor migrants, should focus both on the situation in their home or shelter before going missing and the potential risks of victimisation they are facing after going missing as both may be critical for the child. "There are also considerable risks associated with young people going missing. Even if a 'missing persons incident' is of short duration, young people may still be placed at risk of abuse or exploitation in order to obtain care or somewhere to live; they may become the victims of crime; or they may take to unconventional lifestyles, including crime, drug and alcohol abuse, and prostitution, in order to survive. These possible risks must be balanced against the possible risks in the home or residential care situation which have not been recognised and from which the young person may be running away. Also, it is necessary to consider whether these young persons might be exercising their right to become independent of the adult care that they have been receiving.... Careful analysis of feelings and attitudes in such situations is therefore required" [2].

Furthermore, in cases of kidnappings by strangers, the analysis of offender-victim dynamics such as online or offline grooming of young girls by adults can offer insights into potential locations of the child as well. A study by Warren et al. found only 24% of the analysed 1214 cases of juvenile kidnappings to involve an abduction by a stranger, with the majority of offenders being either family members or acquaintances [22]. Thus, an analysis of the victim and their vulnerability for an abduction can be helpful in creating an activity profile of a potential abduction victim. However, as of yet, there is a serious lack of victimological analysis of abduction cases in the research literature, with the published studies focussing mostly on the profiling of the perpetrator rather than characteristics of the victims. This can possibly be related to data protection issues surrounding the sociodemographic analysis of data relating to underage victims as well as concerns for their psychological wellbeing by being subjected to qualitative research that could be re-traumatising.

For the purpose of the creation of an activity and behavioural profile of the missing child in the context of ChildRescue, an analysis of factors of vulnerability for third-person abductions as well as exploitation in a high-risk environment such as homelessness can be achieved by talking to peers and family members as well as analysing the public information on their social media profiles that might showcase an increased interest in the child's activities by specific adults. In order to fully analyse these dynamics, ChildRescue will implement both the practices of interviews with peers and families that have been identified as an efficient tool in pinpointing potential locations of the child by experts in the field as well as analyses of online data. Victimology thus serves mostly to pinpoint potential factors of vulnerability in cases of missing children which can be vital to expedite the search.

2.1.3 Computational Methods and Tools for Activity and Behaviour Analysis

Following a state-of-the-art research and presentation of the most prominent theories and respective findings to create reliable activity and behaviour profiles of children from a sociological perspective, this section aims to analyse and present how activity and behaviour analysis can be implemented from a technological viewpoint, using methods, scientific fields and respective computational tools that can support such analysis from different angles. These have been analysed into three main

categories/theories that cover the needs of activity and behaviour analysis from a technological viewpoint, in the context of the ChildRescue project.

In that view, Social Network and Sentiment Analysis are analysed together due to their interdependent nature, aiming at further enriching missing child's profile with new, previously unknown insights, either these concern his activities, his mood, his interests, or his social connections. Social Network Analysis is distinguished here from the corresponding one in section 2.1.1.3 of this deliverable, where the term was established from a sociological perspective. Descriptive Analytics, in the context of ChildRescue, offer the scientific and technical base to analyse the existing information about a missing child and help explore and discover behavioural patterns. Finally, Machine Learning techniques are also considered quite relevant with ChildRescue interests to support modelling human behaviour and predicting missing children's potential locations as well as recognising patterns in missing children's behaviour. The theoretical background of each of these three areas is analysed along the following lines, followed by an aggregation of several computational tools, that fall into one or more of these three categories.

2.1.3.1 Social Network Analysis & Sentiment Analysis

Social Network Analysis takes place from information drawn from observation of both online - and physical - network human behaviour [23]. Social Network Analysis software performs both quantitative and/or qualitative analysis of social networks. The most common network graph characteristics that Social Network Analysis tools compute are [24]:

- Connectivity (e.g. shortest path length, diameter and density)
- Clustering (e.g. local clustering, global clustering)
- Centrality (e.g. degree, closeness, betweenness, eigenvector centrality)

Generally, Social Network Analysis software consists of either packages based on graphical user interfaces (GUIs), or packages built for scripting/programming languages¹. While the GUI packages are easier to learn and more user-friendly, scripting tools are generally more powerful and customisable. The most popular software tools used for Social Network Analysis and/or Sentiment Analysis are presented in 2.1.3.4.

The optimal decision on which Social Network Analysis tool should be used, depends on several variables, e.g. the development skills, the network's size, the focus on visualisation or on computing network's metrics, etc. Although, in the general case, most of the packages are more difficult to learn than the privately maintained software, some of these open source packages are growing much faster in terms of functionality and features. Furthermore, for truly large networks (more than 1 million nodes), Pajek (from the GUI packages) and GraphX (from the scripting tools) are more appropriate to use.

As it is obvious, in Social Network Analysis the relationships are important [25], so the relationships might be the reason of a specific behaviour as well the result of it [26]. Thus, Social Network Analysis is extensively used in criminology. Social Network Analysis, Geographical Information Systems, Data

¹https://ipfs.io/ipfs/QmXoyvizjW3WknFiJnKLwHCnL72vedxjQkDDP1mXWo6uco/wiki/Social_network_analysis_software.html

Mining technologies are used for clustering crimes, finding links between crime and profiling offenders, identifying criminal networks, matching crimes and past cases, generating suspects, and predicting criminal activity [27]. It is also used to collect and structure information on criminal networks and analyse complex relationships involving criminal networks [28].

Of course, there also exists another type of social networks, the online ones (emerging from Facebook, Twitter, Instagram, etc.), where a large amount of data is created every second. Users are free to express their interests and opinions whenever and in any way they like. Through Social Network Analysis, recommendation and prediction of user behaviour is possible [29]. Some indicative metrics and useful information are²: friends (and their activities), likes, posts (and dates of posts), followers, actions on page, events, videos, pictures, stories, groups, use of emojis, messages. The data that are available in the web are from semi-structured to unstructured, so various approaches [30] exist such as clustering, face detection, user activities, content analysis and behavioural analysis to profile user(s) in online social networks. As Vasanthakumar et al. claim, "Pattern recognition in behavioural analysis, type and number of activities in analysing the user activities are essential key terms in profiling the user(s) in any online social network" [30].

Sentiment Analysis encapsulates the identification and extraction of subjective information and user-generated content from multiple sources to determine the emotions that are typically evoked to stakeholders and to elicit potentially hidden information about their profile. Sentiment Analysis focuses on emotion recognition [31], where emotions can be analysed and categorised in many different ways, like the ones which Plutchik created in the "wheel of emotions" [32] shown in Figure 2-1. With the rise of social media such as blogs and social networks, interest in Sentiment Analysis has been greatly increased. Online opinion has turned into a kind of virtual currency [33]. Reviews, ratings, recommendations, and other forms of online expression are very useful for businesses to understand and cover the customers' needs.

More detailed information on related methods and algorithms for the analysis of social media, including online social networks and sentiment analysis, is presented in section 2.2.3.

² <https://sproutsocial.com/insights/social-media-metrics-that-matter/>

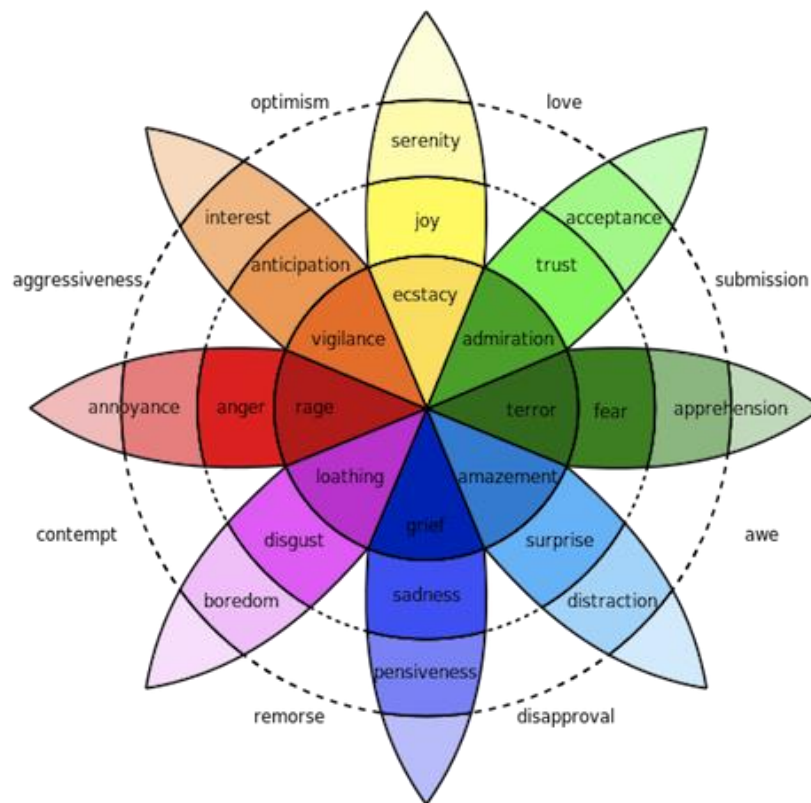


Figure 2-1 Plutchik's wheel of emotions [32]

2.1.3.2 Descriptive Analytics

In today's mainly data-driven organisations, the analysis of the data is crucial to their viability and can be of assistance in achieving their strategic goals.

Descriptive analytics, in particular, are built on the idea that historical data if interpreted and presented in a tangible and easy to comprehend manner, they can be taken into consideration when future actions are estimated³, enabling new business and strategic opportunities. Descriptive analytics are employed in all the aspects of management reporting. Examples of descriptive analytics are the reports produced from companies presenting an overview of the organisation's operations, sales, financials, customers or stakeholders⁴. Descriptive analytics is the way of incorporating lessons learnt, so it may be observed how past actions may affect future outcomes⁵. They are the conventional form of Business Intelligence and data analysis, and they provide a summary view of facts and figures in an understandable format⁶. The two techniques employed in descriptive analytics are data mining and data aggregation.

Of course, the usage of descriptive analytics is not a panacea and it comes together with some issues that need to be taken into consideration, while employing them. As the environment that

³ <https://www.cornerstoneondemand.com/glossary/descriptive-analytics>

⁴ https://www.investopedia.com/terms/d/descriptive_statistics.asp

⁵ <https://www.cornerstoneondemand.com/glossary/descriptive-analytics>

⁶ <https://halobi.com/blog/descriptive-predictive-and-prescriptive-analytics-explained/>

organisations act within becomes increasingly complex and competitive, making decisions only based on the insights of one manager or an individual decision maker is not sufficient [34]. The cultivation of a culture that enables taking into consideration the analysis of the data is required. Thus, the incorporation of descriptive analytics within the decision-making process entails all the challenges that are associated with changes in the organisation's culture. What is more, descriptive analytics mainly depend on the existence of historical data and as a result they cannot be significantly useful in new initiatives and innovations [34]. Descriptive analytics have also been accused of "killing creativity" and managers are required to find the balance between data-driven decisions and creativity or innovation [35].

Furthermore, as it is mentioned above, descriptive analytics fall under the category of Business analytics. Business analytics is a term that can be defined as "a set of all the skills, technologies, applications and practices required for continuous iterative exploration and investigation of past business performance to gain insight and drive business planning"⁷. Business analytics comprise of: Descriptive analytics, Diagnostic analytics, Predictive analytics, and Prescriptive analytics, as presented in the figure below:

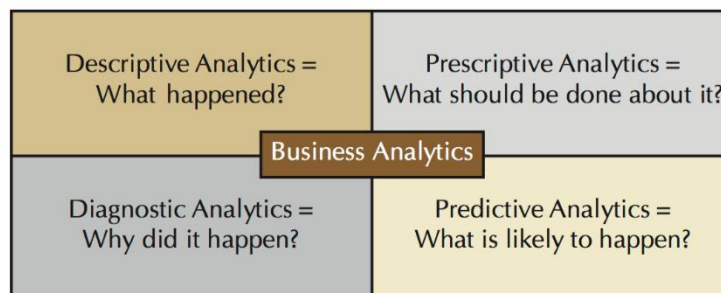


Figure 2-2 the 4 stages of Business Analytics

Descriptive analytics, as already explained, unravel what happened. Diagnostic analytics answer to the question of why something happened, with a focus of finding the roots of what caused a situation. Predictive analytics are employed to seek future actions and predict the potential outcomes of these actions. Prescriptive analytics have the aspect of finding the optimal course of future action, so that the objectives of an organisation can be achieved in the best way possible. In this step, usually decision analysis tools are also employed. In the figure below an interpretation⁸ of how the business analytics cycle operates is demonstrated:

⁷ <https://www.slideshare.net/LightshipPartners/next-generation-business-analytics-presentation>

⁸ https://www.gartner.com/technology/why_gartner.jsp

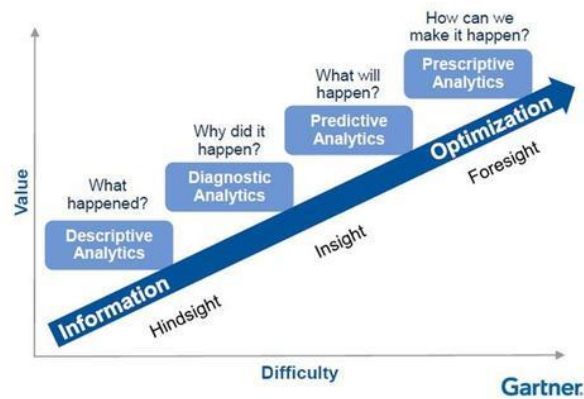


Figure 2-3 The 4 stages of Business Analytics, Gartner's Model.

2.1.3.3 Machine Learning for modelling human behaviour

The term Machine Learning (ML) was coined by Arthur Samuel in 1959 [36]. His definition about Machine Learning is that it is a "field of study that gives computers the ability to learn without being explicitly programmed"⁹. A more recent definition of Machine Learning given by Tom Mitchell [37] is the following: "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E".

Machine Learning is a subset of Artificial Intelligence (AI) which is an umbrella term for any computer program that does something smart¹⁰. Generally, machines learn from and make predictions on data. Machine Learning has endless applications¹¹ e.g. self-driving cars, virtual personal assistants (Siri, Alexa), social media services (people one may know, face recognition), email spam and malware filtering, product recommendations, etc. In the deliverable's context, a review was made in the literature on Machine Learning methods about predicting human behaviour and personality traits. There are a lot of references in the literature about Machine Learning methods and its applications about modelling human behaviour. In this chapter some characteristic examples are described.

In psychology, there has been a great progress in tools that can predict personality traits using digital footprints such as Facebook, Twitter, Instagram. Facebook allows researchers to record information [38] about users' demographic profiles (e.g., profile picture, age, gender, relationship status, place of origin, work, and education history), user-generated content (e.g., status updates, photos, videos), social network structure (e.g., list of friends and followers), and preferences and activities (e.g., group memberships, attended events). Moreover, user-generated text from messages, posts, or status updates can be further processed. Several studies have used Machine Learning to predict various personality characteristics. Most of them focus on the prediction of the "Big Five" personality traits of neuroticism, extraversion, openness to experience, agreeableness, and conscientiousness.

⁹https://www.ibm.com/developerworks/community/blogs/jfp/entry/What_Is_Machine_Learning?lang=en

¹⁰ <https://skymind.ai/wiki/ai-vs-machine-learning-vs-deep-learning>

¹¹ <https://medium.com/app-affairs/9-applications-of-machine-learning-from-day-to-day-life-112a47a429d0>

Until today, Machine Learning approaches to personality assessment have focused on the relationships between social media and other digital records with established personality measures. For example, there have been some studies where users had also completed a Big Five self-report questionnaire. These studies showed that some variables (e.g. Facebook number of friends and favourite books, Twitter words per tweet and number of hashtags) are correlated with at least one of the Big Five traits, which then are used to predict users' personality traits [38].

Further personal attributes that can be extracted from the Facebook likes (which is a mechanism to express positive association with online content, such as photos, friends, status updates, Facebook pages of products, sports, musicians, books, restaurants) are the sexual orientation, ethnicity, religious and political views, personality traits, intelligence, satisfaction with life, use of addictive substances, parental separation, age, gender, relationship status, size and density of the friendship network [39]. The analysis conducted by Hartford et al. (2016) used data (Facebook likes, detailed demographic profiles, and the results of several psychometric tests) based on a sample of 58.000 volunteers obtained through the myPersonality¹² Facebook application. The proposed model used dimensionality reduction for pre-processing the likes data, which then entered into logistic/linear regression to predict individual psychodemographic profiles. It achieved very high scores in predicting various personality traits.

Finally, predicting human behaviour in strategic settings using deep learning is another example. Deep learning (DL) is a subset of Machine Learning and it functions in a similar way¹³. If Machine Learning algorithms return a wrong prediction, then engineers need to make adjustments, but with Deep Learning algorithms they are able to decide on their own if their prediction is correct or not¹³. An example of a Deep Learning application is the prediction of the actions of human players in Go [40]. The approach of [41] evaluates Go board positions and 'policy networks' to select next moves. Deep neural networks are trained by a combination of human expert games (supervised learning), and games of self-play (reinforcement learning).

It needs to be noted though that, although machine learning has the potential to generate significant advantages above traditional assessment tools, machine learning personality assessment models are all initially validated on self-report questionnaires [38], so this is an important issue to be taken into account.

2.1.3.4 Prominent Computational tools

After exploring in a high-level the theoretical background of the three identified categories/theories, some of the most prominent, established computational tools, either proprietary or free, that are employed to enable Social Network Analysis, Sentiment Analysis, Descriptive Analytics or Machine Learning are given in the table below, along with a short description of how they can be used, as well as the category they fall into and their type. The reason for their presentation is to be used as inspiration for development purposes on what concerns activity and behaviour analysis in ChildRescue. Some of these tools may be decided to be used directly in ChildRescue.

¹² <https://sites.google.com/michalkosinski.com/mypersonality>

¹³ <https://www.zendesk.com/blog/machine-learning-and-deep-learning/>

Table 2-3 Computational tools towards Activity and Behaviour Analysis

Tool	Description	Link	Function	Type
Gephi	Gephi is an open-source network analysis and visualisation software package written in Java on the NetBeans platform. Gephi is a visualisation and exploration software for all kinds of graphs and networks.	https://gephi.org	Social Network Analysis	Application
Pajek	Pajek is a program, for Windows, for analysis and visualisation of large networks having some ten- or hundred- of thousands of vertices.	http://mrvar.fdv.uni-lj.si/pajek/	Social Network Analysis	Application
UCINet	UCINet is a software package for the analysis of social network data. It comes with the NetDraw network visualisation tool.	https://sites.google.com/site/ucinetsoftware/home	Social Network Analysis	Application
GUESS	GUESS is an exploratory data analysis and visualisation tool for graphs and networks. The system contains a domain-specific embedded language called Gython which supports the operators and syntactic sugar necessary for working on graph structures in an intuitive manner. GUESS also offers a visualisation front end that supports the export of static images and dynamic movies.	http://graphexploration.cond.org/	Social Network Analysis	Application
ORA-LITE	ORA-LITE is a dynamic meta-network assessment and analysis tool developed by CASOS at Carnegie Mellon. It contains hundreds of social networks, dynamic network metrics, trail metrics, procedures for grouping nodes, identifying local patterns, comparing and contrasting	http://www.casos.cs.cmu.edu/projects/ora/	Social Network Analysis	Application

	networks, groups, and individuals from a dynamic meta-network perspective.			
Cytoscape	Cytoscape is an open source bioinformatics software platform for visualising molecular interaction networks and integrating with gene expression profiles and other state data. Additional features are available as plugins.	http://manual.cytoscape.org/en/stable/Network_Analyzer.html	Social Network Analysis	Application
SocNetV	Social Network Visualizer (SocNetV) is a cross-platform, user-friendly free software application for social network analysis and visualisation.	http://socnetv.org/	Social Network Analysis	Application
Meerkat	Meerkat is an automated Social Network Analysis (SNA) tool used to analyse, visualise and interpret large or complex networks of information, allowing users to examine patterns and investigate relational dynamics.	https://www.amii.ca/meerkat/	Social Network Analysis	Application
muxViz	The Multilayer Analysis and Visualization Platform, MuxViz is a framework for the multilayer analysis and visualisation of networks. It allows an interactive visualisation and exploration of multilayer networks, i.e., graphs where nodes exhibit multiple relationships simultaneously.	http://muxviz.net/	Social Network Analysis	Framework
Netminer	NetMiner is a premium software tool for Exploratory Analysis and Visualisation of Network Data. NetMiner allows you to explore your network data visually and interactively and helps you to detect underlying patterns and structures of the network.	www.netminer.com	Social Network Analysis	Application & Framework
GraphX	GraphX is the new (alpha) Spark API for graphs (e.g., Web-	https://spark.apache.org/graph	Social Network	Library

	Graphs and Social Networks) and graph-parallel computation (e.g., PageRank and Collaborative Filtering).	x/	Analysis	
Oracle PGX	PGX is a toolkit for graph analysis - both running algorithms such as PageRank against graphs, and performing SQL-like pattern-matching against graphs, using the results of algorithmic analysis. Algorithms are parallelised for extreme performance. The PGX toolkit includes both a single-node in-memory engine, and a distributed engine for extremely large graphs.	https://www.oracle.com/technetwork/oracle-labs/parallel-graph-analytix/overview/index.html	Social Network Analysis	Framework
STATNET	Statnet is a suite of software packages for network analysis that implement recent advances in the statistical modelling of networks. The analytic framework is based on Exponential family Random Graph Models (ergm). Statnet provides a comprehensive framework for ergm-based network modelling, including tools for model estimation, model evaluation, model-based network simulation, and network visualisation. This broad functionality is powered by a central Markov chain Monte Carlo (MCMC) algorithm.	http://www.statnet.org/	Social Network Analysis	Framework
IGRAPH	IGRAPH is a collection of network analysis tools with the emphasis on efficiency, portability and ease of use. igraph is open source and free. igraph can be programmed in R, Python and C/C++.	http://igraph.org/	Social Network Analysis	Library
NetworkX	NetworkX is a Python package for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks.	https://networkx.github.io	Social Network Analysis	Library

JUNG	JUNG — the Java Universal Network/Graph Framework—is a software library that provides a common and extendible language for the modelling, analysis, and visualisation of data that can be represented as a graph or network. It is written in Java, which allows JUNG-based applications to make use of the extensive built-in capabilities of the Java API, as well as those of other existing third-party Java libraries.	http://jung.sourceforge.net/	Social Network Analysis	Library
sigma.js	Sigma is a JavaScript library dedicated to graph drawing. It simplifies the procedure to publish networks on Web pages and allows developers to integrate network exploration in rich Web applications.	http://sigmajs.org/	Social Network Analysis	Library
SAS Text Analytics (Text Miner)	Analyses text data from the web, comment fields, books and other text sources	http://www.sas.com/en_us/software/analytics/text-miner.html	Text Analysis	Application
Lexalytics Semantria	Applies text and sentiment analysis to tweets, Facebook posts, surveys, reviews or enterprise content	https://www.lexalytics.com/	Text Analysis, Sentiment Analysis	Framework
Lexalytics Saliency Engine	Is an on premise, multi-lingual text analysis engine	http://www.lexalytics.com/technical-info/saliency-engine	Text Analysis	Framework
Provalis Research (WORDSTA)	Is a flexible and easy-to-use text analysis software – whether text mining tools are needed for fast extraction of themes and trends, or careful and precise measurement with state-of-the-art quantitative content analysis tools	https://provalisresearch.com/products/content-analysis-software/	Text Analysis	Application

T)				
Pingar	It is able to point the trends, topics and issues exposed in documents, posts, articles and emails	http://pingar.com/content-intelligence-2/	Text Analysis	Application
RapidMiner Studio	It is an agile platform for predictive business analytics.	https://rapidminer.com/products/studio/	Predictive Analytics	Application
Text2data	Conducts in-depth analysis of business unstructured data, and trend detection in social media data	http://text2data.org/	Text Analysis, Sentiment Analysis	Application
KH Coder	Is free software for content analysis, text mining or corpus linguistics	http://kncoder.net/en/	Text Analysis	Application
GATE (General Architecture for Text Engineering)	Is a Java suite of tools used for many natural language processing tasks, including information extraction in many languages	https://gate.ac.uk/family	Text Analysis, Natural Language Processing	Framework
Rtm (Text Mining Package)	Offers functionality for managing text documents, abstracts the process of document manipulation and eases the usage of heterogeneous text formats	http://tm.r-forge.r-project.org/	Text Analysis	Library
OpenNLP	Is a machine learning based toolkit for the processing of natural language text	https://opennlp.apache.org/	Text Mining, Natural Language	Framework

			Processing	
Orange Canvas	Is an open source machine learning and data mining software	http://orange.biolab.si/	Data Mining	Application
LingPipe	Is a tool kit for processing text using computational linguistics	http://alias-i.com/lingpipe/	Text Mining	Framework
Apache UIMA	Enables applications to be decomposed into components	https://uima.apache.org/	Text Analysis	Framework
NLTK	Is a platform for building Python programs to work with human language data	http://www.nltk.org	Natural Language Processing	Framework
Scikit	Has simple and efficient tools for data mining and data analysis	http://scikit-learn.org/stable	Data Mining	Library
Scrapy	Is an application framework for crawling web sites and extracting structured data	http://scrapy.org	Data Extraction	Framework
Weka	Is a collection of machine learning algorithms for data mining tasks	http://www.cs.waikato.ac.nz/ml/weka	Data Mining	Library
CoreNLP	Is a set of natural language analysis tools which can take raw text input and give the base forms of words	https://stanfordnlp.github.io/CoreNLP/	Natural Language Processing, Text Analysis	Framework
Tableau	Tableau products query relational databases, OLAP cubes, cloud databases, and spreadsheets and then generates a number of graph types. The products can also extract data	https://www.tableau.com/	Descriptive Analytics, Data Visualisation	Application

	and store and retrieve from its in-memory data engine.			
PowerBI	Power BI is a business analytics service provided by Microsoft. It provides interactive visualisations with self-service business intelligence capabilities, where end users can create reports and dashboards by themselves, without having to depend on information technology staff or database administrators.	https://powerbi.microsoft.com/en-us/	Descriptive Analytics, Data Visualisation	Application
QlikView & QlickSense	QlikView and QlikSense are both products of the software company Qlik, serving different purposes running on the same engine. In QlikView, the user is pursuing her day-to-day tasks, analysing data with a slightly configurable dashboard, most of the data is somehow "pre-canned". On the other hand, QlikSense allows associating different data sources and fully configuring the visualisations, allowing to follow an individual discovery path through the data. In other words, QlikView is for guided analytics; Qlik Sense is for self-service visualisations.	https://www.qlik.com/us/	Descriptive Analytics, Data Visualisation	Application
SAS Visual Data Mining & Machine Learning	SAS Visual Data Mining and Machine Learning, which runs on SAS Viya combines data wrangling, exploration, feature engineering and modern statistical, data mining and machine learning techniques in a single, scalable in-memory processing environment. The solution provides a very visual and highly collaborative workspace that supports a variety of users with different skill sets.	https://www.sas.com/en_us/software/visual-data-mining-machine-learning.html	Descriptive Analytics, Machine Learning	Application
Tulip	Tulip is an information visualisation framework dedicated to	http://tulip.labri.fr/TulipDrupal/	Descriptive Analytics	Framework

	<p>the analysis and visualisation of relational data. Tulip aims to provide the developer with a complete library, supporting the design of interactive information visualisation applications for relational data that can be tailored to the problems he or she is addressing.</p>		
Apache Spark	<p>Apache Spark is a fast, in-memory data processing engine with elegant and expressive development APIs to allow data workers to efficiently execute streaming, machine learning or SQL workloads that require fast iterative access to datasets. Spark is also popular for data pipelines and machine learning model's development.</p>	<p>https://spark.apache.org</p>	<p>Data Analysis, Descriptive Analytics, Machine Learning Framework</p>
KNIME	<p>KNIME is leading open source, reporting, and integrated analytics tools that allows the analysis and modelling of data through visual programming. It integrates various components for data mining and machine learning via its modular data-pipelining concept.</p>	<p>https://www.knime.com</p>	<p>Data Mining, Descriptive Analytics Application</p>
SPLUNK	<p>Splunk is a tool that analyses and searches the machine-generated data. Splunk (the product) captures, indexes, and correlates real-time data in a searchable repository from which it can generate graphs, reports, alerts, dashboards, and visualisations.</p>	<p>https://www.splunk.com/</p>	<p>Descriptive Analytics Framework</p>

2.1.4 Discussion and key-takeaways

In order to identify key characteristics of children to build an activity and behaviour profile on, relevant though scarce literature and applicable theories should be considered alongside empirical data to gather a better understanding of the different situations and the arising varied needs of closer analysis of certain aspects in relation to the case of the missing child. Thus, both strategies of a social science analysis and technical tools should be combined to achieve the best results as outlined in the prior section.

In the context of ChildRescue, Social Network Analysis & Sentiment Analysis are considered as means supporting finding new, useful information about a missing child's personal profile. Some indicative example types of such information for the child could be the connections of the missing child, his/her social network, his/her activities, his/her hidden moods, any depressive or suicidal tendencies, influences, interests, etc. Furthermore, the comparison with past cases could reveal previously hidden insights, such as similarities, and differences that otherwise it wouldn't be feasible to find. In analogy to the use of Social Network Analysis and Sentiment Analysis tools in criminology to create a criminal's profile, ChildRescue aspires to exploit Social Network Analysis and Sentiment Analysis techniques and tools to enrich the children's profile, both in the general case and also from the viewpoint of victimology, in order to identify the profile and behaviour of a children, being a victim. Sentiment Analysis, in particular, is expected to enable also raising of a flag in case there is a "high-risk" of disappearance.

On what concerns Descriptive Analytics in the context of ChildRescue, what is needed is to select and apply the methods, approaches, and techniques of Descriptive Analytics that can be employed to identify behavioural patterns, based on past cases of missing children or unaccompanied migrant minors, as these can be analysed by the existing records kept from those past cases. The analysis and discovery of behavioural patterns can be of assistance, specifically with regards to reducing the required time needed in locating missing children. A specific area of study that can be useful to look into for insights, on how to build these behavioural patterns by employing descriptive analytics is victimology. The algorithms that aim at specifying and finding behavioural patterns in victimology fall into the broader category of descriptive analytics. One of the ways that descriptive analytics are used is to narrow down who has the potential to be a victim and who has the potential to be an offender. In a similar manner and by adjusting the algorithms some behavioural patterns can be formulated for the children that go missing and for the ones that are responsible for them going missing. Of course, descriptive analytics cannot be used without caution, specifically regarding data that involve children and migrant minors.

Finally, machine learning techniques, in ChildRescue, are expected to be leveraged significantly to enable predictions for the missing children's potential locations, exploiting each child's case data gathered and also to enable common patterns recognition based on previous cases. Although, research on the use of machine learning techniques in missing persons investigation is not a new thing [42], there is limited work in academic literature and there are still opportunities for further investigation in the context of ChildRescue. This can be seen as a promising challenge for the project, to open up new perspectives in research and science.

Currently, there exist many, different approaches for predicting users' individual preferences, attributes and behaviour to improve numerous services and products. ChildRescue will create

analogies with and leverage such approaches to develop its methodology on the profiling part and the modelling and prediction of missing children's and/or unaccompanied minors' behaviour. Additionally, if there are enough data, case comparisons can be made with older, historical data to roughly predict the next moves of a missing child and/or to raise some flags in case there is a reason to believe that an unaccompanied minor presents a "high-risk" of disappearance.

In the following section, particular practices and methods utilised in literature are introduced by a thorough research analysis on multi-source data analytics.

2.2 Multi-Source Data Analytics

In the emerging era of Big Data, the analysis of information deriving from multiple sources is motivated by the increasing volume and availability of different types of data and the desire to create data-driven computer applications that can help human beings make complex decisions. The constant sensing and capturing, storing and sharing of personal and other information – knowingly or not – through mobile phones, social media networks or the so-called Internet of Things, creates massive quantities of data produced by and about people, things, and their interactions [43]. As a consequence, the notion of data analytics is getting an increasing amount of attention, in both the academic and business community.

Data analytics is an arbitrary collection of computational methods and techniques that are employed to examine data sets and harvest knowledge and meaningful insights from them.

Data analytics is commonly viewed from four major perspectives, as already defined in section 2.1.3.2: descriptive, diagnostic, predictive and prescriptive, with the last two having attracted special interest when big data are involved [44]. From an applications point of view, however, data analytics can be divided into six major categories, each one representing a relevant set of tools and techniques (Figure 2-4).

In their recent findings, Gandomi & Haider (2015) proposed the following five categories [45]:

- **Text analytics** (or text mining) aims to extract information from documents and textual data. Social media feeds, e-mails, blogs, product reviews, news feeds, are some of the most common examples of textual datasets.
- **Audio analytics** is employed usually to process unstructured audio data, such as human speech. Currently, customer call centres and healthcare are their primary application areas.
- **Image & Video analytics** involves a variety of techniques to monitor, analyse, and extract meaningful information from images and video streams.
- **Social Media analytics** explores social networking channels and platforms and process user-generated content (e.g. text, images, videos), as well as the relationships and interactions among network entities.
- **Predictive analytics** consists of a variety of methods originating, as already mentioned, from the AI field, that seek to uncover patterns, capture feature correlations and predict future trends or actions.

But, considering the current trends in technology, a sixth category can be added, which encompasses a large set of today's tools and applications:

- **Location-based analytics** that focus on geospatial data and aim to improve location-based services.

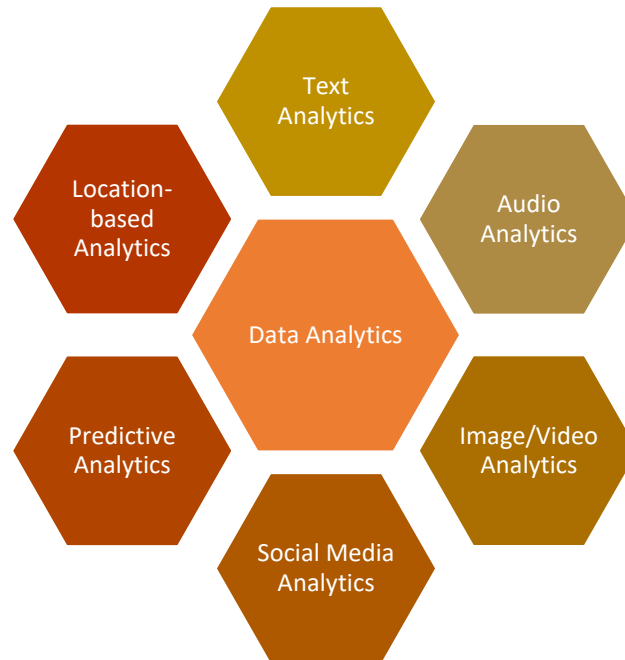


Figure 2-4 Main categories of Data Analytics techniques

It is worth noting that these categories may integrate one another or overlap with each other. For instance, the text and image analytics can play an important part in social media analytics, or location-based methods can be employed to assist in predictive analytics and vice-versa. In other words, a data analytics project may include some or all of the aforementioned categories in order to achieve the desired goals and objectives.

Today, data analytics technologies are widely utilised in commercial industries to improve business performance mostly through targeted advertisement, as well as governmental organisations, to improve on security and control. They can be applied in almost every sector: from agriculture, energy, or economics to healthcare, sports, or transportation. Depending on the particular application, the data that is analysed can consist of either historical records or new information coming from real-time processes.

One area that has been somewhat limited in its acceptance and use of these powerful new techniques is the public safety field, particularly in crime analysis and operations. Surprisingly, this comes in contradiction to what someone would expect, since police analysts, detectives, agents and other operational personnel, base their investigations on many of the principles shared by data mining and knowledge discovery. For example, the behavioural analysis of a violent crime, its characterisation, modelling and related predictions are very close to the computational methods associated with data mining and predictive analytics [46]. Nonetheless, after the events of 9/11 and even more during the last decade, predictive analytics and pattern recognition in crime-fighting have seen a strong surge of interest [47]. Considering several related studies, the data and methods

utilised by what is today called predictive policing, can be divided into three broad categories [48][49]:

- Analysis of human behaviour profiles (to identify groups or individuals at risk of offending or become victims in the future)
- Analysis of time and space (to forecast places and times with an increased risk of crime)
- Analysis of social networks (to find patterns in network relationships and activities)

Although the cases of missing children or unaccompanied migrant minors, which drive the cause of ChildRescue, are not necessarily related to a crime, the sources and the types of data involved, as well as the methods employed, can be quite similar. The analysis of human behaviour and relationships, along with spatiotemporal information, play a pivotal part in the investigation process for both a criminal activity and a missing child case.

Consequently, the main focus of our research study will be directed towards the most significant advancements in human behavioural profiling based on computational learning. In addition, we will explore a wide spectrum of algorithms for spatiotemporal data processing, as well as the most significant methods and techniques for social media analytics. As already mentioned, methods that deal with human behavioural models and patterns may overlap in several cases or be complementary to each another. This is also the case for the three selected areas of research interest. In closing, challenges and perspectives will be discussed for each of these domains, and the main take-aways to be considered by the ChildRescue methodology will be outlined.

2.2.1 Computational Learning in Human Profiling

The computational learning theory is a branch of Artificial Intelligence that studies and explores the ability of machines (i.e. computers) to learn from data. The idea of applying learning to computing devices may go as far back as the Turing machine, but the seminal paper that is usually referenced as the origin of the theory is accounted to L. G. Valiant back in 1984 [50]. In his scientific work, Valiant proposed an attractive general model to study the computational, statistical and other aspects of learning. From then on, an explosive growth of the field commenced with numerous methods, theories and algorithms devoted to computational learning [51]. Today, the applied theories of computational learning have become quite popular by the name "Machine Learning" (for the purposes of this deliverable both notions will mean the same thing) and are employed in a wide range of computing tasks, from email filtering and fraud detection to speech and image processing [52].

Computational learning is also encountered in the analysis of human behaviour and activity. Recent advancements in computer vision have enabled video processing for behaviour recognition, e.g. for surveillance purposes [53], while ambient intelligence and Active and Assistive Living (AAL) applications can automatically classify human physical activities based on wearable (or other) sensor measurements [54]. Human-computer interaction [55], web usage and social networks [56], recommendation systems [57], are some other research domains that human behavioural patterns are extracted from and exploited so that modern applications can offer better and more personalised services.

Modelling and predicting behaviour through AI techniques is still an on-going task for the research community. Several behaviour-based approaches emphasise not on representing or reasoning about intentions, goals, or motivations, but instead rely on how predictions and patterns directly flow from

data [58]. They argue it is impossible to actually find out why some behaviours occur and only mere speculations can be made about the reasons that drove these behaviours. This is in contrast to other AI trends, where the objective is to model cognitive belief states, intentions and internal structures, supporting the idea that an intelligent system can successfully simulate a cognitive creature [59]. In this informal battle, many researchers raise concerns regarding the generalisation ability of the behaviour-based approaches and the problems deriving from it [60]. Truth is, they are partially valid, since the success of search engines such as Google and many other popular applications in behaviour recognition show that much can be done by using “just” correlation patterns.

In this context, the notion of *privacy* has surfaced time and again. However, while the word has remained the same, its meaning never stopped evolving. In the age of big data, and even more so in the future of the Internet of Things, this notion is poised to become all the more important. This phenomenon is taken to another level with “*profiling*”, the use of a person’s data to guess about aspects of his or her personality, generating insights about traits or habits that one may not even know are existing [61].

While extracting data and information from specific individuals is related to identification and control, in the context of data analytics the concept of *profiling* makes it possible to go beyond the personal level and track, monitor or measure various groups of individuals at an aggregated level. So, someone might reasonably wonder, “what is the difference between these two levels?”

The crucial difference is that the *individual level* deals with personal data of a specific individual, and this information is actually observed and recorded, i.e. it is factual knowledge. On the other hand, at the *profiling level*, the knowledge is not usually available. Instead the profile is applied to an individual so as to infer additional facts, preferences or intentions. Therefore, the profile consists of data-driven models that represent correlations, patterns or rules, that apply to a subset of the individuals; for example, missing children who have a family status X and are in the age group Y are more likely to be of missing category Z, if their daily engagement with social media is over W hours.

Profiling provides the means to infer knowledge about an individual that is not actually observed or recorded [62].

The aggregated level example is what Hildebrandt calls *non-distributive* profiles [63], in a sense that the profile properties will not hold true for all individuals that belong to that profile (in contrast to distributive profiles, where a property belongs to each and every member of the group).

In the literature study that follows, we investigate and review papers of the aggregated level, showcasing the fact that the profiling process may include various sources of information and have a huge impact in predicting personality traits and uncovering hidden behavioural patterns.

2.2.1.1 Research Literature Study

The analysis of human behaviour and activity is a quite broad topic with roots in many different sciences. A simple search for these terms (“human behaviour and activity analysis”) in Scopus¹⁴ search engine yields more than 64,000 results, broken down into a large number of categories such as Psychology, Sociology, Criminology, Medicine, Biochemistry, Neuroscience and Computer Science.

¹⁴ <https://www.scopus.com>

We, therefore, need to narrow down our scope, so we query Scopus about “Computational learning and profiles”. The results are then reduced to 1,177 documents, out of which 614 are related to computer science.

In order for the literature study to be efficient, a subset of these approaches and methodologies has been selected so as to present valuable insights and the underlying advancements and challenges of the field in respect to the ChildRescue project. More specifically, the following considerations were taken while reviewing the available publications and selecting the papers and reports to be analysed in this deliverable:

- *Research significance*, by citation count, which is indicative of the article’s research value.
- *Latest trends*, for the specific domain, in order to balance the fact that older publications have naturally more citations. Therefore, a separate search is made to include highly cited articles of the last few years (from year 2013 and afterwards).
- *Computational Learning* algorithms should be employed and evaluated.
- *Applicability*, towards the ChildRescue project. Papers with high relation and applied solutions to the ChildRescue project are favoured among others of similar citation number.

In Annex 4II.2– Computational Learning in Human Profiling Literature, sixteen (16) selected papers have been analysed in depth and compared.

2.2.1.2 Key points extracted from the Literature Analysis

In our literature review we examined and compared a number of methods and algorithms for learning models out of human behaviour data. This automatic process is what we call *profiling*, and the models derived from it are the *profiles*. Profiles can consist of rules or correlations about the relations between features, but they can also divide the group of individuals into a number of subgroups, the members of which share certain properties, or they can be complex functions that compute a value or class label based on some features.

Profiling can be applied on any task that involves human interaction and activity. In our study we encountered various application domains, such as healthcare [64], education [65], economy and marketing [66][67], gaming [68], cyber security [69][70], mobile communications [70][71][72] and, of course, crime analysis and missing persons investigations [42][73][74].

In the majority of these cases, the methodology followed, involves the same general, but sequential, steps:

- First, some type of clustering is performed to shape groups of similar characteristics out of raw data. This process results in an initial profiling model, which in some research studies can be deemed sufficient (e.g. in [65], [68] and [73]).
- The next, usually optional, step introduces a form of dimensionality reduction, either by using a known algorithm or some heuristic method based on domain expertise, in order to decrease the dimension space and keep only the useful and informative input features. This process optimises the profiling model.
- In the last step, the task of classification (or regression) is executed using well-known computational learning algorithms. This process leads to predictions based on the profiling model deduced from previous steps. In cases where a model already exists, as in most personality prediction tasks (e.g. [70][71][72]), the classification or regression process can take place as a first and only step.

Other approaches that deal with textual data, something very common in social media, employ Natural Language Processing techniques, as an additional, albeit necessary, step to the aforementioned routine [39][75][76]. On the contrary, in cases where a sequence of actions or events is used to describe a behaviour, the procedure involved is, in many ways, different. For instance, dealing with web navigation patterns or system intrusion command sequences, as in [77] or [69] respectively, the preferred method is to use stochastic approaches, such as the hidden Markov model.

A sum-up of all algorithms examined in this literature review is presented in the following table:

Table 2-4 List of computational learning algorithms for human profiling

Objective	Algorithms
Dimensionality reduction	<ul style="list-style-type: none"> • Isomap, • Principal Component Analysis (PCA), • Singular-value decomposition (SVD)
Clustering	<ul style="list-style-type: none"> • K-means, • Expectation-Minimisation (EM), • Agglomerative algorithm, • K-medoids, • Archetypal analysis
Classification	<ul style="list-style-type: none"> • Support Vector Machines (SVM), • Logistic regression, • Naïve Bayes, • Decision trees, e.g. C4.5, • K-nearest neighbours (k-NN), • Artificial Neural Networks (ANNs), • Ensemble Bagging, • MultiBoostAB, • AdaBoostM1, • Random Forest
Regression	<ul style="list-style-type: none"> • Linear regression, • Poisson linear regression
Deep Learning (time-series)	<ul style="list-style-type: none"> • Recurrent Neural Networks (RNNs)
Anomaly detection	<ul style="list-style-type: none"> • Hidden Markov models (HMM)
Rule induction	<ul style="list-style-type: none"> • BruteDL

At this point, it is worth mentioning that we deliberately omitted research publications concerning activity recognition and computer vision techniques on human motion patterns and face/body expressions (for examples see [78]), since these methods, most probably, will not be adopted by this project as they raise the most serious privacy concerns.

From a data source viewpoint, it is evident that retrieving social and activity characteristics from humans is not a simple task. The complexity and unreliability of human behaviour, as well as the ethical aspects surrounding it, make the effort even more demanding. Data usually include digital records collected by an organisation, which can be enriched by survey questionnaires and psychometric tests, as well as by external sources. Our literature analysis reveals that these external data sources can be found in wearable sensors and bio-signals, mobile phones, internet applications and social media, among others.

In the special case of missing persons, which relates to the ChildRescue project, a rather low number of research papers have been published that employ computational learning methods for profiling. One explanation for this could be the confidential and sensitive nature of the required data sets, which cannot be available to anyone. Blackmore et al. (2005) in their study [42], present a digital record of a missing individual case that consists of several variables taken from police files (Table 2-5). With the use of data mining techniques, they proceed to train and test a classifier. The results reported in this work, nevertheless, indicate there are some inconsistencies, probably due to some of the variables containing human judgments and estimations (e.g. "Is the missing person known to be socially deviant or rebellious" or "What does the reporting person suspect has happened") that, according to the authors, may eventually affect data quality. A note we must keep.

Table 2-5 Variables based on information in police files describing missing persons [42]

Variables	
<ul style="list-style-type: none"> • Does missing person have any dependents • Residential status • Time of day when last seen • Day of week when last seen • Season of year when last seen • Last seen in public • Is this episode out of character for the missing person • What does the reporting person suspect has happened • Any known risk factors for foul play • Is missing person known to be socially deviant or rebellious • Is there a past history of running away 	<ul style="list-style-type: none"> • Is there a past history of suicide attempt or ideation • Any known mental health problems • Any known drug and alcohol issues • Any known short-term stressors • Any known long-term stressors • Method of suicide • Was the perpetrator known or a stranger to the victim • Was the missing person alive, deceased or hospitalised when located

Regarding external data sources, social networking media have become a major research and business activity field due to the huge amount of public data that gets generated each day, as well as the availability of tools to retrieve and analyse them. It is also a research area undergoing rapid development and evolution, because of the commercial pressure and the potential for using social media data for computational (profiling) research. For instance, Kosinski et al. (2013) in [39] process about 58,000 user profiles from Facebook, exploiting user Likes on music, movies, product pages, etc., and the results of several [79], to predict personality attributes, such as political views, religion or sexual orientation.

It is clear that the domain of social media analytics is quickly rising and encompasses more and more applications, with profiling being just one of them [30]. For this reason, and for its important role in the ChildRescue project, social media analytics will be investigated and discussed separately in one of the sections to follow.

2.2.1.3 Challenges and Perspectives

Computational learning research on profile assessment has been based on a philosophy that emphasises prediction over explanation [80]. It has thus focused almost exclusively on the convergence of predictive analytics models with established personality measures. In contrast, little research has been done to use computational learning and digital records of behaviour to further understand a personality. So, there is appreciable potential in using current machine learning technology to develop improved tools for better understanding what the models are actually measuring, from a psychological point of view [81]. In other words, sociologists and psychologists, apart from a profile prediction, would like to know why and under what conditions something will occur, so that they can act pre-emptively.

Another challenge in profiling is the temporal factor of a behaviour. Human behaviours change through time, depending on external stimulations and certain events. Most research studies seem to ignore this part and prefer to form models of static knowledge in a particular time-frame. More recent studies though, explore the power of deep learning and seem to tackle this issue with success. Work in [66] is a good example.

The major challenge of profiling, however, is data privacy, and that is because the primary task of profiling is the development of models from aggregated personal data, which does raise several privacy concerns. In response to this, from the early days of data mining in the late 1990s, a part of the academic community focused their work on privacy issues leading to the notion of Privacy-Preserving Data Mining [82]. In this framework, and following the advancements in database technologies, many researchers started to study the technical feasibility of realising the data mining methods using perturbed records of individuals (i.e. randomly modified values with sensitive pieces of information) [83].

Nevertheless, after the enforcement of the EU regulation 2016/679 (known as GDPR¹⁵) in May 2018, most social networks and internet platforms have ceased to provide access (through APIs) to personal public data. A period of adjustments and optimisation from all sides seems to have begun, with unknown results. We do hope the desired balance between data security and protection, and scientific progress, will be reached soon.

In the ChildRescue conceptual approach, the profile assessment deriving from multiple data sources is regarded as one of the cornerstones of the project. It is, therefore, crucial to apply the most appropriate profiling methodology, depending on the data available, harnessing the power of predictive analytics and machine learning in the most efficient way. Data anonymisation techniques will be applied as well, so that the process of profiling is in line with the data privacy protection rules and regulations.

¹⁵ <https://eugdpr.org/>

Some of the aforementioned challenges will, most probably, be encountered in the context of the ChildRescue project. It is, therefore, important for the ChildRescue overall methodology to be able to perceive these challenges in time and confront them efficiently. In particular, the explanatory aspect of rule-inferring computational techniques should be considered when selecting the appropriate algorithms. Additionally, the human temporal behaviour will play a significant role in routing estimation and POI suggestion, and as such it will be covered in detail in section 3.3.3 of this document, while privacy issues will be addressed in section 3.2.

2.2.2 Spatiotemporal Data Analysis

As already shown, research about large scale human behaviour patterns is often based on user-generated data either within wireless communication networks such as mobile phone networks, or on social media networks like Facebook or Twitter. In combination with Geographic Information Systems (GIS) these inherently spatiotemporal data enable novel capacities to analyse and visualise large-scale human dynamics in a more integrated manner, even close to real-time.

The first actual use of spatial analysis ever recorded was back in 1854 when a British physician, John Snow, began mapping cholera outbreak locations and eventually noticed that the majority of cholera cases were commonly found along the water line [84]. A century later, from early 1960s to 1980s the advancements in mapping technology, computers and data storage led to what we know today as GIS. Actually, it was Roger Tomlinson, an English geographer, who first used the term "Geographic Information System" in his publication in 1968 [85]. He has been later acknowledged as the father of GIS.

Another account of early spatiotemporal analysis is found during the first decades of the 20th century and is closely related to the migration mobility problem. Many sociologists and demographers put effort on the interpretation of people movement in space using mathematical formulas and statistical analysis to predict the distance or direction of the (internal or not) migration streams, either in rural or urban environments [86].

It is true that human mobility patterns reflect many aspects of life, from the global spread of infectious diseases to urban planning, traffic forecasting, tourism and daily commute patterns. Today, more than ever before, the large availability of human location tracking datasets through location-aware devices and services due to the advancements in social media, mobile telecommunication networks and the large-scale deployment of GPS technologies, has generated growing scientific interest in their possible exploitation and interpretation. As this information is usually timestamped, it is also possible to detect occurrences of activities in temporal relation to each other or to specific daytimes. Examples of these datasets include: mobile phone calls, credit card transactions, bank notes dispersal, check-ins in internet applications and geotagged user-generated content in social networks, among several others. Both personally identifiable information of individuals, but also massive anonymous data are geolocated. User GPS logs or social media posts are examples of the first category, whilst mobile network traffic and anonymous smart-card transactions fall under the second.

Interdisciplinary approaches, generally deriving from knowledge discovery, are utilised towards the deciphering of raw mobility data. Algorithms and techniques from the fields of data mining, machine learning and statistical analysis are employed and in return lead to the discovery of human movement or transportation patterns and event detection [87]. Especially for human trajectories, it is found that

they show a high degree of temporal and spatial regularity, with each individual being characterised by a time-independent characteristic travel distance and a significant probability to return to a few highly frequented locations [88]. Mobility patterns were also found to represent human behaviour in catastrophic events, e.g. the 2010 Haiti earthquake [89].

Our research literature study aims to shed light on the domain of spatiotemporal data analysis (also encountered as *human mobility patterns analysis*) and review particular publications related to behaviour modelling tasks of tracking persons. Predicting their trajectory patterns or locating points of interest (POIs) will be our primary focus as well.

2.2.2.1 Research Literature Study

The target domain of our literature study is the spatiotemporal data analysis. A Scopus search for these terms ("spatiotemporal data analysis") yields more than 13,500 results, out of which the 3,891 belong to computer science. For the sake of a beneficial literature review towards the ChildRescue project, we need to narrow down the publications set, adhering to some criteria that can lead to better and more valuable insights, as well as challenges and perspectives related to the project's cause. More specifically, the following considerations were taken while reviewing the available research studies and selecting the papers to be analysed in this deliverable:

- *Research significance*, by citation count, which is indicative of the article's research value.
- *Latest trends*, for the specific domain, in order to balance the fact that older publications have naturally more citations. Therefore, a separate search is made to include highly cited articles of the few years (from year 2013 and afterwards).
- *Human patterns* should be in the core analysis of each work. We therefore excluded approaches that tracked animals, natural phenomena or objects.
- *Applicability*, towards the ChildRescue project. Papers with high relation and applied solutions to the ChildRescue project are favoured among others of similar citation number.

In Annex 4II.3 - Spatiotemporal Data Analysis, seventeen (17) selected papers have been analysed in depth and compared.

2.2.2.2 Key points extracted from the Literature Analysis

According to Toch et al. (2018) and their proposed taxonomy[87], human mobility pattern analysis can be categorised into three broad categories:

- User modelling, where the object of analysis is a single individual.
- Place modelling, where the object is a geographic area (visited by different individuals).
- Trajectory modelling, where the object is a set of spatial-temporal points created by the same individual.

User modelling applications analyse the mobility patterns of a single individual for extended periods of time. In such applications, the model can predict where a particular user will be at different times of the day or recommend POIs to visit (e.g. [90][91][92]). *Place modelling* applications analyse the characteristics of a geographic location or a set of locations, in order to profile it and classify the type of place according to the mobility patterns of people coming in and out of it (e.g. [93][94]). *Trajectory modelling* applications require a set of spatial-temporal points that reflect a trajectory, defined as a movement pattern through a set of locations or a set of objects and time (e.g. [95][96]). In contrast to user modelling, in trajectory modelling, the identities of the moving objects are not

necessarily a factor in the analysis. All three categories hold a different value on human mobility analysis, but in several cases, the borderline between them is not really clear.

In respect to applications, most of the reviewed studies aim to create some form of recommendation for the user, as part of a location-based service. For example, in [90][92][97] & [98] the next - predicted - point of interest is suggested to the user, based on his or her mobility profile, while in [99] the analysis of GPS trajectories from 107 users infers the right sequence of POIs, which could be tourist attractions, and recommends it to the user. If a man is lost in the wilderness, his route trajectory can be predicted based on the profile of the missing person (age, gender, professions, intention, etc., which translate into direction, distance, and dispersion of travel) and the man can be safely found, or so claim Mohibullah & Julie (2013) [95] and Lin & Goodrich (2010) [100] in their related work. Furthermore, an event or a place of social commotion can be detected using a probabilistic location inference on mobile phone patterns, according to the results of [101].

A very interesting idea, coming from Cho et al. (2011), is to combine social networking connections with place check-ins, following a hypothesis that we tend to go where our friends go, when we are not at work (or school) [91]. It is shown that social relationships can explain about 10% to 30% of all human movement, while periodic behaviour (e.g. going to work and back home every day) explains 50% to 70%. The rest is more or less unpredictable (i.e. no regular patterns).

Another emerging area of human mobility research is urban planning, where spatial and temporal patterns are studied and processed so as to explain urban traffic and adjust traffic lights or plan new transportation routes. Such information can be derived from transportation cards transactions [102] or location-based social media content [103]. Speaking of transportation, even the medium of transport can be predicted for a given mobility pattern, using a large set of recorded human GPS trajectories and urban infrastructure details [96].

From an algorithmic point of view, which defines our main interest, we encountered a clear distinction from the early review of the literature. On one hand, there are methods that make use of probabilistic (stochastic) modelling and analysis, and on the other hand, we have methods that employ machine learning algorithms and techniques.

The common ground on all these methods is the necessary step of data pre-processing which allows for an efficient and producible analysis. Usually a method of trajectory or area segmentation is required at this stage. Then, spatiotemporal data modelling can be accomplished either using a probabilistic or statistical approach (e.g. Markov model), or a machine learning technique (e.g. K-means clustering). In some cases, there can be a combination of both, each applied on different aspects of the same problem, like in [104]. In more recent works, deep learning algorithms have been employed to deal (mostly) with the temporal aspect of mobility pattern analysis, using Recurrent Neural Networks (RNNs) that have the ability to recall past states [92][97].

A sum-up of all methods and algorithms examined in this literature review is presented in the following table:

Table 2-6 Methods and Algorithms for spatiotemporal data analysis

Objective		Algorithms
Probabilistic	(Stochastic)	• Markov decision process and Order-K Markov model,

mobility modelling	<ul style="list-style-type: none"> • Probability density functions, • Bayesian model, • Custom models
Trajectory Clustering	<ul style="list-style-type: none"> • K-means, • Voronoi diagrams (partitions), • Non-Negative Matrix Factorisation, • Expectation-Minimisation (EM) • Density-based (e.g. OPTICS)
Classification	<ul style="list-style-type: none"> • Support Vector Machines (SVM), • Naïve Bayes, • Decision trees, • Artificial Neural Networks (ANNs), • Random Forest
Text based analysis	<ul style="list-style-type: none"> • TF-IDF, • Latent Dirichlet Allocation (LDA), • Dirichlet Multinomial Regression (DMR)
Reinforcement Learning	<ul style="list-style-type: none"> • Inverse reinforcement learning
Deep Learning (time-series)	<ul style="list-style-type: none"> • Recurrent Neural Networks (RNN, ST-RNN, LSTM)
Expert systems	<ul style="list-style-type: none"> • Fuzzy inference system
Recommenders	<ul style="list-style-type: none"> • Collaborative Filtering

Experimentation data originated from a great variety of sources, which constitutes another distinction of the relevant literature publications. Many researchers take advantage of mobile phone data, such as call detail records and GPS log files in order to predict significant locations [94], recommend interesting locations or POIs for next visit [92][97][98], detect social events [101] or travel sequences [99][106], even simulate and predict human mobility behaviour when lost in the wilderness [95][100].

A more recent trend in the research community targets social media data from location-based networks, like Foursquare [103][104], while in some cases the combination of social networks with mobile phone data seems to produce better results [91]. On the spatial aspect of the analysis, the use of open data, such as transportation infrastructure, road networks or land use, enrich the information available and improve predictive capabilities as shown in [93][96][105]. Of course, any other data source that can explicitly or implicitly offer some type of human mobility patterns (e.g. transportation card transactions [102]) can prove very useful for a successful data analysis.

The table below summarises the data sources encountered during our literature investigation, along with the respective datasets, that could potentially be used in the ChildRescue framework.

Table 2-7 Most widely used data sources in spatiotemporal data analysis

Data Sources	Relevant Datasets
--------------	-------------------

Mobile Communications	<ul style="list-style-type: none"> • Mobile phone calls and text messages (location points) • General mobile phone usage (user trajectories)
GPS	<ul style="list-style-type: none"> • Log files
Social Networks	<ul style="list-style-type: none"> • User check-ins (e.g. from Foursquare, Facebook) • Geo-tagged images and videos (e.g. from Flickr, Facebook) • Location reference in a post or activity (e.g. from Facebook, Twitter) • Other geo-referenced content (e.g. Yelp crowd-sourced location reviews) • Network connections
Open data	<ul style="list-style-type: none"> • Transportation infrastructure (e.g. from OpenStreetMap) • List of popular POIs in a particular area (e.g. from OpenStreetMap, national open data) • Transportation time schedule (city open data) • Land use (e.g. MassGIS) • Road networks (city open data) • Weather data
Other data	<ul style="list-style-type: none"> • Transportation card transactions (e.g. Oyster card) • Surveys (e.g. Dutch National Travel survey [107])

2.2.2.3 Challenges and Perspectives

We reviewed a number of different use cases regarding the exploitation of spatial and temporal data in order to infer location-based practical information. In some of these cases, additional data sources were employed, either from social networks (such as user connections or POI reviews) or originating from open data services (such as road networks or weather data).

Statistical analysis has been the most common approach for analysing spatial data, where a large number of algorithms exist including various optimisation techniques. A major drawback of this approach is the assumption of statistical independence among the spatially distributed data. Furthermore, statistical analysis cannot model nonlinear rules efficiently and cannot work well with incomplete or erroneous data, or with categorical values. These disadvantages were partially solved with the advent of spatial data mining [108]. However, traditional data mining techniques and algorithms perform poorly on spatiotemporal tasks, and require significant modifications to exploit the rich spatial and temporal correlations and patterns embedded in the datasets. The unique characteristic of spatiotemporal datasets is that they carry time, distance and topological information which require geometric, as well as temporal, computations. In most cases spatial and temporal relationships are not explicitly defined, and thus, they should be extracted from data, which ultimately causes a processing overhead [109].

Another challenge, related to using open data, is the fact that it is difficult to match the available open data to the datasets one plans to analyse, in respect to the desired location or time-frame. For example, while there are plenty of data for a metropolitan city in US for the year 2015, when you

seek a relevant dataset for a rural area in Belgium or Greece for 2017, you hardly find any. Therefore, careful consideration of the available open data sources should be made before selecting the appropriate ones.

Lastly, as in every aspect of this project, privacy issues play and will continue to play a prominent role when challenges are discussed. Preserving anonymity and protecting personal information should be the first priority. Thankfully, privacy is an essential requirement for the provision of electronic and knowledge-based services in modern e-business, e-government, or e-health environments, and thus, there are several methods and algorithms for protecting location privacy, among others, without losing much of their efficiency and performance [110].

For ChildRescue, the spatiotemporal data analysis will be an essential part of the active missing children investigation process, for its ability to recommend possible POIs the child might visit, and predict his or her trajectory near the area the child was last seen. However, we expect the available spatiotemporal data to be scarce and dispersed, so it is important for the analysts to have selected and modified the appropriate algorithms in such a way so that some probability indications can be obtained. Security precautions will also need to be considered to ensure the protection of data privacy.

2.2.3 Social Media Analytics

Social media can be defined as a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of User Generated Content [111]. Social media, nowadays, constitute an integral part of our everyday lives and affect us in several ways. The communication and interaction barriers have been diminished, while the ability for users to generate content, interact with friends and expose personal thoughts, visiting of places and snapshots of their everyday lives, have made social networks extremely popular, especially among teenagers. At the same time, social media have proven to be a valuable pool of information and trends, widely exploited in the business, bioscience and social science areas, e.g. for measuring customer satisfaction, forecasting stock market fluctuations, predicting social impact, or providing insights into community behaviour [112].

The opportunities associated with the analysis of all these data have helped different organisations generate significant interest in Social Media Analytics, which is often referred to as the informatics tools, techniques, frameworks and applications that collect, monitor, analyse and visualise critical data from social media, so as to facilitate and extract useful patterns and knowledge [113].

Social media analytics is actually an extension of methods and technologies used in web analytics. Over the past two decades, web analytics, that emerged with the coming of the internet and World Wide Web, has been an active field of research, building on the data mining and statistical analysis foundations of data analytics and on the information retrieval and text mining. The advancements in internet technologies and mobile communications led rapidly to the rise of social media platforms and services, with social networking being the most popular online activity. In fact, internet users seem to spend more than 20% of their online time on such platforms, according to recent statistics [112], generating an enormous amount of informative content, interacting with it and being affected by it. Thus, it was only natural for the web analytics research community to divert its interest towards the, most promising, field of social media analytics. This has led to the publication of numerous research works, as well as the creation of novel data services, tools and analytics platforms. The analysed

social media may include blogs, message boards, multimedia platforms, and customer reviews platforms, among others. However, it is the web search and microblogging social media (e.g. Twitter, Facebook, Instagram, or Foursquare) that have attracted most of the interest.

Social media data is clearly the largest, richest and most dynamic evidence base of human behaviour, bringing new opportunities to understand individuals, groups and society. In fact, social media analytics are not only about informing, but also transforming existing practices in politics, marketing, investing, entertainment and news media. Because of this, the scope of social media analytics varies greatly. The largest part of research study is devoted to text mining and processing techniques, like sentiment analysis, topic or influence modelling, impact monitoring, opinion mining or news and trends analysis [114]. Other approaches focus on community network connections, personality traits prediction, venue recommendations or microeconomics and marketing analysis. Even approaches that attempt to track down the spread of infectious diseases through social media have been proposed [115].

In this context, predictive analytics is the primary tool, able to utilise user content (text and images), personal profile information and user preferences, and the social network itself (network connections and their attributes), sometimes combined with open data and information from other sources, in order to provide meaningful insights and content-based analysis [116]. In the lines to follow we will investigate the power of predictive analytics when employed to process the rich and prosperous social media data.

2.2.3.1 Research Literature Study

As a relatively new term, but with a flourishing popularity, the “social media analytics” query in Scopus brings back about 2,241 documents. It is noteworthy to say that the first publications start to appear in 2005 and 2006, and from then on there is a large increase after 2012.

The wide spectrum of methods and applications is also observed. From news and opinion mining to stock market prices and election predictions. It is therefore necessary for us to focus on reviewing publications related to the purposes of ChildRescue, and in particular, under the prism of predictive analytics, we opt to investigate the subdomains of personality prediction, venue recommendation and sensitivity analysis.

For the first two subdomains, we have already had a glimpse in the previous sections. Here, we are going to investigate both domains from the social analytics perspective, covering the most informative, and useful to our cause, cases.

During our review, a few considerations were taken so as to carefully select the papers to be analysed, which are the following:

- *Applicability*, towards the ChildRescue project. Papers with high relation and applied solutions to the ChildRescue project are favoured among others of similar citation number.
- *Research significance*, by citation count, which is indicative of the article’s research value.
- *Latest trends*, for the specific domain, in order to balance the fact that older publications have naturally more citations. Therefore, a separate search is made to include highly cited articles of the few years (from year 2013 and afterwards).

In Annex 4II.4– Social Media Analytics, twenty-one (21) selected papers have been analysed in depth and compared.

2.2.3.2 Key points extracted from the Literature Analysis

The methodologies reviewed in the field of Social Media Analytics can be grouped into three main application fields: Personality prediction, Recommender Systems, and Sentiment Analysis.

Personality Prediction is a subcategory of predictive analytics, concerned with the classification of distinct personality traits, which are serving as the class categories. This classification can be performed on users of social media by sole use of openly available profile information. The purpose of this process is to determine hidden personality traits or attributes, perhaps unknown to a person's close family or friends. In the majority of the here presented articles, these categories are provided by the Big Five model [79]. Although the goal is the same, the feature set utilised by these methods varies greatly. In the most typical approach, personality prediction is conducted using social media posts, i.e. short public messages, that usually describe some activity or express some feeling, upon which a method of linguistic analysis is applied [117]. As an improved alternative, some researchers suggest the usage of meta-attributes that characterise a post, such as the number of words, length of text, emoticons, or exclamation marks [118], while others promote the exploitation of details that characterise a user profile, such as personal preferences, activities and networking structure [119]. In more recent publications, the combination of two or more social media platforms is proposed, which seems to improve the error rates of the classifiers [120][121].

A drawback of the big-five approach is the fact that the users, the profiles of which are used as dataset, have to answer a questionnaire survey in order for their personality to be categorised and for the algorithms to be trained. On the other hand, the exceptions to the big-five-centric approach, focus on particular emotions and intentions. For instance, De Choudhury et al. tackle the challenge of depression detection [122], while a more recent study investigates reddit forums for traces of suicidal ideation [123].

Recommender Systems seek to predict a user's preferences so as to suggest an item or service that the user will appreciate (and perhaps purchase). So, the unprecedented opportunity offered by the social media, for gaining insights into the needs and desires –current and future- of millions of potential customers, could not be left unexploited. A large number of the so-called recommender systems was developed over the last years, combining available user information (demographics, family status, purchases, personal preferences) and recommending items that were calculated as the most appealing for the user. An advancement in recommendation systems was the aggregation of social network parameters in order to refine the resulting recommendations with influence of the social circle. The prevailing techniques in recommender systems are the Content-based Filtering (CBF) and the Collaborative Filtering (CF), as well as combinations of the two.

In respect to location or POIs recommendation based on social media data, a variety of approaches can be found in literature. Some suggested systems exploit not only direct satisfaction indicators, such as ratings, but also user location history and check-ins to calculate user preferences and perform user profiling [124][125]. In other cases, collaborative filtering was adapted for increased estimation accuracy and minimised complexity. This was achieved by limiting the user space, for example only to the friends of the user based on behavioural studies claiming that friends are more likely to share tastes and preferences, but also drag one another to an event/venue [126]. Another approach was the selection of "local experts" and inferring a venue score based on their opinions [127]. Sentiment analysis has also been utilised for extracting user preference scores and assisting in location

recommendation [128]. In more recent studies, the aggregation and collaboration of multiple recommenders is examined in multi-dimensional contextual information [128][129][130].

The most popular of the three techniques in the studied literature was collaborative filtering, with the reason probably lying on the selected field of interest -location recommendation-, which is slightly differentiated from tangible items recommendation.

Sentiment Analysis is a subject of study, both as an applied technique used in other applications of Social Media Analytics (for example in predictive analytics), but also as an application itself. As defined in [131], sentiment analysis is the computational study of people's opinions, attitudes and emotions toward an entity. As such, sentiment analysis has been employed in many research studies that analyse social media data, especially for brand building and marketing monitoring applications. Sentiment analysis techniques are divided into two main categories, which can also be intersected in hybrid models: the machine learning approach (un-/semi-/supervised) and the lexicon-based approach [132]. Generally, both techniques achieve high performances.

The usual applications of sentiment analysis involve some form of text processing and linguistic analysis. Twitter is one of the most commonly studied social platform mainly for two reasons: the feasibility to easily access user content through the Twitter API and the character limitation of tweets. The number of related research works, prove this claim [133][134][135][136]. Facebook however, although not used as corpus often, is also a microblogging platform, and as such, its posts and metadata can be analysed in a way similar to tweets [137].

Sentiment analysis can also be expanded on images, a field relatively less covered. For instance, in [138], the visual information of an image is exploited for sentiment classification purposes. When the image sentiment analysis is enhanced with contextual information, such as tags and comments, it is demonstrated that the positive effect of textual information in the classification of images is limited up to a threshold, after which the calculations are subject to overfitting [139].

According to Kalampokis et al. (2013) [140], a common methodology for predictive analytics when dealing with social media data in general, involves the following three steps/subprocesses:

1. Data conditioning, i.e. collection and filtering of data, determining time-window and location/area, identifying profile characteristics.
2. Feature selection, i.e. choosing appropriate prediction variables usually based on some feature selection or feature transformation algorithm.
3. Predictive analytics, i.e. data modelling, pattern extraction and performance evaluation.

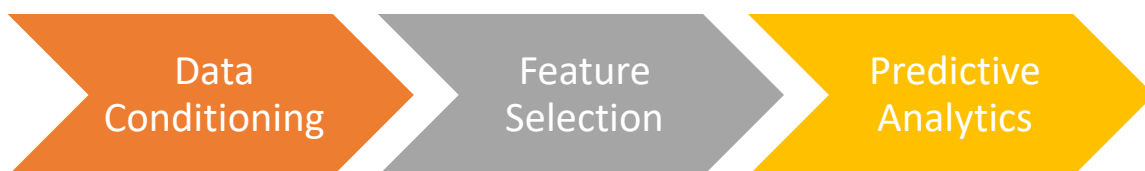


Figure 2-5 General methodology for predictive analytics in social media

The first phase, data conditioning, includes not only the collection of appropriate data, but also the selection of social media attributes that will lead to meaningful insights. Depending on the task, a selection should be made among volume variables (e.g. the number of posts, frequency), sentiment-

related variables (measuring the sentiment load of the data), variables depending on profile characteristics (e.g. number of friends, preferences, activities), as well as location and time-window specifics. The second phase, called feature selection, deploys an algorithmic strategy that excludes variables of low information, or creates new variables by transforming the data. The last phase of predictive analytics consists of: 1) the creation of the predictive model, such as linear regression, Markov models etc., 2) the incorporation of external predictors, like mobile phone data or GPS logs, and 3) the evaluation of the developed model using a generally approved metric.

The methods and algorithms encountered during the study of the social media analytics literature were differentiating depending on the given task. Table 2-8 summarises all of the examined approaches.

Table 2-8 Methods and Algorithms for social media analytics

Task	Objective	Algorithms
Personality prediction	Linguistic Analysis (NLP)	<ul style="list-style-type: none"> • Linguistic Inquiry and Word Count (LIWC), • Latent Dirichlet Allocation (LDA)
	Dimensionality reduction	<ul style="list-style-type: none"> • Principal Component Analysis (PCA), • F-statistic subsampling
	Clustering	<ul style="list-style-type: none"> • K-means
	Regression	<ul style="list-style-type: none"> • Ordinary Least Squares, • Linear regression, • Decision Trees, • Random Forest, • Support Vector Machines (with RBF kernel)
	Classification	<ul style="list-style-type: none"> • Naïve Bayes, • Logistic regression, • Support Vector Machines, • Multi-Layer Perceptron (MLP)
Recommenders	Location Recommendation	<ul style="list-style-type: none"> • Model-based Collaboration Filtering, • Memory-based Collaboration Filtering, • Custom Collaboration Filtering approaches
	Clustering	<ul style="list-style-type: none"> • Non-negative Matrix Factorisation
Sensitivity Analysis	Linguistic Analysis (NLP)	<ul style="list-style-type: none"> • Lexicon based clustering, • Lexicon based classification

	Deep Learning (images)	<ul style="list-style-type: none"> • Convolution Neural Networks
	Deep Learning (word-embeddings)	<ul style="list-style-type: none"> • Convolution Neural Networks
	Clustering	<ul style="list-style-type: none"> • K-means, • Non-negative Matrix Factorisation
	Classification	<ul style="list-style-type: none"> • Decision Trees (C4.5), • Support Vector Machines, • Naïve Bayes, • Maximum Entropy, • Random Forest

Although methods and algorithms may differ, the social media data types that are available for research are, more or less, the same. Text messages, user profile details, check-ins, reviews, activities and network connections are only some of them. In the following table, the reader can view all types of data sources utilised by methods and algorithms in the literature.

Table 2-9 Most widely used data types in social media analytics

Data types	Relevant Social Media
Text messages	<ul style="list-style-type: none"> • Twitter, Facebook, Reddit, Digg, MySpace
User check-ins	<ul style="list-style-type: none"> • Facebook, Foursquare, Gowalla, Brightkite¹⁶
Profile details	<ul style="list-style-type: none"> • All, but Facebook has the most information
Customer Reviews	<ul style="list-style-type: none"> • Foursquare venues
Activities (Likes, shares, etc.)	<ul style="list-style-type: none"> • Facebook, Twitter, Instagram
Images	<ul style="list-style-type: none"> • Instagram, Flickr
Audio-Video	<ul style="list-style-type: none"> • YouTube

2.2.3.3 Challenges and Perspectives

It is by now clear that social media analytics can be a powerful tool to the scientific toolkit. For some, the introduction of social media data analytics has had impacts in the study of human behaviour similar to the invention of the microscope or the telescope in the fields of biology and astronomy: it has produced a qualitative shift in the scale, scope and depth of possible analysis. Such a dramatic

¹⁶ Gowalla and Brightkite used to be location-based social platforms that both closed in 2012. Various datasets extracted from these two social networks are still utilised for research experiments.

leap, however, raises several concerns to the researchers that derive from the multi-faceted and complex nature of human communications and socio-cultural interactions [141].

One such issue is the prevalence of single social platform studies (e.g. Twitter, or Facebook), which overlook the wider social ecology. Even in cases where multiple platforms are examined, their data are processed separately, for comparison reasons. Therefore, multi-platform analyses should be sought and multi-methods should be examined, that will be able to capture the overall user social interaction and diffusion. In addition, whenever possible, the analysis of social media data should be paired with surveys, interviews, ethnographies, and other methods so that biases and short-comings of each method can be used to balance one another and arrive at richer answers.

Another major difficulty is the undesirable computational overhead set by the large volume of social media data. Therefore, the systematic filtering of irrelevant and noisy social media data prior to the analysis, is of high importance for the accuracy of the resulting predictions and the elimination of unnecessary complexity.

Furthermore, the raw data encountered in social media is usually of poor processing quality. Fake accounts, type errors, or on purpose lies and false data, create a headache for the analyst, adding an extra overhead to the overall procedure of extracting useful information. Clear examples are the attempts to predict presidential elections, as explicitly shown in [142], where the methods tested perform slightly better than clear chance. According to the same paper, for such an election prediction system to be reliable, it must have a strong algorithmic background, consider the idiosyncrasy of the area and the possible manipulations from spammers and thirdly, identify why the system produces the results it does. These conclusions could be generalised for predictions using social media analytics in other fields, as well.

As already noted in previous sections, during the last few years, there are raising concerns about social media public data harnessing. Researchers worldwide are facing the fact of the increasing access restriction set by the companies (be it for user privacy or business protection) towards social media data. The new EU regulation (GDPR - effective since the 25th of May 2018 worldwide), in particular, has already made the most popular social networks restrict the retrieval of public data through their APIs, which were freely available before.

This brings forth a major issue for the ChildRescue project, since social media analytics aim to provide significant added value to the current status of missing children investigation. In what way the project will handle the limitations in accessing social media information, has become an open challenge, which ChildRescue further discuss and address in section 3.4 of the presented deliverable.

2.2.4 Discussion and key-takeaways

The ability of analysing data coming from multiple sources, in order to model human behaviour and make applicable predictions, has been investigated through a literature overview.

The results show that, on one hand there is great multi-disciplinary research interest in this field, but on the other hand, the potential benefits are hindered by the overall human complexity and unpredictability. Traditional data origins, like forms, surveys, or digital records, as well as modern sources of information, such as open data, mobile phone activity, GPS logs, and social networks and applications, were encountered during this investigation, exposing the vast availability of human behavioural data.

Most of the methods and algorithms that were studied and analysed in depth, derive from the field of computational learning and data mining, whereas probabilistic techniques were applied when the research was related to spatiotemporal datasets.

Some important key-takeaways taken from the literature analysis are the following:

- Multiple data sources produce better results than a single source.
- The selection of the most appropriate features in a dataset increases performance.
- The quality of data plays a significant part on data analytics in general.

The traditional methods of collecting and processing behavioural data and classifying the data-subjects used to rely on sociological methods, manual analysis and interpretation. Nowadays, the process is highly automated and dependent on computer technology and many organisations routinely use data mining technology and techniques to analyse the large amounts of data available. However, adapting to an era of data-driven decision making is not always a simple task. Some companies have invested heavily in technology but have not yet changed their organisations so they can make the most of these investments. Others are struggling to develop the talent, business processes, and organisational muscle to capture real value from analytics. But the most common mistake is the insatiable and aimless gathering of unstructured and uninformative data just for the sake of data collection, which results in low quality datasets that require a lot of manual labour in order to be qualified for data analytics. Thus, in practice, the integration and analysis of multiple data sources coming from different organisations becomes even more challenging.

For sure, ChildRescue has a difficult task lying ahead: The modelling of missing children and unaccompanied migrant minors' behavioural patterns, taking the most out of available data sources. Using this profiling model, estimations and predictions regarding the child's whereabouts can be made. If successful, a missing child can return home, to its parents, in a faster, safer and more reliable way. The sections 3.3.1 to 3.3.3 are devoted to this exact mission.

2.3 Privacy and Anonymisation

The present section aims to provide an overview of the main techniques for anonymising and encrypting data depending on some privacy challenges which are also highlighted in Section 2.3.1. To this end, the definition of the current methods and models that ensure data privacy is presented. The extent to which these techniques can be used to conform to GDPR requirements will also be described, before moving to the details regarding how these techniques will be used in the context of ChildRescue.

More specifically, the approach will heavily rely on data encryption and data masking of personal sensitive data, since in the case of ChildRescue, many messages need to be broadcasted publicly or in an extended network of people. Missing children alerts, and evidence created by citizens acting as social sensors are examples of such messages. Performing data masking to these data ensures that only relevant information is transmitted publicly, whereas irrelevant data that contain personal information are hidden. Apart from the case of the personal sensitive data, various techniques and methods are also specified that focus on the privacy protection for the users of location-based services.

Before proceeding with an analysis of the existing techniques, the definition of the terms pseudonymisation and anonymisation is going to be provided as these are used in most legal documents and especially in the GDPR.

- **Pseudonymisation** means the transformation of data in such a way so that personal data cannot be retrieved without the usage of additional identifiers, not present in the transformed set.
- **Anonymisation** means the transformation of data in such a way so that personal data cannot be retrieved in any way from the transformed set.

The definition covers all types of data transformation so, technically, encryption can be used both for pseudonymisation (by keeping the decryption key in a separate database) and for anonymisation (by dropping the key or using one-way hashes). However, in the present section, we will cover encryption separately and we will focus on data masking techniques only when discussing pseudonymisation and anonymisation.

2.3.1 Research Literature study

The techniques and methodologies that can be used in the case of personal sensitive data along with those in the case of data privacy of location-based services, will lead to the definition of certain privacy methods and approaches, aiming to compose the complete research literature study. The results of this research will be presented in *Annex II: Analysis of Research Literature*, through a comparative table by denoting the main differences among the privacy models and highlighting their main objectives.

It is often the case that data that concern individuals need to be communicated publicly. Examples of such cases are medical studies, poll results and, in the case of ChildRescue, data about missing children. While personal information is inevitably disclosed in these cases, it is necessary to restrict the amount of data published or stored to the bare minimum needed for each case. When personal identification is not needed, the data disclosed should be such that no identification can be performed by reverse engineering, ensuring effectively that the people described by data remain anonymous.

At the same time, the increasing capabilities of position determination technologies (e.g., GPS) in mobile and hand-held device facilitates the widespread use of Location Based Services (LBS). Although LBSs are providing enhanced functionalities and convenience for ubiquitous computing, they can introduce new vulnerabilities that can be exploited to target violation of security and privacy of users, since LBS enabled applications require knowledge of the location of the individual/user. This may pose a major privacy threat on its users; consequently, for LBS applications to succeed, privacy and confidentiality are key issues. In this context, several approaches have been proposed for protecting location privacy of a user. The fundamental idea behind all techniques is to prevent revelation of unnecessary information and to explicitly or implicitly control what information is given to whom and when [143].

Location obfuscation is a technique used in location-based services to protect the location of the users by slightly altering, substituting or generalising their location in order to avoid reflecting their real position. Obfuscation, or closely related ideas, have already been suggested in the literature. Hong et al. within the Confab system, use landmarks and other significant locations to refer to user's location instead of coordinate-based geographic information. The effect of this "reverse gazetteer" is

to provide less information about a user's exact location. Similarly, Snekkenes suggests adjusting the precision of an individual's location as part of a mechanism for policy-based protection of location privacy [144]. Data transformation is another approach that can also be considered, including several models which will be extensively described in the next section.

2.3.2 Techniques & Methodologies

Data anonymisation can be performed by a variety of techniques, that are not needed to be performed exclusively [145]. The first step in anonymising the data set is performing removal or encryption of personal identifiable information (PII) [146]; these include information like name, address, id number etc. The mapping to the original data can be maintained in a separate database in case of pseudonymisation or be entirely discarded for anonymisation.

Removal of PII often is not enough for ensuring privacy however, since combinations of other information can still lead to the identification of a person, especially if said combinations occur rarely in the original data set. Combinations of *Personal Characteristics* data (also called Quasi-identifying attributes), like ethnicity and sex, are typically such combinations. If, for example, a single combination of a certain nationality, sex and marital status appears in a data set, this certain person can be identified by only requiring the extra knowledge that she/he is a member of the data set. Anonymising a set that contains personal characteristics data typically involves one or more of the techniques of **data masking**, namely a) encryption, b) shuffling, c) substitution, d) variance, e) masking and f) pruning.

Encryption, as already mentioned, is covered in a different section.

Shuffling is the technique of randomly or via an algorithm, changing the values of a data column. The transformed data contain entries that cannot lead to the person's identification since the information contained in each row are now invalid. The transformed data sets may also, depending on the case, keep its main statistical characteristics and still be useful for statistical analysis. Shuffling has the drawback that when performed as the single anonymisation technique, can leave the transformed data set open to reverse engineering attempts, especially when the shuffling algorithm is known.

Substitution is similar to shuffling, the main difference being that the substituted values do not originate from values occurring in other entries of the data set, but rather in external lists. A list of all the ethnicities for example may be used to randomly change the ethnicity of an entry. In contrast with shuffling, this may lead to the appearance of values not appearing in the original data set.

Variance is typically applied to values of numerical nature and consist of adding or subtracting a random noise to the value; an example is to compute a random number between -5% and 5% of a person's height and add/subtract this to the original value. The variance technique can too lead to the formation of data sets that keep their original statistical characteristics, if the mean value of the random noise added is zero. The higher the order of the moments that is required to remain close to the original dataset, the more customised the noise distribution must be.

Masking is the substitution of all, or part of, the value of a field with null values or a standard character. It is typically used in credit card numbers where the unmasked fields can be used to identify details like type of card (VISA or MasterCard), but numbers unique to the individual person

are being left out. Masking also applies to non-text data, like images or video. Substituting the face of minors in media pictures is a typical case of image masking.

Pruning is simply the deletion of a record. Pruning can be used in records that have rare combinations and are thus easier to be traced back to a certain person. As these combinations are rare, the statistical properties of the data will not be affected much, however there is the drawback that such combinations often have scientific value. A special case of pruning is the deletion of only certain values of a record by removing columns or assigning null values; this last technique is also called nulling. Nulling is efficient for discarding personal data that are not needed for processing but has the drawback that it can draw suspicion that data masking has been performed on the data set.

Except from Masking and Pruning, all anonymisation techniques can be used in such a way that the transformed data set, though technically contains false data, appears realistic to an outside viewer.

Figure 2-6 depicts an example of data masking, with each column being affected by the transformation being denoted by the purple outline. Except from the masked credit card number and the null ages, all other values appear realistic. However, the combinations of name and sex lead to a male Helen and a female Jack which do not appear realistic. Data masking algorithms can be designed such that realistic looking combinations and values are produced in the transformed data set. The set of all the transformations performed defines a transformation matrix; the inverse of this transformation matrix may be used to restore the original data. When the transformation matrix is stored in a separate database, the data set is considered pseudonymised, whereas, if it is deleted, it is considered as anonymised.

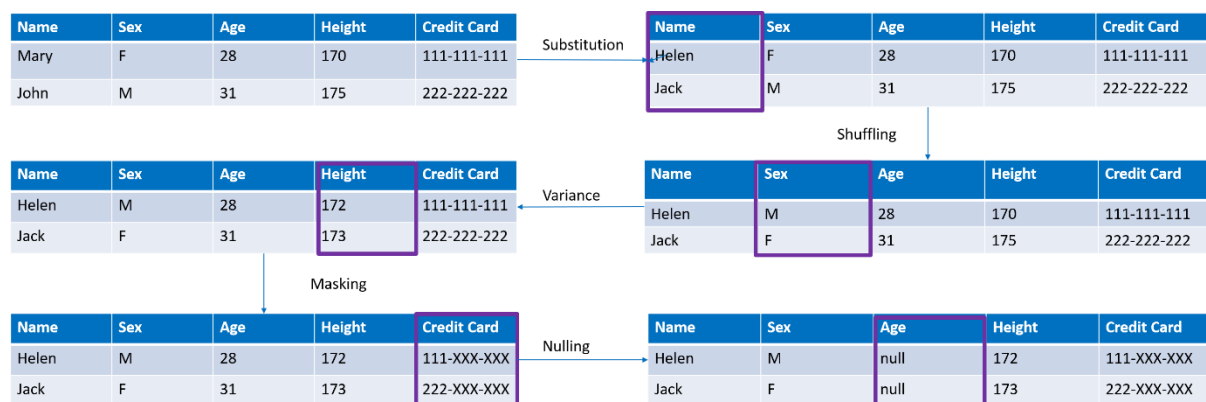


Figure 2-6 Example of data masking

2.3.2.1 Data anonymisation models for dealing with attackers having access to multiple data sets

Though in the previous sections we analysed the common masking techniques, the degree of anonymisation often depends not only on the efficiency of the transformation themselves, but to the extent that an attacker possesses knowledge of extra datasets containing information of the data subjects. These datasets may be also anonymised, but when combined they could be used to de-identify the subject(s); a typical example involves combination with clinical data released as anonymous with a public voter dataset [147]. To handle these cases, a set of privacy models have been suggested; each one attempts to quantify the degree of anonymisation to one or more metrics and transform the data set in a way that target values of the metrics are achieved.

In the following sub-sections, a brief overview of the most commonly used models will be given.

2.3.2.1.1 k-anonymity

A data set is k-anonymous [148] if each entry from the released data set cannot be distinguished from at least k-1 other entries. The distinction is based on the number of attributes identified as personal characteristics (also called quasi-identifiers). Each group that shares the same values of the quasi-identifiers is called an equivalence class. Figure 2-7 depicts an example of a k-anonymous set with k=2. Here, a set of four attributes have been designated as quasi-identifiers, namely the race, birth year, gender and postal code (which is masked).

To achieve k-anonymity two techniques are commonly being employed:

- Suppression, is similar to masking and assigns a generic single value to a variable. This can be like the masked postal code of the example or a fake value if a realistic look of the data is needed
- Generalisation, in which groups of values are merged into one. For example, birthdates can be grouped into decades so that any birthdate from 1980 to 1989 being assigned the value of 1980.

Race	Birth	Gender	Postal Code	Diabetes
White	1980	male	20*	Yes
White	1980	male	20*	No
White	1980	male	20*	Yes
White	1982	female	18*	Yes
White	1982	female	18*	Yes
Black	1982	male	18*	Yes
Black	1982	male	18*	No

Figure 2-7 Example of a k-anonymous set with k=2 and three equivalent classes

One of the main problems of k-anonymity is that efficient generalisation depends on the proximity of values. If values are very dispersed, broad categories should be used to achieve a high value of k , and this leads to degradation of data (the data loses its usefulness to the recipients instead of only to the attackers).

Another issue is that what is considered a non quasi-identifier by the holder of the data may be effectively be a quasi-identifier in the hands of an attacker. In the example given, diabetes is not part of the quasi-identifier set and was not transformed under k-anonymisation. However, attributes not transformed may exhibit a lack of diversity; if this is the case the attacker may use this to expose sensitive information. In the example given, the attacker may know that "Alice" is white and was born in 1982. Though the attacker cannot distinguish which of the two rows (4 or 5) correspond to Alice,

she/he will know that Alice has diabetes. The problem can be remedied by considering *Diabetes* to be quasi-identifier too, however this tends to aggravate the problem of efficient generalisation described above.

2.3.2.1.2 k-map

Similar to *k-anonymity*, the *k-map* [149] requires that the entries cannot be distinguished from at least $k-1$ entries; in the case of *k-map* however, the records counted correspond to the larger population and not only to those present in the data set and thus the computed k values are adjusted accordingly along with the characteristics of the relevant transformations. Consider for example the postal code as a quasi-identifier and that we perform k -anonymity with $k=5$. Suppose further that a postal code appears in the dataset that corresponds to a town with 20 registered residents. An attacker can identify this, and reduce her/his search space greatly. To remedy this, we can mask the postal code. While this transformation may not alter the k -value relative to the original data set, it will greatly increase it relevant to the larger population.

2.3.2.1.3 l-diversity

With l -diversity only equivalent classes that have a certain amount of diversity in their sensitive data are contained in the data set. The diversity is measured by the parameter l and there are various ways to define the measure of “ l -diversity” [150]:

- distinct: where each equivalent class is required to have at least l distinct values in their sensitive data
- entropy: where the entropy of each equivalent class E denoted by $Entropy(E)$, is greater or equal to the logarithm of l . $Entropy(E)$ is defined as: $Entropy(E) = -\sum_s p(E,s)\log(p(E,s))$ with $p(E,s)$ being the fraction of records in E that have the sensitive value s .
- (c- l) recursive: where the counts $n(E,s_i)$ are computed for each equivalent class and each distinct sensitive value s_i (so that $n(E,s_1)$ is the number of times that s_1 appears in E , $n(E,s_2)$ is the number of times that s_2 appears in E , and so on). The counts are then sorted in descending order and a constant c is chosen. If n_{max} is the first element of the list of sorted counts, (c- l) recursive criterion requires that: $n_{max} \leq c \sum_{n \neq n_{max}} n$

For the example of Figure 2-7, if distinct diversity is chosen, the data set has a diversity equal to one. Though l -diversity achieves better results in terms of anonymisation than simple k -anonymity, it too can have its limitation depending on the characteristics of the data set. If a sensitive value is very rare (e.g. a rare disease), it is difficult to achieve a high value of l . Furthermore, the equivalent classes may be low when compared to the count of the data entries to achieve a satisfactory value of l . For 10000 entries for example, we must have a maximum of 100 equivalent classes to achieve an l value equal to 2. As a final note, *l-diversity* does not distinguish between the semantics of sensitive values. This has as a result that entries are considered diverse although semantically, they are not. Consider the example depicted in Figure 2-8. Although the equivalent class has a diversity equal to 2, an attacker with partial information may infer that “Alice” has a serious medical condition.

Race	Birth	Gender	Postal Code	Condition
White	1982	female	18*	HIV+
White	1982	female	18*	Breast Cancer

Figure 2-8 Example of diverse, yet semantically linked, sensitive information

2.3.2.1.4 (α, k) -Anonymity

The k -anonymity model is proposed in order to prevent the re-identification of individuals in a released data set. However, it does not consider the inference relationship from the quasi-identifier to some sensitive attribute. In addition to k -anonymity, (α, k) -Anonymity model requires that the value of the frequency (in fraction) of any sensitive value in any equivalence class set is no more than α after anonymisation [151].

2.3.2.1.5 t -closeness

Intuitively, an equivalent is defined to have t -closeness [152] relative to a sensitive attribute, if the distribution that the sensitive attribute has in the equivalent class, is similar to the one it has in the whole data set. More precisely, if the distance between the distribution of the sensitive variable in the equivalent class and the distribution of the sensitive variable in the data set is at most t , then the equivalent class has t -closeness relative to the attribute. If all of the equivalent classes have t -closeness, then the whole data set is said to also have t -closeness.

t -closeness ensures a good level of anonymity on the data; however, it has the common problem of all anonymisation techniques, mainly that what is considered sensitive data may in fact be a quasi-identifier in the hands of the attacker. For example, an attacker could, by external means, know that a person has some, yet unknown, medical condition and search the data set to identify the exact condition or at least narrow down the list.

2.3.2.1.6 δ -presence

For an equivalent class, δ_E is defined as the ratio between the size of an equivalent class that is present in a dataset and the size of the same equivalent class as this is considered in the general population. Consider as an extreme case that a town with postal code equal to 12345 has a total of five residents and that a dataset contains 4 subjects with the same postal code. Then we know that 4 out of the 5 who live in that town are part of the dataset. In numbers, the δ_E parameter for this equivalent class has a value of $4/5$, or 80%. In effect, δ_E is the ratio between the k -anonymous and k -map equivalent classes.

It is evident that for anonymised data sets, the δ_E for each equivalent class, should be the smallest possible. In essence, if we define the δ of the data set to be the maximum value of the δ values of each equivalent classes, then δ should be as low as possible.

Although δ is a great indicator for anonymisation, the main difficulty lies with estimating the data set that represents the larger population. This data set is, in many cases, not available so that estimations may be given by the data experts.

It is noteworthy that δ -presence, as originally was proposed [153], involved the computation of two δ parameters: δ_{max} , which is the same described above and δ_{min} , which was the *minimum* value of the δ values of each equivalent classes. This was to compensate for symmetry attacks, namely attacks that were based on the knowledge that someone was *not* in the dataset. However, this case is not encountered frequently in practice.

2.3.2.1.7 p-sensitivity

Several methods have been proposed to support location-based services without revealing mobile users' privacy information. There are two types of privacy concerns in location-based services: location privacy and query privacy. Existing work, based on location k-anonymity, mainly focused on location privacy and are insufficient to protect query privacy. In particular, due to lack of semantics, location k-anonymity suffers from query homogeneity attack. To this end, p-sensitivity is introduced as a novel privacy protection model that considers query diversity and semantic information in anonymising user locations [154].

2.3.2.1.8 historical k-anonymity

Most of the defenses presented so far pay less attention to historical attacks, namely those attacks that take advantage of the acquisition of a history of requests that can be recognised as issued by the same (anonymous) user. The conditions enabling this kind of attacks are very likely to occur in LBS. In this context, several algorithms for providing historical anonymity as a defense against historical attacks have been presented [155].

2.3.2.2 *Data Anonymisation via aggregation*

A special case of data anonymisation is when only aggregated data are needed and information corresponding to specific individuals is of no concern. If, for example, a survey needs to check for correlations between nutrition habits and specific health markers, once the correlations are obtained the original data in row format are no longer needed. It is, thus, possible to store only the aggregated data needed for correlations and discard the original data set; the aggregated data can be considered as anonymised.

One common technique for data aggregation is the Data Cube [156]. A Data Cube as an $m_1 \times m_2 \times \dots \times m_n$ array, with n being the number of aggregated variables and m_i being the number of distinct values the variable indexed by i can take. Figure 2-9 depicts an example of a Data Cube for the case of $n=3$, storing the aggregate values concerning disappearance incidents of children. The variables are *Financial Status*, *Family Status* and *Age* and indexed appropriately. For example, an index of zero for the Financial Status can mean that the child's family is in extreme poverty, an index of 1 for Family Status can mean that the parents of the child are divorced and an index of 1 for the Age may correspond to the age of 6. If we name the Data Cube matrix as DC , in such a scheme the value stored in cell indexed as $DC[0][1][1]$ will contain the number of disappeared children, who came from a family with divorced parents, who live under extreme poverty and who were of age 6 when they disappeared. In similar way, $DC[0][1][2]$ may mean the number of disappeared children, who came from a family with divorced parents, who live under extreme poverty and who were of age 7 when they disappeared, and so on. If a certain combination has a very low count when compared to the average value, then we discern that this combination is a rare one and can lead to

identification of the subject's details. Such combinations are thus dropped. For further security a random small noise with a mean value of zero can be added to each cell, so that the data set is further obfuscated.

2.3.2.3 Location obfuscation

Location Obfuscation is the process of degrading the quality of information about a person's location, with the aim of protecting that person's location privacy. It is the process of slightly altering, substituting or generalising the location in order to avoid reflecting real, precise position. The most common techniques to perform obfuscation are pseudonyms, spatial cloaking, adding random noise and dummies and redefinition of possible areas of location [143][144]. The most representative of these techniques are presented in the following subsections.

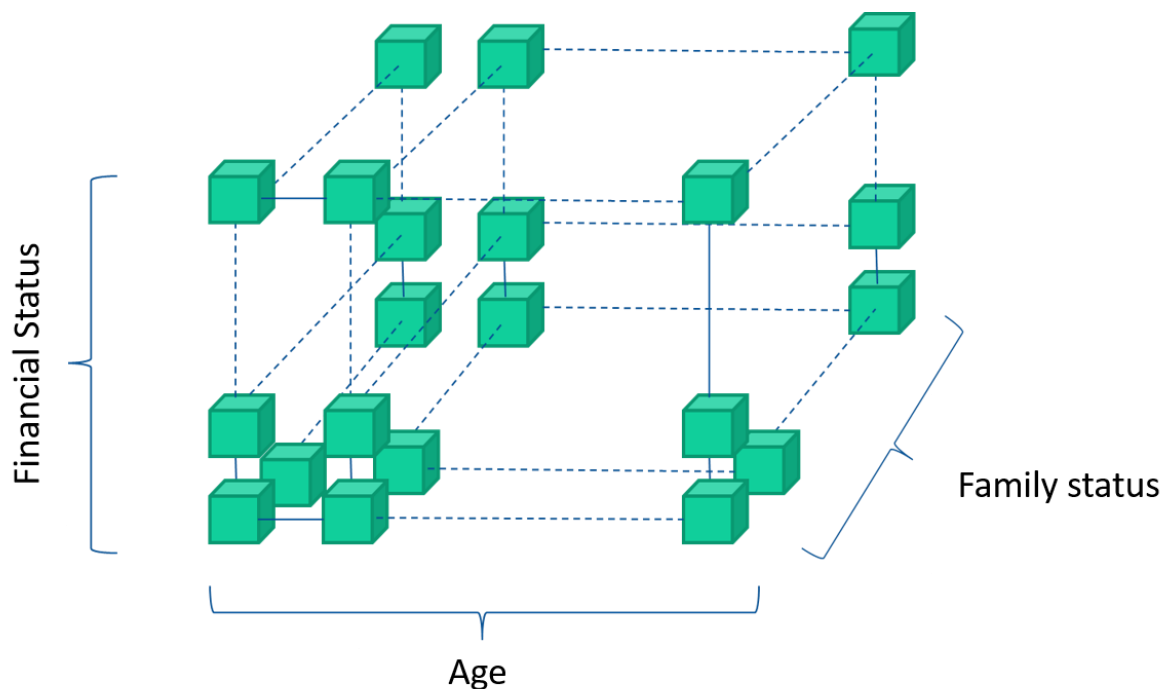


Figure 2-9 Data Cube Example

2.3.2.3.1 Spatial k-anonymity

Spatial K-anonymity (SKA) exploits the concept of K-anonymity in order to protect the identity of users from location-based attacks. The main idea of SKA is to replace the exact location of a user U with an anonymising spatial region (ASR) that contains at least K-1 other users, so that an attacker can pinpoint U with probability at most $1/K$. Simply generating an ASR that includes K users does not guarantee SKA. Previous work defined the reciprocity property as a sufficient condition for SKA. However, the only existing reciprocal method, Hilbert Cloak, relies on a specialised data structure [157].

2.3.2.3.2 Generation of dummies

Another approach for location privacy under the category of obfuscation is generation of dummies. To add dummy locations and noise to user's position proposed an idea of sending additional set of

dummy queries along with the actual query. The obfuscation region consists of the distinct locations included in the query set sent to the LBS [143][158].

2.3.2.3.3 Pseudonyms

Broadly speaking, a pseudonym is a set of artificial identifiers that replace PII. In contrast with anonymisation of PII covered in the previous section, pseudonyms are mainly applicable to users of a service of platform and are typically used to cover the personal details of a platform user, as she/he appears on the platform. These personal details do not only cover the typical PII, like name, address etc., but also details of usage that can be linked back to the user. The subject's IP address for example, can be used to trace back her/his location and should therefore be hidden when using a pseudonym. Usage of pseudonym is required when there is a need that a subject appears under a false identity. The identity of an informant contacting the police for example, should remain secret to the public, however the police should be able to identify her/him.

Apart from GDPR compliance, the usage of pseudonyms is critical when the identity of a user needs to be protected; in ChildRescue this is very important for when a citizen, who contacts the organisations or authorities for providing evidence or new information, needs to be able to retain her/his anonymity.

Typically, pseudonyms are generated via an IP masking infrastructure such as TOR. Users negotiate a key to establish a secure connection with the pseudonym issuing authority; this authority is typically the platform that implements the pseudonym infrastructure. The authority issues a certificate for the pseudonym, which is then used by the user for authenticating her/his pseudo-identity. The authority keeps the correspondence between the personal identity and the pseudo-identity and appropriately authorises the pseudo-user and audits her/his activities to discern any attempt of pseudonym hijacking.

2.3.2.4 *Spatial Data transformation methods*

In this setting the data has been transformed using some encoding methodology like Hilbert curve etc. prior to transmitting it to the LBS. An authorised client has the secret transformation keys. This client issues an encoded query to the LBS. Both the database and the queries are unreadable by the LBS. In this way, location privacy is protected [143].

In the context of location privacy of spatial data, the transformation of the original locations of POIs is needed. An ideal space transformation method should be a one-way function, which is easy to compute but difficult to invert. Meanwhile, to maintain the query efficiency of encrypted spatial data, the space transformation method should respect the spatial proximity of the original space. In the following sections, the standard Hilbert curve (SHC) is introduced, which is a representative spatial transformation method. Then, the Density-Based Space Filling Curve method is presented, which achieve better security than SHC [159].

2.3.2.4.1 Standard Hilbert Curve (SHC)

Space filling can be applied in spatial data transformation to protect location privacy for outsourced spatial data. Space filling curve passes through every partition of a closed space and has no intersection with itself. In this way, each point in multidimensional space will be mapped as a value to one-dimensional space. Z curve, Gray curve, and Hilbert curve are all space filling curves, which can

be used for space transformation. Compared to Z curve and Gray curve, Hilbert curve is widely applied due to its superior clustering and distance-preserving properties [159].

2.3.2.4.2 Density-Based Space Filling Curve

DSC partitions the spatial domain according to the capacity, which is the maximum number of POIs a partitioned region contains, denoted as C . It uses fractal rules of Hilbert curve to determine the visiting sequence of each partitioned region.

Two steps are involved in the DSC generation. (1) According to the capacity, the spatial domain is partitioned by quad tree structure. And the generated partitioned regions will be represented as quad tree nodes. (2) Based on the curve orientation, starting point, and scaling factor given by DO , each partitioned region is traversed sequentially in accordance with the fractal rules of Hilbert curve, then the sequence number of each partitioned region is generated, and we call this number DSC value, which is used to build indexes of POIs [159].

2.3.2.5 ARX

ARX is a comprehensive open source software for anonymising sensitive personal data. It supports a wide variety of (1) privacy and risk models, (2) methods for transforming data and (3) methods for analysing the usefulness of output data [160]. Due to the fact that this section focuses on privacy models and ARX supports several combinations of privacy models, some of them are presented in the following subsections.

2.3.2.5.1 Population uniqueness

This privacy model aims at protecting datasets from re-identification in the marketer model by enforcing thresholds on the proportion of records that are unique within the underlying population. For this purpose, basic information about the population has to be specified. Based on this data, statistical super-population models are used to estimate characteristics of the overall population with probability distributions that are parameterised with sample characteristics. ARX supports the methods by Hoshino (Pitman), Zayatz and Chen and McNulty (SNB). Different models may return differently accurate estimates of the number of population uniques. As a rule of thumb, the Pitman model should be used for sampling fractions lower than or equal to 10% [161]. ARX also implements a decision rule proposed and validated for clinical datasets by Dankar et al [162].

2.3.2.5.2 δ -Disclosure privacy

This privacy model can also be used to protect data against attribute disclosure. It also enforces a restriction on the distances between the distributions of sensitive values but uses a multiplicative definition which is stricter than the definition used by t -closeness [163].

2.3.2.5.3 Differential privacy

In this model, privacy protection is not considered a property of a dataset, but a property of a data processing method. Informally, it guarantees that the probability of any possible output of the anonymisation process does not change "by much" if data of an individual is added to or removed from input data. Consequently, it becomes very difficult for attackers to derive information about specific individuals and datasets are protected from membership, identity and attribute disclosure.

Differential privacy does further not make any strong assumptions about the background knowledge of attackers, e.g. about which attributes could be used for linkage. Instead, all attributes should be defined to be quasi-identifying [161][164].

2.3.2.5.4 Average risk

This privacy model can be used for protecting datasets from re-identification in the marketer model by enforcing a threshold on the average re-identification risk of the records. By combining the model with k-anonymity a privacy model called strict-average risk can be constructed. ARX further supports a variant which permits some records to exceed the risk threshold defined by k [161][165].

2.3.2.5.5 Profitability

This model implements a game-theoretic approach for performing cost/benefit analyses of data publishing to create output datasets which maximise data publisher's monetary benefit [161][166].

2.3.2.6 *Encryption*

Encryption of data [167] refers to a set of techniques that can be used between two parties to exchange information in a secure and reliable way. This means that data exchanged between them are:

- Encrypted: No other party can make sense of the data unless the other party has possession of the private key needed to decipher the information.
- Signed: The identity of a sender can be verified in a way that the recipient is sure that the sender is who she/he claims to be. Upon receipt of the message, the sender cannot deny that the message originated by her/him and cannot claim that the contents of the message were others than those received by the receiver.

Apart from data exchange, encryption can also be used when storing data. There are two kinds of encryption, the symmetric/private key and the public key encryption. In symmetric key encryption, the key for both encryption and decryption is the same and is shared between communicating parties. Public key encryption on the other hand relies on the existence of two keys: one public and one private. The public key is used for encryption and, as its name implies, is publicly shared. Anyone who wishes to send an encrypted message to the receiver, does so by using the public key. The private key is kept at the receiver's side only and is inaccessible by the public. The receiver uses the private key to decrypt the message.

In general, public key encryption is used more widely. It has several benefits over symmetric encryption, most notably the fact that communicating does not require the exchange of keys beforehand, a process that can be both time-consuming and introduce additional risk. The general principles of generating and using public keys for cryptography are depicted in the next three figures.

Figure 2-10 depicts the process of generating a private/public key pair. The subject uses a key generation algorithm which generates a key pair. Key pair generation algorithms typically make use of large random numbers; the main idea behind most public key cryptography algorithms is that though multiplication of two large numbers is a computationally easy process, the reverse procedure, factoring a product of two large prime numbers, is an intractable process. After generating a key pair, this key pair can be used for secure communication and digital signing as depicted in Figure 2-11 and Figure 2-12 respectively. In the first case, anyone wishing to communicate with the receiver, uses the

receiver’s public key to encrypt the message. Upon receipt of the encrypted message, also called a cipher, the receiver can use her/his private key to decrypt the message. In the digital signing case on the other hand, the sender uses her/his own private key to sign the message. The receiver, upon getting the signed document, uses the sender’s public key to obtain the original document. If any modification was made while the message was en-route, the verification procedure will fail.

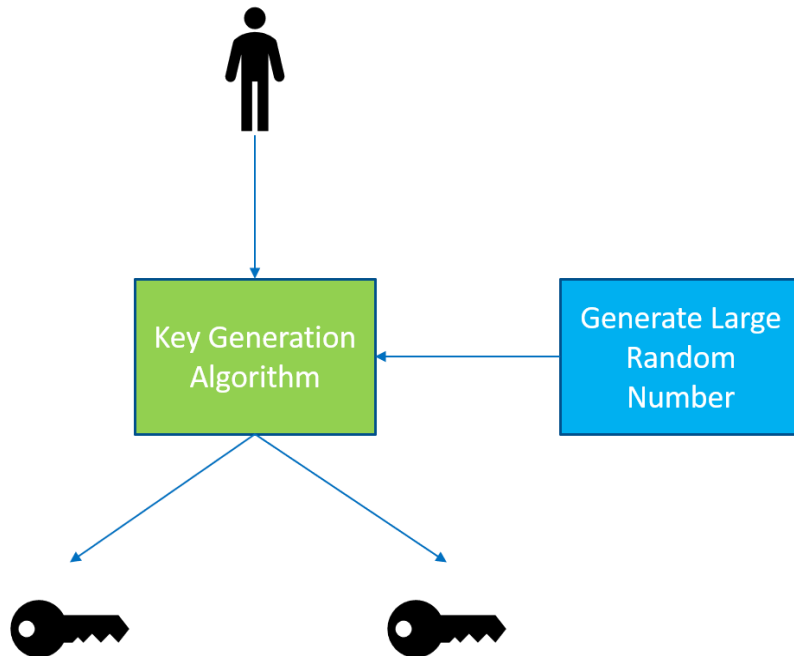


Figure 2-10 Public Private Key pair generation

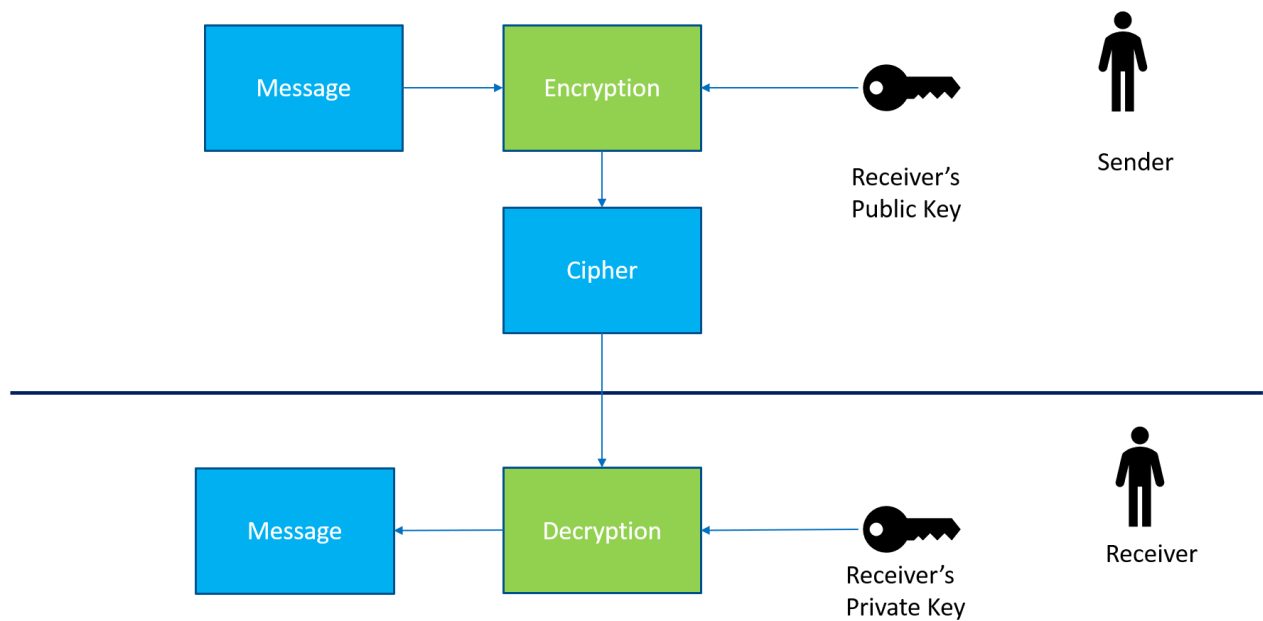


Figure 2-11 Encryption and Decryption of a message under the public key encryption scheme

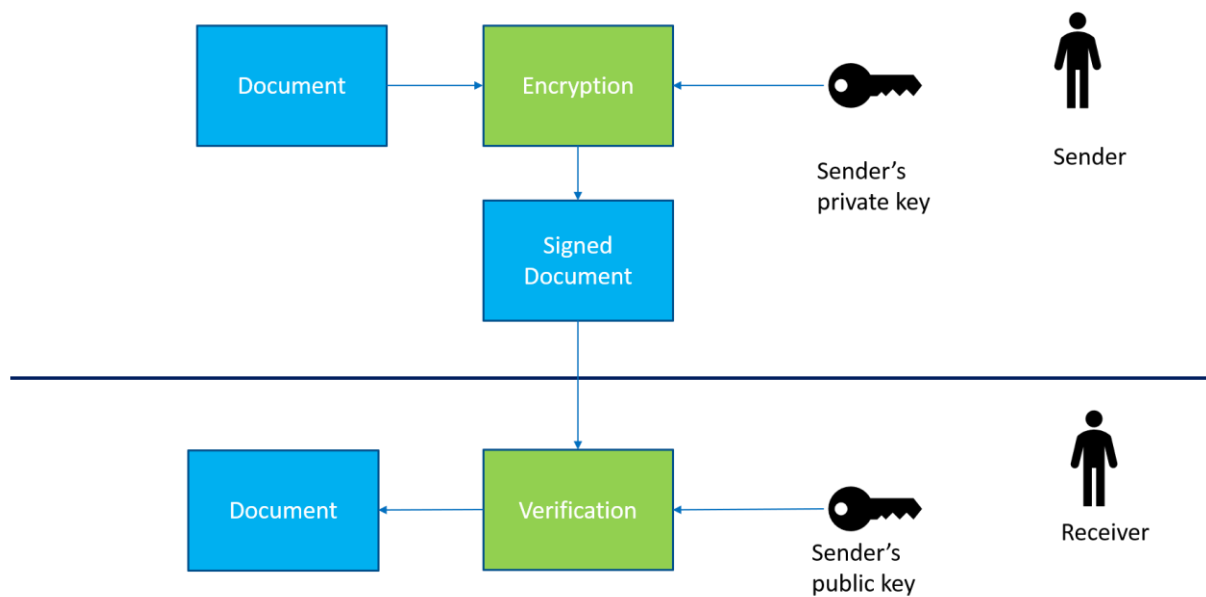


Figure 2-12 Digital signature under the public key encryption scheme

Although the main mechanisms of public cryptography are simple, when a large network of users are being connected and wish to exchange information securely, it can be quite difficult to distribute public keys in such a way such that the end users can be sure that the party publishing the public key is the actual owner of the public key. In order to distribute public keys securely, networks are usually called to implement what is commonly known as a Public Key Infrastructure (PKI). A PKI can create, distribute and revoke digital certificates, which are digital documents that prove the ownership of a public key.

The establishment of a binding between an entity and a public key is verified by the Certificate Authority (CA). The CA issues certificates, provided that registration was carried through correctly. The confirmation of correct registration is done by the Registration Authority (RA), which in many cases can be the same organisation as the CA. A third entity, called the Voucher Authority (VA), can vouch on behalf of the CA for the validity of the entity's information. Though a VA can reject a party's request if validation fails, it cannot issue or revoke the certificate; this is solely the responsibility of the CA.

Once the communicating parties can share public keys in a trustworthy manner, a security protocol can be used to encrypt communication. The Transport Layer Security (TLS) and the soon to be deprecated Secure Sockets Layer (SSL) are the most common examples of such protocols. Under TLS, before initiating exchange of data, a handshake between parties must first take place. In the handshake at least one party (the server) provides its certificate thus proving its identity. After validation, both parties generate and exchange keys and encrypt their messages using a private key algorithm. Figure 2-13 depicts an example of a PKI setup.

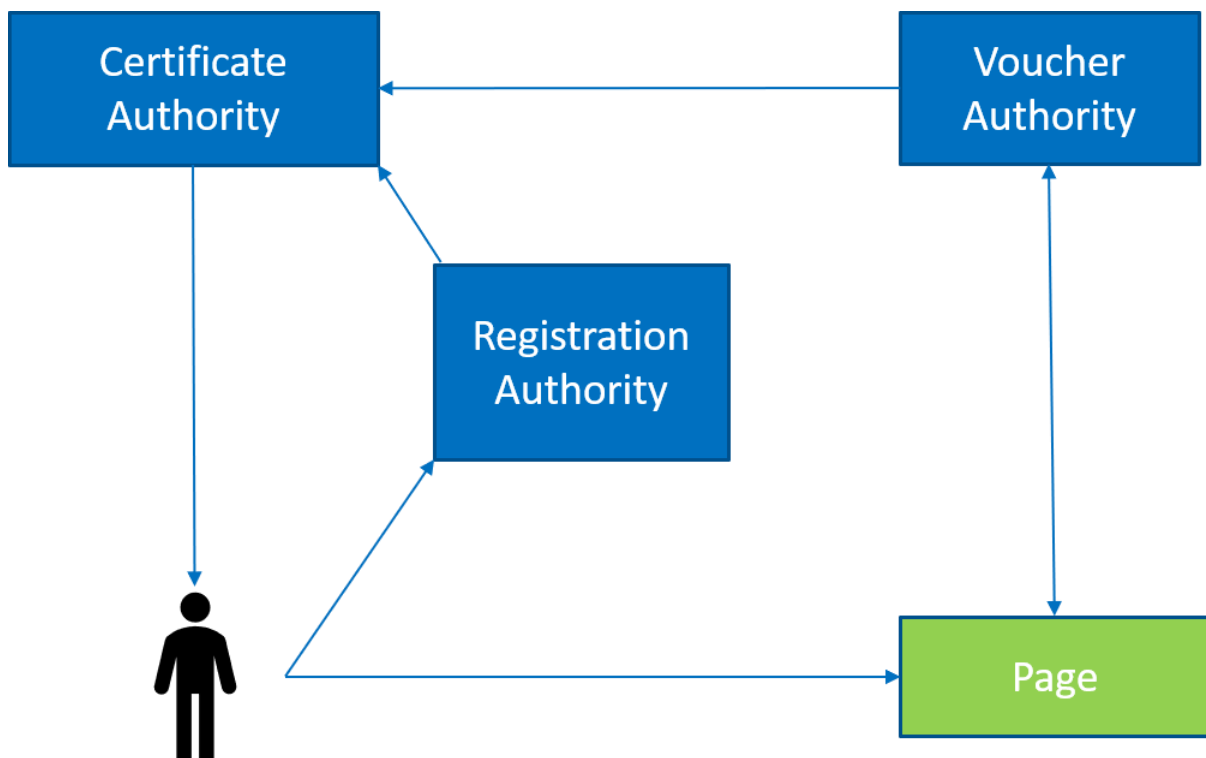


Figure 2-13 Typical PKI Infrastructure

2.3.2.7 Know Your Customer (KYC)

It is required by law for many legal entities to verify the identity of each potential customer/applicant before engaging to any transaction with her/him. Identification is a very old process and in the most common case, it involves the presentation of an official identification document (typically an id card) that is presented by the customer in person. The entity compares the visual appearance of the person to the one depicted in the picture of the id card and, if matched, confirms that the data contained in the id card are valid and the person is who she/he claims to be.

Such processes, typically called "Know You Customer" (KYC), though simple, may be difficult to implement electronically. Though no standards exist yet, the most straightforward approach is to perform the exact same steps required for physical identification, with any exchange of information being performed electronically. The verification procedure itself may be performed either by an operator that will compare the images provided, or automatically via an algorithm; even in the latter case, however, the algorithm's result needs to be verified by a physical person.

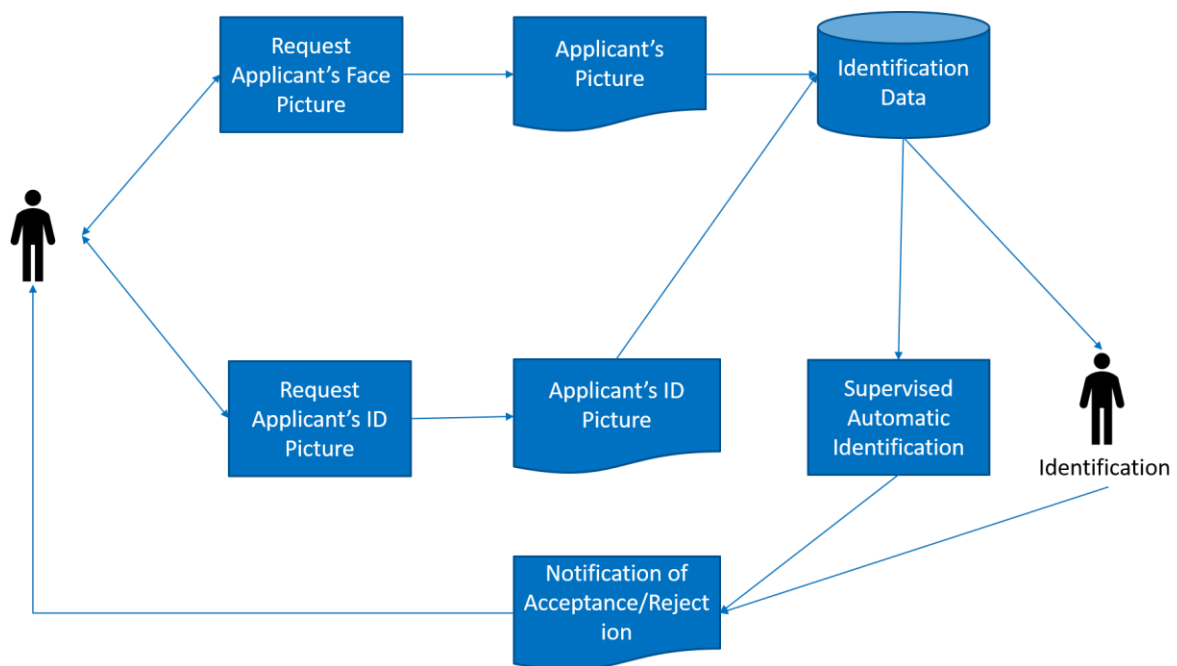


Figure 2-14 An example of a KYC Infrastructure Setup.

Figure 2-14 depicts a sample KYC Infrastructure Setup. The platform/application asks the applicant for a picture of herself/himself; for a mobile application this can be easily done via a selfie. Then a picture of an ID is required, which typically involves taking two pictures for the two faces of the identification card. The images are uploaded in the identification database; the identification is carried out either by an automated process or by the operator; both cases are depicted in the Figure. The applicant is informed accordingly for the result of the identification.

2.3.3 GDPR Compliance

Since the adoption of GDPR by all member states of the EU, the processes of storing, handling and transmitting data need to conform to specific rules in order to be considered GDPR compliant. Under the GDPR any entity that handles personal data is fully accountable for conforming to the GDPR and responsible for any data breaches.

Entities that store and process personal data are referred to in the GDPR as *data controllers*. The exact definition as this appears in Article 4 of the GDPR reads: "*controller*' means the natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data; where the purposes and means of such processing are determined by Union or Member State law, the controller or the specific criteria for its nomination may be provided for by Union or Member State law;"

An entity that does not collect personal data itself, but processes it on behalf of a controller is termed under the GDPR as *processor*. The exact definition of a processor as this appears in the GDPR is given as: "*processor*' means a natural or legal person, public authority, agency or other body which processes personal data on behalf of the controller;"

While listing the complete set of rules of GDPR is outside the scope of the present deliverable, controllers and processors should follow a set of guidelines. These guidelines are summarised as follows:

- **Legitimate interest:** To store data that contain personal information, the controller should have a legitimate and valid reason to do so. In the context of ChildRescue this is typically covered by the existing procedures of the participating organisations and are followed “as is” in ChildRescue. SoC for example has a legitimate reason to publish a child’s personal data under an Amber Alert, a procedure that is regulated by the existing legislature.
- **Provision of Consent:** Storing personal information of an individual requires that the said individual will give her/his consent.
- **Right to be forgotten:** An individual cannot only revoke his consent at any time, but she/he may request that any personal information already stored be removed upon request. In case of erroneous data, the individual may request correction of data.
- **Right of access:** Any processing of an individual’s data must be approved by her/him.
- **Data breaches:** When any compromise of the system leads to personal data exposure, the supervisory authority should be notified promptly within 72 hours both of the occurrence of the breach and of the details concerning which data have been exposed.
- **Anonymisation:** GDPR encourages controllers to perform pseudonymisation on data before they are stored.
- **Sharing by third parties:** Explicit consent must be given before personal data can be shared with third parties. The consent may be revoked at any time.

As already mentioned, the establishment of Legitimate Interest in the case of ChildRescue derives from the existing procedures, agreements between the participating organisations and the public sector and the existing legislature. Details can be found at the Regulatory Framework of ChildRescue that is documented in deliverable D1.2. Legitimate Interest will therefore not be covered in the present deliverable. Sharing by third parties will also not take place in ChildRescue, so that point is also omitted.

For each of the rest aspects, we will give a brief review and describe how ChildRescue will conform to its requirements in the following subsections. The details of the implementation of the proposed techniques will be given in Section 3.2.

2.3.3.1 Pseudonymisation and anonymisation in GDPR

Pseudonymisation in the context of GDPR is defined as “the processing of personal data in such a way that the data can no longer be attributed to a specific data subject without the use of additional information.” Under GDPR, data controllers are encouraged to perform pseudonymisation on any stored data. Data that are truly anonymised can be used, under circumstances, beyond the strict confines of “specific, explicit and legitimate purposes”. Furthermore, access, rectification and erasure need not be provided to data subjects if their data is anonymised.

As already mentioned, the GDPR makes a distinction between pseudonymisation and anonymisation by distinguishing between data that can be re-identified and data that cannot be re-identified. Re-identifiable data are those covered by the term “pseudonymisation”. Pseudonymised data are typically stored using data masking and/or encryption. When encrypted, parties that do not have possession of

the private key cannot discern the informational content of the original data set. When data masking is performed on the other hand, parties have access to data that have no relevance to the original data set. In both encrypted and data masking cases however, it is vital that the key and the data transformation matrix is kept somewhere, preferably in an external and secure data base. If they are not, the original data set may not be reconstructed in any way and the data set may not be classified as "pseudonymised" but rather as "anonymised".

Data sets that are truly "anonymised" are those for which the encryption keys and/or the transformation matrices have been deleted. Since there is no way to obtain confidential information from these data sets, anonymised data are not covered by the GDPR. However, anonymised data may still be very useful, especially when only statistical information is needed from a raw data set. In such scenarios, the data set may be transformed by using a combination of the techniques put forth in Section 2.3.2 so that although the data set is obfuscated, the statistical characteristics remain mostly intact.

ChildRescue will conform to the guidelines put forth by the GDPR by performing pseudonymisation to any personal data that are allowed to be stored under the Regulatory Framework or under the data subject's explicit consent. When data expire, or consent is revoked, ChildRescue will provide the functionality to either remove the pseudonymised data or anonymise them.

2.3.3.2 Data breaches

Under the GDPR, when a data breach results to exposure of a data subject's personal data, the data controller needs to notify the supervisory authority no later than 72 hours after the breach occurred (or give reasonable justification in case more time is needed). Quoting from the GDPR, the notification needs to:

- 1. describe the nature of the personal data breach including where possible, the categories and approximate number of data subjects concerned and the categories and approximate number of personal data records concerned;*
- 2. communicate the name and contact details of the data protection officer or other contact point where more information can be obtained;*
- 3. describe the likely consequences of the personal data breach;*
- 4. describe the measures taken or proposed to be taken by the controller to address the personal data breach, including, where appropriate, measures to mitigate its possible adverse effects.*

When the above data cannot be collected at the same time, partial information may be given without delay as missing information is collected.

ChildRescue shall incorporate and implement all best practices regarding safe registration, authentication, authorisation and data storage. Databases will be audited and monitored so that the details of every access are logged. In the case of a data breach, mitigation measures will be carried out, depending on the extent and nature of the breach. In the case that the breach occurs in anonymised and/or pseudonymised data, access to the offending addresses will be blocked, whereas in case that a data breach occurs in raw data (e.g. pseudonymised data plus the database that contain the re-identification data), databases will be quarantined, and access will be restricted only to trustworthy parties. In all cases, the supervisory authority will be notified with all the information

required by the GDPR. In the case of SoC and Hellenic Red Cross, the supervisory authority is the Hellenic Data Protection Authority (DPA)¹⁷. In the case of Child Focus, the supervisory authority is the Commission de la protection de la vie privée¹⁸.

2.3.3.3 *Right of access and right to be forgotten*

The GDPR dictates that before any processing of a data subject's data takes place, the data subject must give her/his permission. Moreover, the data subject shall be aware at any time of whether her/his data are being used for procession. If personal data are indeed being processed, the data subject must be able to know the purposes of the processing, the exact kind of data that are being processed, the parties to which the results of the processing will be disclosed and whether the data is used in any automated decision-making. In any case, the data subject will be able to revoke permission to the controller to perform any processing on her/his data and even to delete any personal information already stored.

In the case of ChildRescue data subjects can be distinguished in two categories. The first category concerns individuals like missing children or parents of unaccompanied minors which are part of the stakeholder's group of ChildRescue. The second category consists of registered users to the ChildRescue platform.

Concerning stakeholder's personal data, these are absolutely needed only as long as the case is open. Often, as is the case with an Amber Alert, these data are released publicly and, in this sense, cannot be practically deleted since they have already been disseminated to the entire world. In any case, ChildRescue will delete any sensitive information that is stored locally in mobile devices and will remove sensitive information from any part of the platform that is displayed publicly, once the case is over. Users will also be notified that they should delete any data that they, by their own actions, have stored, such as phone screenshots containing the child's face. The users will be notified both via an application notification and via an e-mail (in case they do not longer use the application) that they are required to delete any relevant data that they may have stored. In case however, a user stops using the application and has, without updating her/his profile, altered her/his communication detail there will be impossible to contact this user. Under this circumstance, Article 19 of the GDPR applies, which covers cases when contacting the holder of the data is impossible or requires a disproportionate effort. It is to be noted that the above applies to data that are shared publicly. The subject's data that are stored in the organisation's premises follow a different policy. This happens because these data contain part of the cases both ongoing and closed. For ongoing cases the reason for keeping data is obvious; however even in close cases the data may be needed again. This can happen when a similar case is opened (e.g. the same child disappears again), or if a legal authority (e.g. the Police or District Attorney) requires for data from past cases for some legal reason (e.g. to press charges on a person that is suspect of being part of an abduction network that had a part to some of the past cases). Thus, stakeholder data that are part of past and ongoing cases, are never deleted. The same applies to stakeholder data that were never intended to be disseminated publicly, such as unaccompanied minors. In this case, the personal data remain solely within the data controller, which

¹⁷ <http://www.dpa.gr/>

¹⁸ <http://www.privacycommission.be/>

in ChildRescue consists of SoC, Hellenic Red Cross and Child Focus and are similarly kept for an indefinite amount of time.

Regardless of the extent to which stakeholder's personal data have been disseminated to the public, ChildRescue will provide the right of access and erasure by performing anonymisation on undisclosed personal data. In case the stakeholder expresses her/his desire to revoke access and/or erase the data, the already stored pseudonymised data will be converted to anonymised data by moving the relevant entries from the pseudonymised data set to the anonymised data set.

2.3.3.4 User Identification

When authorities examine whether to disclose children's personal data to the public, they weight in the pros and cons and decide to the best interest of the child. There is always the danger that disclosed information may be used with malicious intent; this danger however may be deemed acceptable depending on the case. As ChildRescue will be used to broadcast information concerning ongoing cases of missing children, the same danger also lurks; namely that ChildRescue will be used by abusers to locate missing children.

ChildRescue supports two kinds of registration, anonymous and named. In anonymous registration, the user will receive the public notification of the authorities, like public alerts and will of course be able to contact the authorities in case of a finding; however, she/he will not be able to participate in the investigation via ChildRescue by providing evidence to the platform. In that sense, ChildRescue will not provide to anonymous users any more information than that already given to the general public and already deemed acceptable by the authorities.

For named registration, ChildRescue will require that users will undergo an identification procedure which will require from them to prove their identity before being able to use the platform. Section 2.3.2.7 describes how "Know Your Customer" techniques can be employed to identify the physical identity of users online. ChildRescue will use such an infrastructure and will store permanently the relevant data in case they will be required by the authorities. This circumvention of the "Right to be forgotten" of the GDPR is justified under Article 23 and will clearly be stated to the User before using the platform.

2.3.3.5 Consent

Storing and processing any kind of personal data requires the explicit consent of a data subject. It is therefore necessary to provide to each user a consent form written in clear language that definitely describes both the kind of personal data that are going to be collected and the purposes of data processing. Since the exact kind of information collected and processed may depend on the profiling and decision-making algorithms that are going to be developed in the context of ChildRescue, it is not possible at this moment to define exactly the kind of data that the consent form will cover. However, the consent form will clearly indicate:

- The identity of the data controllers and data processor. In the case of ChildRescue, controllers and processors coincide and consist of SoC, Hellenic Red Cross and Child Focus. In case an organisation wishes to use a third party for processing after commercialisation of ChildRescue, the consent form will be updated accordingly.
- The kind of personal data that will be collected from each data subject.

- Which of the data are going to be used for profiling, decision-making and statistical analysis purposes and in what way.
- Clear indication that the data subject can ask and be granted permission to view what kind of personal information is stored at any time.
- Clear indication that the data subject may at any time revoke her/his consent and that after revocation, their personal data will be removed from the platform within a reasonable time. This only concerns data that do not constitute legal evidence; the case for legal evidence data is considered in the next paragraph.
- Clear indication that in case ChildRescue makes any change to the described way that personal data are handled, the subject will be notified and consent will be asked again.

The above GDPR requirements are typical with any application that uses personal data; ChildRescue however is not typical in the sense that data subjects are part of an official police investigation. Not only may it be required that the police contact them directly, their data may also constitute legal evidence and as such may not be completely removed. This exception will also be stated in the Consent form and the exact kind of data that will be exempt from their right of erasure will also be clearly stated.

It must be noted that consent covers mainly the users of the platform, whether these are anonymous citizens or registered users. Subjects of ChildRescue however include also missing children, for which an Amber Alert has been issued, and unaccompanied minors.

Missing children cannot of course give consent for storing, or even publishing personal data. These cases follow a different procedure with consent being given by the appropriate authorities like the District Attorney or by setting a legal basis like the "protection of the vital interests of the data subject" (GDPR art, 6.1.d). ChildRescue will only disseminate Subject's information that is already present in the Amber Alert and have thus been already been made publicly available by the Authorities. Though consent is not given explicitly to ChildRescue to also store or disseminate this information, doing so will facilitate the citizens to act as social sensors thereby improving the chances of locating the child. As long as this processing increases the chances of finding the child, consent may not be explicitly asked to process the subject's data. This is covered by Article 6.1.d as well as recital 46 of the GDPR. In order to better safeguard the interests of the missing child, additional information may be protected. For example, the last name of the child is seldom used when locating a child; in the vast majority of the cases the child is recognised by the picture. Thus, the last name of the child may be omitted by the notification pushed by the ChildRescue Platform.

Unaccompanied minors on the other hand need to register in shelters, where they also provide consent for the usage of their data by the corresponding Legal Entity that is responsible for their sheltering (e.g. Hellenic RED CROSS). ChildRescue needs to collect these data too, in order to be able to perform predictions regarding the possible location or route of an unaccompanied minor, in case she/he has disappeared. The data needed are exactly the same as that, that the original consent form already contains. However, explicit consent must be given to ChildRescue to also use these data. Therefore, a copy of the consent form, applied to ChildRescue Platform, will also be given to the subjects.

For the users of the ChildRescue platform and app, an example of a consent form can be found in *Annex V: Sample Consent Form*.

2.3.4 Discussion and key-takeaways

One of the main conclusions of the discussion followed through the previous subsections, is that no single approach can guarantee anonymity of data for every case. Various factors, such as the distribution of data and the similarity of the dataset to a generic dataset taken from the populace can lead to “anonymised” data sets which, although possess good values in representative metrics (such as k -value), are however vulnerable to reverse engineering attacks. Moreover, the extent to which an attacker possesses inside information regarding the dataset is something which cannot be evaluated with accuracy in many cases.

For the purposes of ChildRescue, we must distinguish the various cases both between data subjects and between recipients of data. Table 2-10 summarises the broad categorisation of data subjects and data recipients for ChildRescue. For the case of children data, as these are disclosed to the public, it is of course easy to identify the child; this is indeed the meaning of a signal like Amber Alert. While ChildRescue does not anonymise these data, it makes sure that only the bare minimum of the required information is transmitted publicly. These are the information contained in the Amber Alert in the case of Registered and Unregistered Users of the platform plus the feedback or information provided by other users in the case of Volunteer Users. Volunteer Users have been registered with the respective organisations and are authorised to assist the authorities in ongoing investigation. All relevant data are in any case deleted from the user’s devices as soon as the investigation is completed.

Table 2-10 Data subjects and data recipient summary for ChildRescue

Data Subject	Data Controller	Data Recipient
Child	ChildRescue Organisation	Unregistered User
		Registered User
		Volunteer User
User	ChildRescue Organisation	ChildRescue Organisation
		Authorities
		User

For the case of personal data that are collected by the users, the anonymisation techniques covered so far will be used as to provide the desired level of anonymisation. It will be assumed, that the characteristics of the user’s that are using the ChildRescue application do not differ significantly from those of the general populace. Therefore, no specific consideration will be given to compute custom metrics. In particular, information that is not necessary for processing purposes will not be collected, or it will be pseudonymised with the re-identification data stored separately in an off-line database in case it is information that may be needed by the authorities in future investigation. Information that is needed for processing, will be transmitted anonymised with a minimum value of $k=3$ for k -anonymity. The consortium will also try to obtain general populace data from voting registries; it this is the case the δ -presence metric will also be taken into account when transmitting anonymous data.

3 Methodological Approach

The issue of missing children and unaccompanied migrant minors affects European citizens not only at a national level, but also at the level of the European Union. Despite the huge efforts from organisations, governments and EU or International institutions to effectively address this issue, many challenges still remain open. One of them is the insufficient aggregation, explanation and analysis of available information coming from a number of heterogeneous sources.

The proposed methodology aims to tackle this exact issue, by designing a proper way of collecting profile data, suggesting ways of safekeeping them according to EU and national laws, and use the power of data analytics to make useful assumptions and predictions that could help in the tracing of a missing child.

The investigation of the related landscape can be considered as the initial, but also necessary, step in order to form a sound and effective methodology. The theoretical background on human behaviour profiling was examined from both the perspectives of social and computer science. In addition, computational methods and best practices were presented in the context of multi-source analytics, while anonymisation techniques, which ensure data protection and privacy, were described with technical details. All these compose the foundations upon which the ChildRescue methodology is built.

The proposed approach is therefore described by three axes that operate, more or less, in parallel:

- Setting up an appropriate behavioural and activity profile based on the results of the conducted interviews with experts, the literature review, and the suggestions of the pilot partners, that covers the most important aspects of a missing child case.
- Ensuring privacy and anonymisation, in compliance with GDPR, since each profile includes several personal and sensitive details.
- Performing predictive analytics, using the aforementioned profiling on past cases, as well as other informative inputs, such as open data or social media, in order to enhance the investigation processes in a timely and effective manner.

Additionally, a semantic data model is defined that unifies the information available in different institutions in an attempt to enhance cross-organisation, as well cross-border, cooperation.

3.1 Setting up a Behavioural and Activity Profile

In order to review the choices of theories and assumptions formed from the analysis of the research, interviews with experts in the field were conducted. The interviews underlined the importance of the context of the disappearance of the child for the quick and successful resolution of the case. In accordance with the proposed differentiation by Missing Children Europe, ChildRescue follows the classification into runaways, abductions by third persons, international parental abductions and missing unaccompanied migrant minors, with the addition of lost, injured or otherwise missing children for cases with an unknown outcome. For cases of lost, injured or otherwise missing children no theoretical approach or specific important information can be pointed out due to the lack of information in these cases and the resulting strong uncertainty. Consequently, all available information should be collected and processed to help create the best available profile.

Due to the vastly different circumstances of missing children cases and the varying degrees of agency of the child involved, which can critically influence the success of a behaviour prediction, it seems

sensible to utilise different theories and apply a risk assessment-tool, first in the process of gathering data on the individual case, which also establishes a probable link to one of the categories. After the initial categorisation, different information can be of specific importance for the creation of the behavioural profile of the child. For example, asking for the intended location, as identified in the interviews by the Hellenic Red Cross, seems vital for cases of runaways or unaccompanied minor migrants, but less so in cases of child abductions as the missing child will not necessarily be able to influence that choice. The findings of the interviews have yet again stressed the importance of creating the most comprehensive profile possible while stressing the importance of certain pieces of information for specific case types. For example, the use of dating apps is an important piece of information for cases of runaways and missing unaccompanied migrant minors – especially for young girls – but offer little insight in cases of parental abductions. Further insights from the interviews can be found in section 3.1.2 and the complete logs in *Annex III: Interviews*.

3.1.1 Forming Profiles based on Theories

The choices of theoretical approaches that substantiate the practical response to the cases, cannot be unified, but rather depend on the individual case and its categorisation. All cases require a timely response that should be initiated through the risk assessment and followed up by a case-specific approach of uniting all existing knowledge on the missing child through interviews with family members and friends as well as a review of the existing casefile and an analysis of their social media profiles. The case specific approach then depends on the probable circumstances of the case. While a wrong categorisation can potentially include unnecessary information or lead to wrong behavioural predictions, the procedure also offers the possibility of a timelier response and a more efficient recovery of the child which can reduce their risk of victimisation. Thus, the proceeding explained in section 2.1 stands as valid. In cases of runaways, when the child has a relatively high amount of agency, the approach of analysing the social network of the child is most promising as it might reveal key persons or locations of interest to the child. Furthermore, utilising subcultural theory and the interview results with members of the NGOs working with street youth in Germany, it can be argued that so-called 'hotspots' for runaway children exist within the cities, which can be identified through the expertise of social workers, interviews with peers or existing case files on earlier recovery locations of the child. Additionally, Activity Theory can be applied to understand the behavioural patterns of the child within its environment. In cases of missing unaccompanied minors who left the shelters willingly, subcultural theory and activity theory can also be applied. Additionally, collective behaviour theory can be enlightening as the children live in shelters and both the interviews as well as the research have highlighted the importance of 'group intelligence', meaning the information that is passed through the trusted channels of other minor migrants, for the decision to relocate.

On the contrast, in cases with little agency on the child's side, such as parental abductions or third-person abductions, the prediction of the behaviour of the child based on their previous behaviour is more difficult and the implementation of algorithms focussing on the child's profile can be restrained in its effect. Yet, in cases of parental abductions, the location of the child might be revealed on its social media site and the social network theory can be applied more generally to identify key contact persons for both the parent and the child such as other family members or new partners, which could be sources of information on or be identified as the location of the missing child. In cases of third-person abductions, an analysis of the social network of the child could reveal the perpetrator in cases

where a child was specifically targeted. Thus, the proposed choice of theories was further supported by the results of the interviews and should be applied after the initial risk assessment and categorisation of the individual case.

Table 3-1 Indicators and Theories applicable to cases

Type of missing children case	Applicable theories	Important indicators for the creation of the profile
Runaways	Subcultural Theory, Activity Theory, Social Network Analysis	Timing of incident (e.g. last day of school), information on family situation, history of abuse at home, current phone number, online social media profile (romantic involvement abroad, potential radicalisation), past recovery locations, mental health state, police warrants
Third-person abductions	Social Network Theory, Victimology	Online social media profiles, contact to non-family member adults, character of the child (trusting toward strangers, safe spaces)
Parental abductions	Activity Theory, Social Network Analysis	Time of disappearance, location of family abroad, police warrant
Missing unaccompanied migrant minors	Subcultural Theory, Activity Theory, Collective Behaviour Theory	Intended destination, time of disappearance, immigration status, country of origin/trafficking risk, accommodation history, data on current events in the area, existence of profile on dating app, mental state, police warrants
Lost, injured or otherwise missing child	Not applicable due to lack of information on the case	All indicators considered until category of case can be determined

Cases of lost, injured or otherwise missing children are hence the most complicated as no theoretical approach or a prioritisation of indicators can be completed until the case can sensibly be allocated to one of the other four categories. Until this point, all available information should be considered to be of high importance and be collected.

3.1.2 Prioritising profile indicators

The indicators that could be utilised in data analytics are based on the results from both the interview data as well as best practices found in documents of the partners and the state-of-the-art research review that was conducted for ChildRescue.

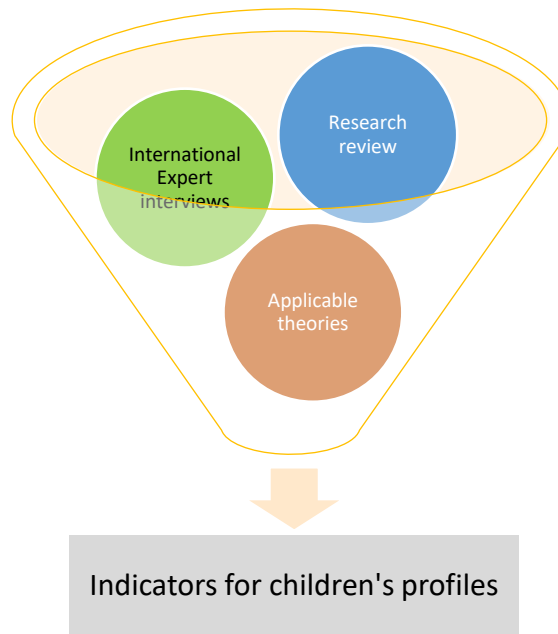


Figure 3-1 Different sources contributing to the indicators of the profiles

Different indicators were identified, but the clear re-emerging theme was family conflicts and issues in the structures of the families (due to problematic divorces). This should thus be prioritised as children with challenging homes are also less likely to return and children involved in custody disputes could be located at their other parent's home. Adapting the profiling methods of the FBI used in the search for offenders for a prediction of the behaviour of missing children [168], the following profiling characteristics emerge when disregarding all information related to criminal offending:

Table 3-2 Applicable information from FBI offender profiling for missing children cases

Basic information extracted from the FBI offender profiling

- Emotional state
- Socio economic standing
- Hobbies
- Lifestyle
- Social/sexual competence
- Age
- Gender
- Vehicle
- Place of residence (so far)
- Behaviour
- Norms and believes (cf. *ibid.*)

In the interviews conducted by Smile of the Child, the Hellenic Red Cross and the Frankfurt University of Applied Sciences, the following indicators were identified:

Table 3-3 Different indicators named in the interviews

Smile of the Child	Hellenic Red Cross	Frankfurt University of Applied Sciences
General info (Full name, gender, phone number, school, nationality, language)	General info (name, gender, nationality)	General info (full name, gender, phone number)
Age	Age	Age to determine likelihood of runaway vs. kidnapping
Family situation	Immigration status	Family situation
Health status (substance abuse, disease, emotional and mental state)	Health status (mental state)	Health status (substance abuse; disease; emotional and mental state)
Time the child was last seen	Time of disappearance	Timing
Key persons (family, friends, others)	Context of recovery site (red-light district, border ...)	Access to dating apps
Location the child was last seen	Intended location	Police warrant

This overlaps partly with the findings from the literature review, namely:

Indicators in research literature

- Family structures ([2], [11], [169] / Family situation (divorce, pubescent conflict, physical/ mental/ sexual abuse, neglect, substance abuse)
- Risk of socioeconomic status [7]
- Substance abuse
- Age (to know if child ran away or was taken)
- Timing (last day of school)
- Announced leaving (unaccompanied minor migrants)
- Emotional/mental state (risk of suicide or harm of others)
- Limited access to alternative help resources/ forced restraint from seeing family or friends (Crosland et al. 2018, S. 38)[7]

While the categorisation of cases can lead to a prioritisation of certain indicators in specific cases, all available information should be registered for all cases, as there is vital basic information that is needed in all cases and the classification of the case might change from the initial risk assessment to

a later stage of the inquiry and information that is deemed atypical can still turn out to be essential for the case. Additionally, the nature of data that is helpful for all cases can be divided into socio-demographic information on the child (child profile) and situational information on the case (case profile), as presented in table:

Table 3-4 Most important profile indicators for ChildRescue

Socio-demographic information on child (Child Profile)	Situational information on case (Case Profile)
Full name	Financial status (carrying money, bank account)
Age	Timing (last day of school/date of deportation)
Gender	Place last seen
Appearance (height, weight)	Clothes worn on day of disappearance
Health status (pre-existing disease)	Emotional/mental state (substance abuse, suicidal tendencies)
Relationship status	Access to dating apps/ social media platforms
Police warrants	Hobbies/interests
Key contact persons/access to other persons	Family structures/ issues

However, this distinction is not as clear-cut as it might appear. While family structures are listed as a situational information due to their importance in categorising a case properly, it could also be seen as part of the socio-demographic information on the child, such as their living situation (changing homes due to a shared custody agreement, living with non-parental guardian etc.). This structure should thus not be seen as a rigid construct, but rather a construct that helps to grasp the complex situation of cases of missing children in theory.

3.1.3 Discussion and Limitations

The identified information from the interviews as well as the research review should be included in the profiles of the children and thus reasonably can be incorporated into the algorithms. Applying the indicators to the case will enable a timelier recovery of the child. However, in order to increase the response time, the initial categorisation of the case and the resulting use of the indicators also has to be completed in a timely manner. The interviews conducted with different stakeholders in the field internationally have shown a great overlap in the information that was identified as crucial.

Thus, in order to evaluate the behaviour of the missing child and to make sensible predictions on potential places of interest, the listed indicators should be analysed. Additionally, the lifestyle and hobbies that were not specifically identified for missing children cases might also be enlightening as missing children may have contacts through those and could be "couch-hopping" there. Further, the norms and beliefs can have an influence on the behaviour of the child and their likeliness to seek help from official state actors, which thus should also be reflected in the input form of the ChildRescue platform.

The biggest challenge that ChildRescue will face when creating the profiles of missing children is the lack of or false information delivered by parents and friends. In order to correctly predict the behavioural pattern of the child and for ChildRescue to minimise the time of the recovery of the child, the necessary information needs to be available and correct. In some cases, this can prove to be challenging. This is especially true in cases of unaccompanied minor migrants who have not spend a long time in their shelter before going missing and thus do not have close friends who could give a testimony and the parents are not reachable due to a lack of contact information or unwillingness to cooperate. In these cases, the initial categorisation might be difficult as it could both be runaway or kidnapping cases.

Furthermore, if key contact persons cannot be reached, the timeliness of the response of ChildRescue is limited as it requires the initial assessment through the child's profile. Additionally, the complete lack of trustworthy information can hinder the successful creation of a profile, which renders a prediction of their behaviour very challenging to almost impossible. The same issue can be faced in cases of missing children with uncooperative parents and friends, who do not comply with the protocol of filling in the information sheet. Additionally, friends of runaways, who are teenagers themselves, might feel like they betray their friends if they volunteer information on their whereabouts, if the missing child does not want to be found, resulting in a lack of trustworthiness of given information.

Also, due to the different nature of the relationship with parents and friends and the resulting different behaviour of the child in the interaction with these different actor groups, the testimony of the friends and the parents can differ greatly in relation to the character and behaviour of the child. This can make the correct prediction of behavioural patterns of the child more difficult. In cases of kidnappings the behaviour of the child is only relevant if there was online grooming of the perpetrator on the child's social media profiles as it may point to the location of the child.

In cases of spontaneous kidnappings of children due to an opportunity to take an unsupervised child, the profiles based on the child's behaviour unfortunately won't be successful in aiding the recovery. Still, in cases in which the child has some degree of agency, the methodology of creating profiles of the children based on their characteristics and behavioural patterns can be applied successfully, rendering the approach of ChildRescue especially helpful in cases of runaways, which also make up the majority of the caseload of missing children.

3.2 Ensuring Privacy and Anonymisation

In order to identify the specifications of the privacy and anonymisation framework ChildRescue, the following deliverables were used as input: "D1.1 – User Requirements", "D1.2 – Regulatory Framework for Data Protection, Privacy and Ethical Issues" and "D1.3 – ChildRescue Integrated Methodology, Release 1". D1.1 and D1.3 provide the functional and non-functional requirements and use cases that the privacy and anonymisation framework need to address, while D1.2 provides the main outlines that the framework needs to follow in order to be compliant with the GDPR.

The following tables show how the functional and non-functional requirements and use cases from D1.1, as well as, the procedures described in D1.3 are mapped to the technical requirements of the ChildRescue Privacy and Anonymisation Framework.

Table 3-5 Requirements mapping to technical requirements of the Privacy and Anonymisation Framework

Requirement	Description	Framework requirement	Relevant data
FR_5	ChildRescue must allow users to anonymise data	The framework must provide anonymisation of the data	User data, case data
FR_11	ChildRescue must provide a user with the ability to register with a pseudonym	The framework must provide pseudonymisation	User data
NFR_4	ChildRescue must be able to protect any personal data stored	The framework must provide anonymisation and security of the data	User data, case data
NFR_5	ChildRescue must be able to encrypt any data transmitted	The framework must provide encryption of the exchanged data	User data, case data, other data
NFR_7	ChildRescue must be able to archive any significant data required by the user	The framework must provide anonymisation and encryption of the data	Case data

Table 3-6 Use cases mapping to technical requirements of the Privacy and Anonymisation Framework

Use case	Description	Framework requirement	Relevant data
VU.08	to send feedback in the form of text, image or video the Organisation concerning a missing child by filling-in an appropriate form	The framework must provide encryption of the exchanged data	User data, case data, other data
SU.01, RT.01, VT.01, FM.01	to be able to specify some basic, optional personal profile information during the initial registration process	The framework must provide pseudonymisation of the data	User data
FM.02	to be able to fill in extended profiling information about an	The framework must provide pseudonymisation of the data	Case data

	unaccompanied migrant minor residing in the premises		
CM.01	to be able to fill in the basic information about a new missing child case	The framework must provide pseudonymisation of the data	Case data
CM.02	to update the information of a case with more details or specify crucial information (like location last seen, latest info gathered, etc.)	The framework must provide pseudonymisation of the data	Case data
OO.8	to have an overview of the duration of all historical cases and the human resources assigned on each one	The framework must provide pseudonymisation of the data	Case data

Table 3-7 ChildRescue processes mapping to technical requirements of the Privacy and Anonymisation Framework

Process	Description	Framework requirement	Relevant data
Preparation and Profiling	Construction of a multilayer profile of the missing person	The framework must provide pseudonymisation of the data	Case data
Coordination and Collaboration	Real-time sharing of information and communication among team members	The framework must provide pseudonymisation of the data	User data, Case data
Action	The investigation process followed after report of a missing child or a tracing request for an unaccompanied minor	The framework must provide pseudonymisation of the data	User data, Case data
Archiving	Case closure in a secure and a privacy-respectful	The framework must provide anonymisation and encryption of the data	User data, Case data

	manner		
--	--------	--	--

In the aforementioned tables the user data refer to the personal data of the end users, while the case data contain the missing children/unaccompanied minor personal information and any additional case or evidence data.

Based on the above tables ChildRescue should provide anonymisation and pseudonymisation techniques for the user data and the case data. Additionally, encryption should also be implemented during the exchange of information among the various stakeholders through the ChildRescue Platform. The ARX Anonymisation Framework addresses the aforementioned requirements, while Public Key Infrastructure covers the needs for encryption of the data. A detailed presentation of the methodologies and the technologies that are going to be used in ChildRescue is provided in the following sections.

3.2.1 Pseudonymisation and Anonymisation

As already mentioned, pseudonymisation and anonymisation refer to any technique that obfuscates or encrypts data with the process of pseudonymisation being reversible, while that of anonymisation being irreversible. Both techniques can, in the context of ChildRescue, employ encryption to perform the required transformation; the encryption infrastructure will be described in section 3.2.3, here we only present the process by which anonymisation is performed.

The overall architecture of the module performing pseudonymisation and anonymisation is depicted in Figure 3-2. The module includes the following components:

- **Consent database:** A database which stores the data subjects who have provided consent to the ChildRescue platform.
- **Framework database:** A database which contains the PII's of all the data subjects whose data disclosure is covered by the Regulatory Framework of D1.2.
- **Re-identification database:** A database which contains the original data of the data subjects or other data which can be used to match the pseudonymised (or anonymised) data to the data subjects. These data need to be pseudonymised (or anonymised) and their access is restricted only to the authorised personnel.
- **Exposed database:** A database which contains the pseudonymised data which are accessed and disseminated to the various parties which use the ChildRescue platform.
- **Pseudonymisation:** A component that will perform pseudonymisation transformations on the data
- **Anonymisation:** A component that will anonymise the data
- **Data adapter:** A software component which is responsible to implement the pseudonymisation of the data.

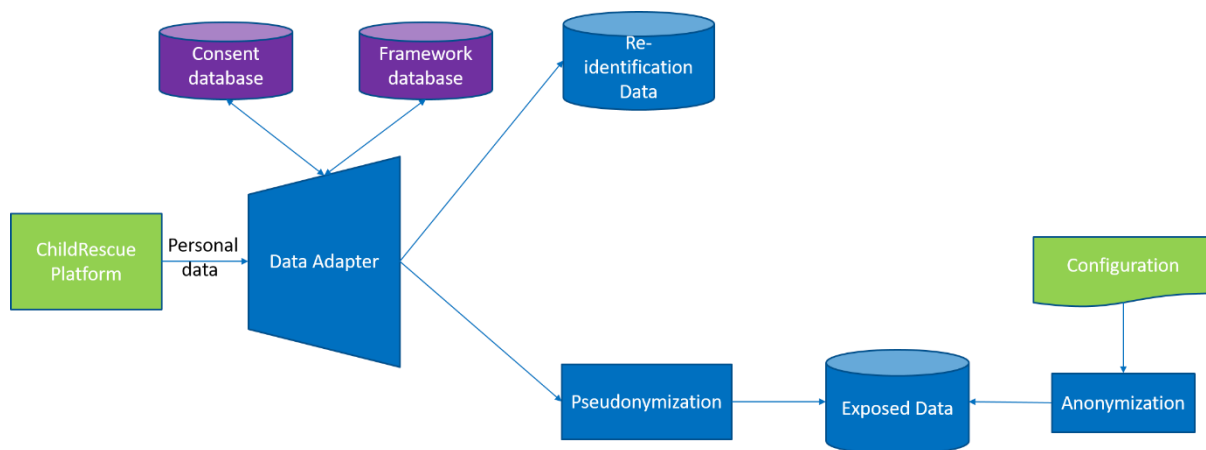


Figure 3-2 Architecture of the Pseudonymisation module

The pseudonymisation process is briefly described below:

When collecting personal data, the Data Adapter will query the Consent database and the Framework database. The consent database will have stored a map of all subjects that have provided consent to the ChildRescue platform. As mentioned above, the Framework database will contain the PIIs of all data subjects whose personal data disclosure is covered by the Regulatory Framework as it is described in D1.2 (e.g. a child for which a distinct attorney has provided permission to issue an amber alert). If confirmation from the consent or the framework database occurs, the Pseudonymisation Module will perform pseudonymisation on the data; it will store the pseudonymised data in an open dataset that can generally be accessed by parties being in communication with ChildRescue and will store the re-identification data in a separate database; the Re-identification database. The Re-identification database will not be publicly accessed but will be used and maintained by each of the data controller's explicitly trained personnel. When re-identification is needed at run-time (e.g. when the e-mail of a user needs to be verified), the Pseudonymisation Module will communicate with the Re-identification database to obtain the original data; apart from this case, access to the re-identification database will be restricted.

After storage, an extra *Anonymisation* module will provide the functionality of generating anonymised data from the exposed data set. It will be based on the **ARX Framework**¹⁹ and will produce a data set with high *k-value*, *l-diversity* and *t-closeness* parameters. In case the platform operator imports a population table, the *Anonymisation* module will also produce a low value of δ (for the specifics of *k-value*, *l-diversity*, *t-closeness* and δ -difference, please consult Section 2.3.2.1). The anonymised data set will contain all useful information regarding user actions and cases and can still be used to compute analytics and provide useful feedback. Since data subjects cannot be de-identified from the anonymised data set, it can be stored or archived regardless of the status of consent forms.

In case that a subject is removed from the framework database or a consent is revoked, the Pseudonymisation Module will remove for this subject the re-identification data from the re-identification database. The pseudonymised data will be automatically converted to anonymous data

¹⁹ <https://arx.deidentifier.org/>

upon this removal, so they can still be stored in the Exposed database. Upon revocation of consent, the deletion of re-identification data may take some time due to the system having to poll the consent database and the technical expert receiving the notification to delete re-identification data. This will be explicitly noted in the consent form.

The pseudonymisation module will perform a combination of techniques as these were documented in Section 2.3.2. The administrator of the platform will be able to define which transformations are needed to ensure proper pseudonymisation or anonymisation.

The set of transformation offered will consist of both one-way hashes²⁰ and two-way encryption (possibility to encrypt and decrypt the data) as well as all the data masking techniques, except from shuffling. The reason that shuffling is being excluded, is because it couples data of multiple subjects. If one subject revokes consent, it is difficult to undo the transformation without affecting data corresponding to other subjects.

Table 3-8 shows some sample transformations that will take place in the context of ChildRescue. By definition, reversible transformation leads to pseudonymised data, while irreversible leads to anonymised data. As noted, the module will be configurable by the platform administration regarding which transformations to perform on which fields. Typically, fields that contain personal data, that are however useful for the operation of the system (e.g. e-mail addresses) will be encrypted, with the decryption key stored in the re-identification database. Non-operational personal data will be data masked with the optional possibility to perform encryption upon the masked data. As previously stated, anonymising the data will be performed by erasing the relevant information for the re-identification database, these being the encryption key in case of encryption and the transformation entry from the transformation matrix in the case of data masking.

Table 3-8 Pseudonymisation techniques offered by ChildRescue

Technique	Type of output	Reversible/irreversible	GDPR Compliance
Nulling	Empty	Irreversible	Yes
Substitution	Plain	Reversible	As long as consent is given
Text Masking	Plain	Irreversible	Yes
Image Masking	Picture	Irreversible	Yes
Hashing	Cipher	Irreversible	Yes
Encryption	Cipher	Reversible	As long as consent is given
Variance	Plain	Reversible	Not by its own
Pruning	Empty	Irreversible	Yes

3.2.2 Pseudonyms

Pseudonyms in ChildRescue will be used to provide to the registered users the option to appear with a realistic pseudonym in the platform. Under the present process, a citizen may provide anonymous

²⁰ A one-way hash function, also known as a message digest, fingerprint or compression function, is a mathematical function which takes a variable-length input string and converts it into a fixed-length binary sequence. Furthermore, a one-way hash function is designed in such a way that it is hard to reverse the process, that is, to find a string that hashes to a given value (hence the name one-way.)

information without disclosing any personal information, not even to the call-centre operators. This will continue to apply in ChildRescue by allowing anonymous registration. Pseudonyms will be used in the case where the organisation and the authorities need to be able to discern the identity of the user, but the rest of the community that uses ChildRescue must not. An example of such a case is a citizen who has registered to the platform and wishes to provide evidence for an ongoing case, but feels that the knowledge that she/he provided evidence for, may endanger her/him or the case's subject. In order to obfuscate her/his identity in evidence viewed by the ChildRescue community, the identity of the citizen may be substituted by a pseudonym that is negotiated between her/him and the ChildRescue platform.

The architecture for the pseudonym generation module, as this will be implemented in ChildRescue, is depicted in Figure 3-3. The main components that are used during the process are described below:

- **Pseudonym client:** A s/w component which provides an interface to the end user in order to support the pseudonym generation process.
- **Negotiator:** The negotiator component is used to agree with the user upon a key that will be used for encryption of the communication throughout the pseudonym generation process.
- **Pseudonym Generator:** The pseudonym generator generates the pseudonym under which the user will appear.
- **Pseudo certificate creator:** The pseudo certificate creator generates the certificate which will be used by the user throughout future communications. The pseudo certificate certifies the pseudo-user; that means the user under the pseudonym.
- **Identity Management Proxy:** A component responsible for delegating the traffic data to the logging and auditing module. The component also verifies that the pseudo user corresponds to the actual user for which the pseudo-certificate was issued.

The pseudonym generation process is as follows: A user at first accesses the Pseudonym client. The Pseudonym client is connected to the Negotiator via a Web Anonymisation infrastructure which hides the IP of the user and allows her/him to connect anonymously. The user agrees upon a symmetric key²¹ with the negotiator which is transmitted back to her/him. The user can then use this key to communicate securely with the Pseudonym generation module to create a pseudo certificate. The pseudo certificate data and the corresponding re-identification meta-data are stored in the module's internal storage. The user can then communicate via a 2-way SSL²² Channel with the platform by using the pseudo-certificate. The traffic is proxied by the responsible Identity Management Proxy. The

²¹ Symmetric-key algorithms are algorithms for cryptography that use the same cryptographic keys for both encryption of plaintext and decryption of ciphertext. The keys may be identical or there may be a simple transformation to go between the two keys.

²²Two-way SSL or Mutual authentication refers to two parties authenticating each other through verifying the provided digital certificate so that both parties are assured of the others' identity. In technology terms, it refers to a client (web browser or client application) authenticating themselves to a server (website or server application) and that server also authenticating itself to the client through verifying the public key certificate/digital certificate issued by the trusted Certificate Authorities (CAs)

logging and auditing module logs the traffic for auditing purposes and it further delegates the traffic data to the authorisation module which confirms the validity of the pseudo-certificate and responds back to the user by using the reverse process.

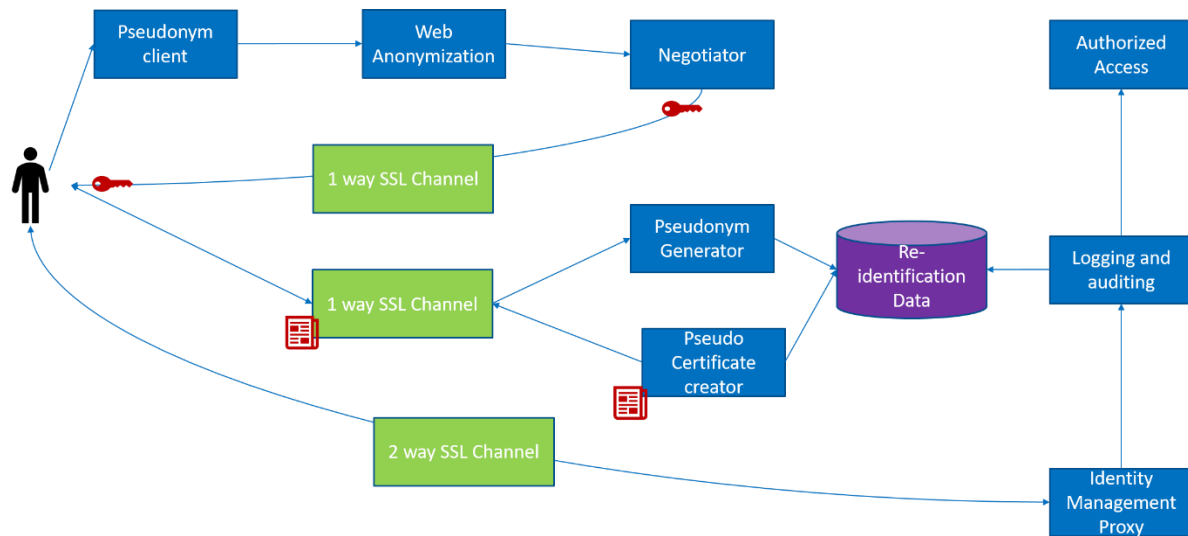


Figure 3-3 Pseudonym generation module architecture

3.2.3 Encryption

Encryption in ChildRescue will be performed both in the level of communication and in the level of storage, where it is needed.

For the purposes of communication, ChildRescue will use a typical Public Key Infrastructure²³ that will ensure that certificates shared between parties are trusted and will apply common cryptographic protocols to ensure that data are transmitted securely. The specifications of how a typical PKI is setup was given in Section 2.3.2.6; ChildRescue will use this setup.

Since most of the traffic will be over the Web, the ChildRescue platform will be deployed under the HTTPS platform and appropriately verified certificates will be distributed to all client applications that need to connect.

Encryption will also be performed for ensuring that proper authorisation and authentication is performed for each user, without the danger of account hijacking. This encryption will take place on top of the usual traffic encryption and will ensure that no information regarding user password credentials is stored unencrypted. Since authentication of a user requires that the hash of the given password matches the hash of the stored password, no decryption is needed. A one-way encryption scheme for user passwords will thus be used.

²³ A Public Key Infrastructure (PKI) is a set of roles, policies, and procedures needed to create, manage, distribute, use, store, and revoke digital certificates and manage public-key encryption. The purpose of a PKI is to facilitate the secure electronic transfer of information for a range of network activities such as e-commerce, internet banking and confidential email.

Encryption will also be used when storing certain data. Data for which anonymisation is required from the start, will be encrypted using a one-way algorithm, namely the Secure Hash Algorithm (SHA). Pseudonymised data on the other hand will be stored by using a public key algorithm, namely the RSA. Since the key pair in the case of RSA is only needed for re-identification, this will be exclusively stored in the ChildRescue Platform, and when data expire it will be deleted. The procedure is depicted in Figure 3-4.

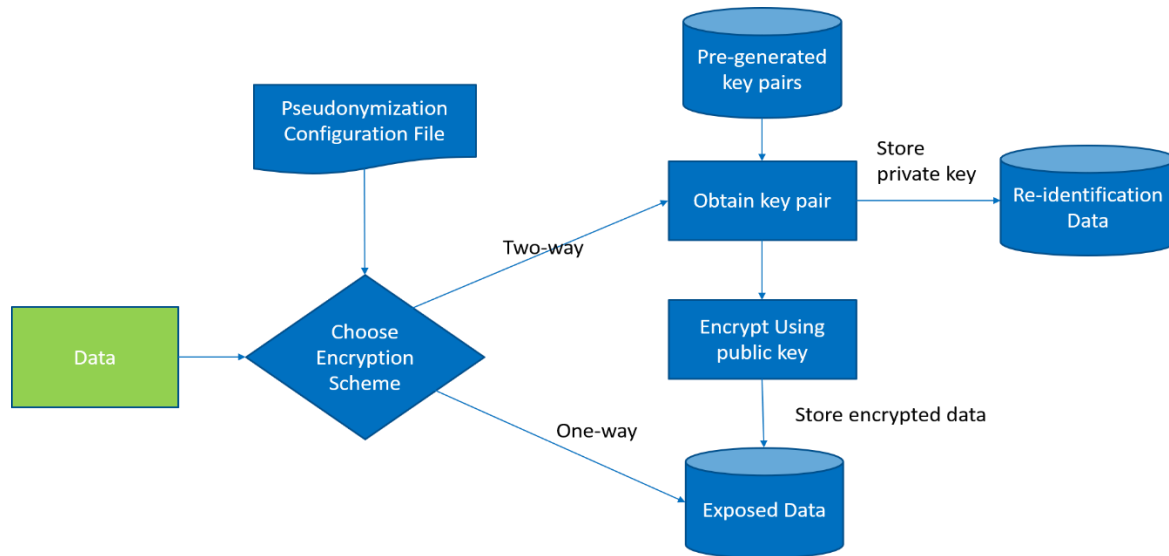


Figure 3-4 Encryption for data fields.

3.2.4 Discussion and Limitations

ChildRescue needs to respect the ethics requirement and existing legislature regarding the privacy of all its data subjects (ranging from children to users) as well as conform to existing procedures that organisations follow. Smile of the Child for example, allows citizens to use the call-centres without them having to provide any personal information. Though not obliged under existing law, this process has been shown to encourage citizens to contact Smile of the Child more often.

Likewise, ChildRescue needs to make certain that no information regarding the data subjects is collected unless that which is absolutely necessary and for which the data subject has consented. Moreover, ChildRescue should continue to facilitate citizen's contributions anonymously. The presented methodologies will be applied to this end.

More specifically, encryption will ensure user's that any personal information they have wittingly disclosed will only be accessible by the organisations and not by any other third party. Pseudonymisation and anonymisation techniques will be used in tandem to make sure that any data that the application needs to hold concerning individual's information will be stored in such a manner so that de-identification will not be able to take place without the relevant re-identification tables; these tables will not be accessible to outside users.

Finally, the ability to use the platform anonymously will be retained by allowing users to register anonymously or to register via a pseudonym.

3.3 Performing Predictive Analytics

ChildRescue is a project with social impact that aims to benefit children and their families, relying on new technologies and modern data processing techniques. Therefore, one of the project's core objective is the efficient, and scientifically sound, analysis of data. Data coming from various sources and usually containing sensitive pieces of information. In this section we are going to investigate what are, or can be, the various types of data sources for ChildRescue, how they can be combined and on what type of tasks the selected algorithms can be applied, so as to produce useful information and help decision-making for all ChildRescue pilot cases through predictive data analytics.

Data analytics in ChildRescue could be viewed as two separate processes with discreet features. The first process involves datasets that are already existing at the beginning of an investigation, and takes place during the PROFILING phase, while it is supported by the ARCHIVING procedures. The second process analyses, in a more dynamic fashion, incoming information from the social sensors (i.e. citizens sending informative messages or pieces of evidence through a mobile app) and can be considered as a continuous effort, involving mostly spatiotemporal data. It lasts for the duration of the ACTION phase, assisting also in the COLLABORATION operations.

In order to conceptualise, develop and adapt the multi-source data processing methodology, the following 6-step approach was adopted that spans the foundations building, the pilot cases analysis and the specifications design axes, as follows:

- Setting up the possible sources and tasks: During this preparatory step, the scope of the data acquisition and processing framework was discussed and agreed.
- Studying the state-of-the art: Following the definition of the overall scope, a state-of-the-art analysis was carried out in order to screen the landscape along 3 dimensions, namely: Human Profiling, Spatiotemporal Data analysis and Social Media Analytics. The research perspective was explored in order to extensively review and classify existing approaches / methodologies and to form a thorough listing and comparative analysis of the most popular algorithms in all the aforementioned high-level areas. Based on the state-of-the-art analysis conducted, useful conclusions on each of these areas were drawn while open issues and gaps were effectively identified.
- Iteratively identifying the data sources available in the pilot cases and discussing with the pilot partners to get a better understanding of the data essence.
- Preparing the data templates in order for the ChildRescue pilot cases to start recording in a consistent way the new data, and deliver historical data and any other operational data in a uniform manner.
- Designing a high-level ChildRescue data model incorporating information from the state-of-the-art analysis and the ChildRescue specific pilot cases and data requirements in order to create a common understanding in the consortium.
- Elaborating on the data analysis along with the data manipulation and processing perspectives in order to bring forward the preliminary ChildRescue considerations, requirements and constraints for the next steps of the project implementation.

Data sources

The data sources available from pilot partners include the archived documents of past cases for missing children, unaccompanied migrant minors' records and tracing requests. Digital files, such as filled-in forms and photos, as well as hand-written documents, comprise the usual set of a case file. Some of these cases contain a link to a social media account.

Data Templates

The purpose of these templates is to present a set of required data fields (data schema) that could be employed by the recommended methods and algorithms in order to produce useful insights and predictions. We considered two types of templates: The Profiling template and the Events template (in order to collect *Profile data* and *Events data*, respectively).

The development of the Profiling template was determined by two main factors: The available formats of data structuring and archiving already utilised by pilot partners, and the results of the research landscape analysis of Task 2.1 and Task 2.2 of WP2. In other words, we compiled an initial list of 96 attributes with the data fields the pilots already use and the fields we additionally want.

The next step was to ask the pilot partners to look into their archives and specify which of these data fields always contain some form of information for all cases (mandatory), in some cases (optional) or in none (unused). Finally, we resulted in a more flexible list, limiting the field number to *40 attributes*, according to the replies of the partners and the expected contribution of each feature to the data analytics. The outcome of this process is presented in Table 3-9.

Table 3-9 List of data fields included in the Profiling Template

Name of field	Type of reply	Importance	Category	Possible Values/List (examples)
Case ID	Alphanumeric value	Mandatory	Case data	e.g. IA23891023D
Area/Location child was last seen	Text value or coordinates	Mandatory	Case data	e.g. "Around Limassol castle, Limassol"
Date and time the child was found	Date time	Mandatory	Case data	e.g. 15/08/2017 11:30
Date and time of disappearance	Date time	Mandatory	Case data	e.g. 16/07/2017 23:00
Conditions of disappearance	<i>Select 1 from list</i>	High	Case data	On way to school; Returning home from activity; Was out with friends; visiting a friend; on trip to another city;
Possible reasons of disappearance	<i>Select multiple from list</i>	High	Case data	argument with family; victim of school bullying; argument with boy/girlfriend; visit friend in another hosting facility; in search for relatives; recently moved to new hosting facility;
State of child when found	<i>Select multiple from list</i>	High	Case data	abused; normal; shocked; dead; wounded; etc.
Currying mobile phone	Yes / No	High	Case data	Yes / No
Currying money or credit card	Yes/No	High	Case data	Yes / No
Has area knowledge	Yes / No	High	Case data	Yes / No
Rescue teams utilised	Yes / No	Low	Case data	Yes / No
Volunteers utilised	Yes / No	Low	Case data	Yes / No
Transit country/-ies	Separate	Low	Case data	e.g. Turkey; Greece; Albania

reached or intended to be reached	values using semicolon [;]			
Date of arrival at hosting facility	Date time	Low	Case data	e.g. 16/07/2017 23:00
Type of disappearance (Category)	<i>Select 1 from list</i>	High	Case data	runaway; parental abduction; criminal abduction; lost/injured/otherwise missing; missing U/A minor; tracing request; unclear;
Multiple-times case	Number	High	Case data	e.g. 2
Family members	Number	Medium	Case data	e.g. 5
Probable destinations (location/city/country)	Separate values using semicolon [;]	Medium	Case data	friend's house in Brussels; Music event in Apollon Limassol stadium;
Clothing with scent (for dogs)	Yes / No	Medium	Case data	Yes / No
Home / Facility Address (area only)	Text value or coordinates	High	Demographics	e.g. Ritsona Refugee Camp, Vathy, Euboea, Greece
Education level & current participation in educative activities	<i>Select 1 from list</i>	Low	Demographics	first grade; second grade; third grade; none; unknown;
Languages spoken (number of)	Number	Low	Demographics	e.g. 1
Nationality	<i>Select 1 from list</i>	Low	Demographics	e.g. Syrian
Place of birth / Country of origin	<i>Select 1 from list</i>	Low	Demographics	e.g. Syria
Age (or Birthday)	Number (Date)	High	Demographics	e.g. 15 (or 12/9/2003)
Gender	<i>Select 1 from list</i>	High	Demographics	Male/Female
Health issues (current)	<i>Select multiple from list</i>	High	Medical Profile	allergies; substance abuse; diabetes; asthma; heart disease; pregnant; etc.
Medical treatment required?	Yes/No	High	Medical Profile	Yes / No
Social media accounts	Platforms /apps separated with semicolon	High	Personality/ Social Profile	e.g. Facebook; Snapchat; Instagram
Protection concerns/Vulnerabilities	<i>Select multiple from list</i>	High	Personality/ Social Profile	child headed household; disabled; medical case; street child; girl mother; living with vulnerable person; abuse situation; trafficking/exploitation risk; early marriage; other;
Specific personality characteristics/psychological disorders	<i>Select multiple from list</i>	High	Personality/ Social Profile	antisocial, suicidal, autistic, depressive, schizophrenic, other mental or emotional disorders
Family situation	<i>Select 1 from list</i>	High	Personality/ Social Profile	(living with both biological parents, living with single parent (divorced, separated, widowed, other), living under relative's care, living in foster care, living in hosting

				facility/other institution, separated child, unaccompanied child)
Parents' (Tracing enquirer) profile evaluation	<i>Select 1 from list</i>	High	Personality/Social Profile	Excellent; Good; Sufficient; Not good; Really Bad;
School grades, Absences	<i>Select 1 from list</i>	Low	Personality/Social Profile	A' student with zero absences; A' student with normal absences; A' student with excess num of absences; B' student with zero absences; B' student with normal absences; B' student with excess num of absences; C' student with zero absences; C' student with normal absences; C' student with excess num of absences;
Interests/Hobbies	<i>Select multiple from list</i>	Medium	Personality/Social Profile	music; football; basketball; dancing; painting; singing; etc.
Relationship status	<i>Select 1 from list</i>	Medium	Personality/Social Profile	single; in a relationship; married; its complicated; other;
Religion	<i>Select 1 from list</i>	Medium	Personality/Social Profile	Orthodox Church; Catholic Church; Protestantism; Other Christian; Islam (Sunni); Islam (Shiite); Other Islam; Hinduism; Buddhism; Nonreligious; Other;
Weight	Number	Low	Physical data	e.g. 42
Height	Number	Low	Physical data	e.g. 140
Distinguishing features	<i>Select multiple from list</i>	Medium	Physical data	tattoos; scars; skin marks; missing teeth; front teeth gap; body piercing; other;

As one can observe, the fields required are divided by type, importance and general category. The type of reply is related to the technical implementation of the respective field and the expected way the data will be acquired by the ChildRescue platform. Where a *List* is denoted, its contents will be populated by the aggregated values existing in the compiled past cases, along with the suggestions of experts. The importance is an arbitrary estimated value for the respective data field's contribution to data analytics. The categorisation column groups the fields of similar characteristics into 5 classes.

For the Events Template, things were more straightforward. We created a template that records the sequence of events (spatiotemporal input) and incoming proofs of evidence, starting from the event of the disappearance (place and time the child was last seen) until his or her finding.

Table 3-10 List of data fields for the Events Template

Name of Field	Example value
Step ID	E.g. 1,2,3...
Events in chronological order / Tracing steps (Date time)	E.g. 12/07/16 18:30
Location child was seen	E.g. Panepistimio metro station, Athens, Greece

Transportation means the child used to reach location	E.g. On Foot, Subway, Bus, unknown, etc.
Evidence True?	E.g. TRUE/FALSE
Reasons for location selection	E.g. Relative's home, there is a subway station, etc.
Child Physical status	E.g. Clothing have changed, Had a hair-cut, etc.
Extra info	E.g. Child was accompanied by middle-aged woman

Data Model

The data coming from multiple sources need to end up into a form of data model that completely describes what ChildRescue is about. From the development and compilation of the data templates, it soon became evident that the core elements of a high-level model should be three: The Case Profile, the Child Profile, and the Events Log (Figure 3-5).

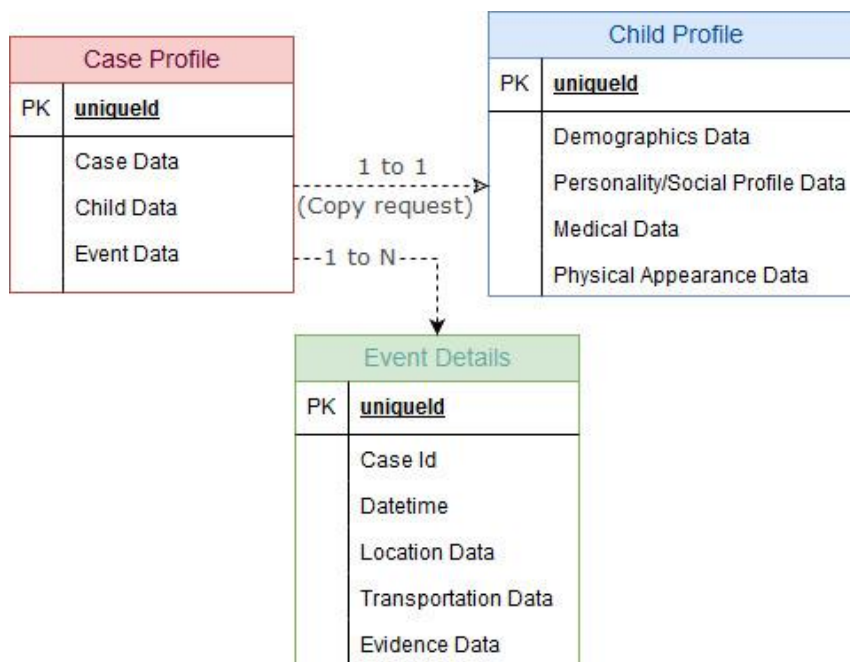


Figure 3-5 The proposed preliminary Data Model for ChildRescue

The main entity of our model is the Case Profile. The Case Profile includes a) case data, related to the current case, b) child data, that are transferred to the case from the current (i.e. updated) profile of the child, and c) a list of Events. Each Event includes, at least, a date and a time of the event, location data, transportation data and details on the related piece of evidence.

The Child Profile includes demographics, personality traits, medical and physical data, and maintains the up-to-date version of the child (where, in comparison, the case data keeps the child version of the time the case took place). The Child Profile transfers a copy of itself if and when requested by the Case Profile (i.e. when a new case is opened for an existing, in the database, child. Otherwise a new

child profile should be created in advance). In addition, when a case involves more than one child, say x children, then x Case Profiles should be created so that each child is tracked separately.

Following this practice may lead to the same information being kept multiple times, but this is a small sacrifice to pay, if we aim to cover each case thoroughly and correctly. After all, due to the very nature of the project, it is the first priority of ChildRescue to propose and follow a rigorous and sound methodology of data analysis which produces robust, accurate and secure results. Other factors, such as the computational speed or cost of resources, although important, come at a second place.

Nevertheless, when cases are related, either when there is a simultaneous disappearance of more than one child and they are somehow related (e.g. two runaways that are siblings or close friends), or when there is a natural disaster involving a number of missing children, or when it is a repeated case by the same individual, then these cases should be marked as connected, so that an easier recovery or update of associated information can be accomplished.

Data Analysis

The last step of the proposed methodology is concerned with the data manipulation and analysis, in order to bring forward the preliminary ChildRescue considerations, requirements and constraints for the next steps of the project implementation. This procedure involves several stages, from data ingestion and transformation to the actual data analysis and the visualisation of the results.

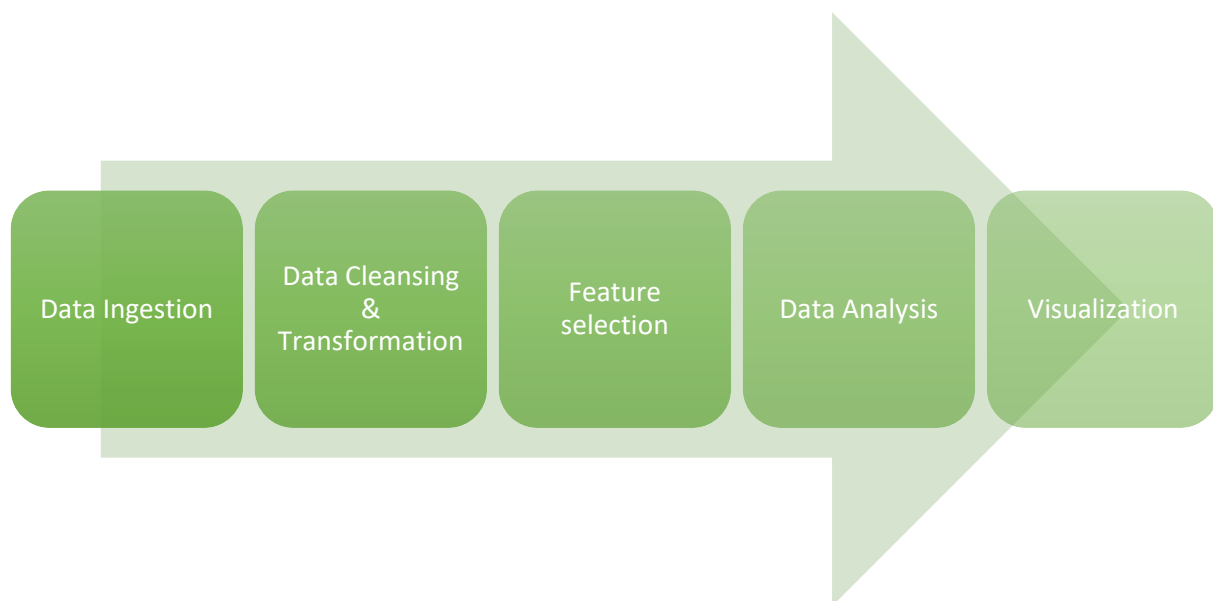


Figure 3-6 A simple data analytics flow

Judging from the data available and the knowledge we can infer from them, we conclude that there can be three types of data analytics that can be performed within the context of ChildRescue. These are depicted in the following table, with each one being further detailed in the lines to follow.

Table 3-11 Relation of ChildRescue investigation phases and Analytics types

Analytics type	Investigation Phase	Data sources
Predictions based on	PROFILING,	Past cases profile data, Current child

Behavioural and Activity Profile Data	ARCHIVING	profile, Social media
Evidence Analysis and Evaluation	ACTION, COLLABORATION	Evidence data, User profile data, Past cases events data
Real time Route/Destination Estimation	ACTION, COLLABORATION, PROFILING	Past cases data, Evidence data, Open data (transportation, events, etc.), Linked data

3.3.1 Predictions based on Behavioural and Activity Profile Data

As described in section 2.2.1 of the present document, profiling is all about predicting future behaviours of an unobserved individual based on aggregated knowledge from a large group of observed individuals. According to ChildRescue workflow (see deliverable D1.3), during the PROFILING phase, information coming from multiple sources is collected and refined so as to create a more complete and informative profile. This can be applied on both the missing children and the unaccompanied migrant minors' cases.

Using the concept of our proposed data model, our definition of profile requires all the elements that comprise the specific case, as well as the child's details. Once we have the individual's profile data, we can then compare it with the aggregated set of past cases and discover hidden patterns and correlations.

For any personality attributes that are missing or not yet acquired, we can take advantage of social media analytics and retrieve them, in the same fashion as described in [170].

More specifically, the objectives of this procedure should be:

- 1) To assess the profile model of the case-child (descriptive analysis),
- 2) To extract missing information using social media analytics,
- 3) To employ behavioural predictive analytics based on past cases.

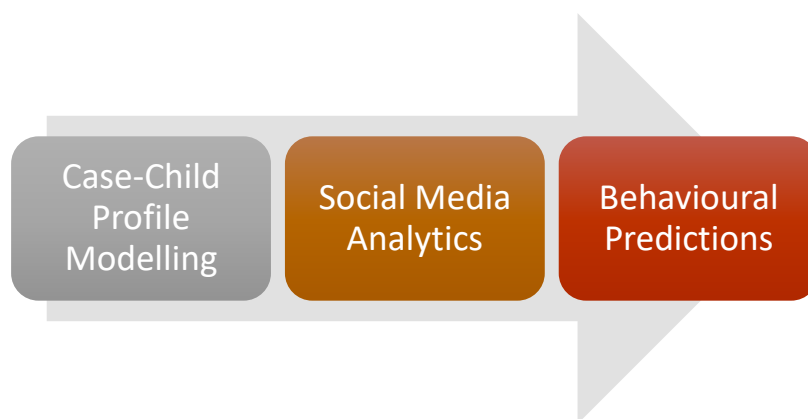


Figure 3-7 Behavioural Prediction process

The behavioural prediction concept could be further broken down into more specific tasks, examples of which can be:

- 1) The classification of a new profile to a Case category,
- 2) The clustering of behavioural characteristics, either on child or case level, or
- 3) Personality classification based on the big-five model.

If the social media account of the missing child is not confidential (for anonymity reasons), we can also enrich the profile with his or her social account preferences, activities and networking circle. However, an obstacle to this procedure would be the fact that we would need to acquire the same pieces of information for the archived cases, as well.

3.3.1.1 Possible sources

The data required for the profile assessment and pattern extraction can be derived from the previous investigation cases and, optionally, from social media provided data. For this purpose, each ChildRescue pilot partner was asked to compile a number of past cases from their archives, in order to modify their data according to the Data Template presented in the beginning of this section.

In "Annex IV: Past cases list of reference", a sample list of 122 past cases, compiled by all pilot partners, is presented, along with 6 complete examples.

The requirements for every case of this sample were:

- to hold as much information as possible,
- to be relatively recent, and
- for the sample itself, to be statistically representative of the Case categories.

The list contains 122 past cases, out of which 8 are still open. All cases refer to actual events that took place in the last few years (since 2009 and onwards) with the majority being in the last 4 years (since early 2015). This dataset will be employed as input to train the appropriate algorithms in order to create a profiling model. The resulted model will then be used to assess any new profile. The optional use of social networks data, as already mentioned, has to do mostly with the prediction of the child's missing attributes, like religion, relationship status, etc.

The individual's social data and networking connections would be also of great value to the profiling assessment, in a sense that many hidden traits and habits could be located and utilised, that are otherwise very difficult to obtain (see for example [122] and [123]).

The table below presents the data sources required by this part of the methodology.

Table 3-12 Data sources to be used for profiling

Type	Data sources
Mandatory	Past cases data, current Child Profile
Optional	Social media data (aggregated)
Optional/Privacy issues	Social media data (individual profile)

3.3.1.2 Methods and algorithms

In this section we present the methods related to profiling predictive analytics. The three objectives mentioned, namely profile modelling, social media analytics, and predictive analytics, usually utilise data mining techniques and computational learning algorithms.

The following table includes all algorithms and methods encountered in literature analysis, that in one way or the other, can effectively deal with the ChildRescue profile data and play an integral part in predictive analytics. The specific tasks are yet to be determined since the actual data availability, structure, and quality are not finalised at this preliminary stage.

Table 3-13 Summary of algorithms for profile modelling and predictions

Objective	Related Family of algorithms	Popular algorithms
Profile modelling	Clustering, Classification, Correlation analysis	K-means SVM, k-NN, decision tree, Random Forest Ordinary Least Squares regression
Social media personality prediction	Classification/Regression, Linguistic Analysis	SVM, k-NN, Random Forest NLP techniques
Behavioural Analysis	Classification/Regression	SVM, k-NN, Decision tree, Random Forest, Naïve Bayes
Sensitivity Analysis	Linguistic Analysis, Classification	NLP techniques Decision trees (C4.5), Naïve Bayes, SVM

3.3.1.3 Profiling Algorithms for ChildRescue

A preliminary approach on the ChildRescue profiling process is based on the five (5) missing child profiles described in Section 3.1.1, where is clearly indicated, and supported by relevant theories, that each of these profiles has its own distinct characteristics. Therefore, an investigation can be affected and directed according to the profile at hand. The task is to be able to classify a case before knowing the real outcome of the investigation, so that the organisations and authorities involved in the case can act in the most appropriate fashion.

Having this in mind, five algorithms are proposed, one for each distinct profile type, as depicted in Table 3-14 that involve a correlation analysis step as well as a classification process.

Table 3-14 Profiling approach

Algorithm	Inputs	Tasks	Output
Profiling #1	Past cases details	Correlation analysis, Classification	Runaway
Profiling #2	Past cases details	Correlation analysis, Classification	Third-person abductions
Profiling #3	Past cases details	Correlation analysis, Classification	Parental abductions
Profiling #4	Past cases details	Correlation analysis, Classification	Missing unaccompanied migrant minors

Profiling #5	Past cases details	Correlation analysis, Classification	Lost, injured or otherwise missing child
--------------	--------------------	--------------------------------------	--

Each of these algorithms will follow a similar approach, which includes a feature correlation analysis first, in order to select the appropriate input details that are more significant for the given profile. This is an important step from a psychological point of view as well, because it can be used to explain the motives and behavioural patterns behind each profile.

On the classification part, since the total number of the available past cases are already labelled with one of the missing child categories, the supervised learning approach is the most appropriate and efficient method to be utilised. In such a case, SVM, decision trees and random forests seem to be the best candidate methods to apply.

In the months to follow, extensive experiments should be performed, so that a well-trained model is produced for each profile. The comparative results and the produced models are to be presented in the next iteration of this deliverable in D2.5 [M24].

3.3.2 Evidence Analysis and Evaluation

In a broad scope, evidence can be defined as anything presented to support an assertion. In ChildRescue, pieces of information are accumulated and assembled to support a missing child's investigation process during the ACTION phase (see deliverable D1.3). In other words, it is the evidence collection and analysis that construct, piece by piece, the tracing path towards the missing child.

The incoming pieces of evidence may include errors, or can be completely false, either accidentally or on purpose. One piece may hold truth in respect to location, but have the time wrong, while another can be 100% correct, but it is sent by an anonymous, and thus, less trustworthy, user.

It is, therefore, a crucial task for ChildRescue to be able to rapidly and reliably analyse the data coming from exterior sources (i.e. citizens, also known as "social sensors" in this project) in an intelligent and efficient manner.

The main goal of the task is to assess the quality and reliability of the incoming information, so that the extracted data offer valuable assistance in the investigation for a missing child and do not mislead or otherwise, hinder the whole process.

More specifically, the objectives of the evidence analysis should be:

- 1) To assess the quality of evidence
- 2) To assess the credibility of the sender
- 3) To weigh the value and reliability of one particular piece of evidence against another or the credibility of a sender against another (contradicting information)
- 4) To validate a piece of evidence through some form of collective intelligence (i.e. Crowdsourcing)

In this framework, ChildRescue should adopt a clear methodology of evaluating various forms and types of evidence by analysing not only the content of a message but also the source. In other words, in ChildRescue evidence evaluation implies the analysis of both the provider and the information provided. In our opinion, this evaluation process should involve three steps:

- 1) User evaluation through his/her historical behaviour
- 2) Content evaluation, through cross-checking with similar (in time and space) information
- 3) Evidence validation, through Crowdsourcing, by sending relevant info to the appropriate and eligible receivers (usually in order to verify or dismiss it)

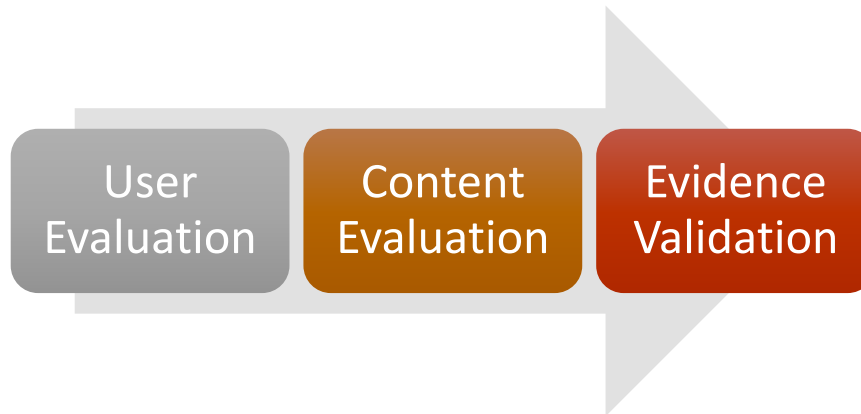


Figure 3-8 Evidence Evaluation process

In the following subsections, we are going to describe thoroughly these steps, by specifying the sources and the means that can achieve as much accurate evaluation of evidence as possible. The theories of Crowdsourcing and collective intelligence will also be presented as they will play a key part in the validation process.

3.3.2.1 Possible sources

In the case of evidence, ChildRescue primary contributors will be the active “Social Sensors”, i.e. all the users, registered or not, who have downloaded the ChildRescue mobile application and wish to contribute to the investigation process by sending proof of evidence regarding the case.

The incoming information can be of several types and formats. Text messages, images, videos, voice recordings and geolocation data are the most common types supported by modern applications. In ChildRescue we will consider all of these types as possible evidence, albeit not all of them can be used by algorithms.

The data that will be used as input for the evaluation process will contain, apart from the submitted evidence, its accompanying metadata (timestamp, device details, geolocation, etc.). The results of the predictive algorithms (i.e. the calculated POIs) and current case information, along with the user’s profile history, will be combined with the evidence in order to provide a reliable and accurate evaluation.

Table 3-15 Data sources to be used for evidence evaluation

Type	Data sources
Mandatory	Text message, user geolocation, evidence timestamp and location, user profile history
Optional	Image, Video, voice recordings, User profile details, User device details, Past cases

	related info
--	--------------

3.3.2.2 *Methods and algorithms*

In this section we present the methodological approach towards evidence analysis. The three steps recognised previously, are further analysed and matched to known algorithmic tasks. The algorithmic families along with widely used algorithms that will be considered for the implementation of the evidence evaluation are presented in Table 3-16.

User evaluation through past behaviour

The expected user evidence will be received mainly through the ChildRescue mobile app. The app users are graded in different groups with different access to case data, different permissions and functionalities. All of the above-mentioned users will be able to submit evidence. It is obvious however, that this evidence could (and should) not be equally ranked, as data coming from a trustworthy, authenticated source should generally be prioritised.

User history is another important factor when considering user credibility, as a user who has previously provided valid leads is considered more trustworthy. For this evaluation to happen, it is necessary for the system to be able to assess the provided evidence after closure of the case (with human expert intervention). Apart from user history, the profiles are characterised from other attributes also, such as demographics, place of residence, etc.

When it comes to new users without a history of actions it becomes more complex. In order for their evidence to be evaluated, we could utilise clustering algorithms, which will enable us to “predict” the user evaluation factor before she/he has even provided evidence. The clustering could be performed on user profiles and the “prediction” could be based on the previous performance of similar users.

However, because demographics do not usually suffice for extracting safe results, we should also consider applying a fixed cold-start credibility value for every newcomer. After having enough history data, we then can proceed with the ranking of this user. An issue for further consideration on this approach, is the starting point: should we consider new users as having the highest credibility and maintain or lower it according to their provided evidence, or should we set the default credibility for a newcomer to zero and let him/her build piece by piece the required trust with the system? In our opinion, the European culture strongly favours the former approach and this is what we recommend: every user should be by default trustworthy.

Content evaluation, through cross-checking with similar (in time and space) information

Content evaluation consists of finding contradictory elements inside the uploaded evidence, or in relation to other case information, which has already been evaluated and is considered valid. These elements could concern both time and space and indicate a possibly incorrect and misleading piece of evidence. For example: a user states that she/he witnessed the child half an hour ago and pinpoints the location of last seen. However, the geolocation of the user, which is also included in the submitted information, shows that she/he is in another city, and her/his statement of witnessing could not be physically feasible. In cases such as the above, the submitted evidence will not be filtered or excluded. They will however receive lower ranking and be appropriately tagged.

Contradicting evidence could be perceived as data differentiating from the expected results (outliers). An outlier is a data pattern not conforming to the expected behaviour and anomaly detection is the locating of such spikes in a data set [171]. Thus, an analogy could be drawn between contradicting evidence and spatiotemporal anomaly detection, a rather well-studied field in the literature. In our case, the expected data values could be calculated by factors such as the location the missing person was last seen, the expected location, etc. Nevertheless, such algorithmic approach requires a sufficient amount of data to be able to discern anomalies. For instance, if there are only two patterns contradicting each other, then each one sees the other as outlier and are both right (or wrong).

Evidence validation, through Crowdsourcing, by sending relevant info to the appropriate receivers (usually in order to verify or dismiss it)

Following the positive aspects of the collective intelligence, which will be briefly presented below, a Crowdsourcing process will take place in order to validate the evidence at hand. The power of the crowd will be exploited in two ways. On the one side, the active participation of the ChildRescue users will be needed. The examined evidence (part or the whole of it) will be disseminated for verification to the legitimate receivers. The extent of dissemination, location and content will be defined accordingly. It could vary from a simple alert to all mobile app users nearby the evidence location, for activation of the local community, to the uncensored evidence being sent to authorised volunteers and S&R members, who will consecutively go to the indicated place for self-inspection and verification or dismissal of the information.

The second step of the validation could be an automated process and thus, active validation from the users would not be required. Indirect confirmation of the evidence through a clustering algorithm can take place in such case. Text analysis and location proximity among submitted evidence could be the major similarity factors used for the clustering. If a considerable quantity of similar evidence is accumulated, then it should acquire a higher validity factor.

Table 3-16 Algorithms related to Evaluation Steps

Objective	Related Family of algorithms	Popular algorithms
User evaluation (profiling)	Clustering, Classification	K-means SVM, k-NN, Random Forest
Anomaly detection	Regression/Classification, Stochastic analysis, Linguistic Analysis	SVM, k-NN, Random Forest Markov models NLP techniques
Crowd sourced Evidence validation	Clustering of incidents	K-means, Non-negative Matrix Factorisation

Collective Intelligence

Collective intelligence is the accumulated and shared intelligence resulting from the collaboration of a loosely organised group towards a purpose. An expression of collective intelligence is through Crowdsourcing platforms, like Wikipedia or iStockphoto²⁴, where a call for individuals' contribution (regardless of their expertise) leads to a remarkable result. According to Surowiecki (2004), a successful solution often emerges from a large basis of people [172]. It is also stated that the aggregated ideas contributed by a large group can exceed in intelligence those of the smartest individuals. Except for content uploading, another example of a popular collective intelligence harnessing method is by explicitly asking the users to rate or vote for an item [173].

In [174], which also addresses the problem of information correctness assessment in social-sensing applications, we also read about the collective response to a social purpose, varying from the reporting of potholes to the participation in rescue missions. The collective behaviour response to disasters has also been identified and studied previously [175]. Considering the instantaneity of information diffusion through the internet and mobile phones, combined with the active response of citizens to numerous emergencies or disasters, as for example in the Virginia Tech Shootings [176], we can see the emerging potential for ChildRescue.

Crowdsourcing in ChildRescue will not be limited to just asking for and gathering potential evidence from users. It will also be utilised in the evidence evaluation process, were the collective intelligence will be extracted as was described previously in the methodology (both directly and indirectly).

3.3.3 Real time Route/Destination Estimation

The task of route and destination estimation was encountered in literature by the name "trajectory prediction" and "POI recommendation", respectively. Actually, this task is the most complex of the three tasks assigned to ChildRescue analytics and the reason is threefold: Firstly, we have to deal with multiple heterogeneous sources of data; secondly, these data are both spatial and temporal in nature; and thirdly, ChildRescue does not rely on using mobile phone or GPS log data since a child that has gone missing usually does not carry a cell phone, and even then, the call data or GPS logs are not accessible.

In particular, the objectives of this complicated process should be:

- 1) To use profiling information and extract an initial set of related POIs for the child.
- 2) To predict next POIs or locations the child will visit based on dynamically assessed information from verified evidence and transportation data.
- 3) To simulate and predict possible movement trajectories (routes) of the child.

Because of the apparent low probability to locate a moving individual within a city's limits or within a country using the aforementioned data sets, the actual purpose of this task is not so much to locate the child, but to estimate a centre (POI) and a radius of an action circle where the available Social Sensors, volunteers and/or rescue teams will receive appropriate notifications to act.

²⁴ <https://www.istockphoto.com/>

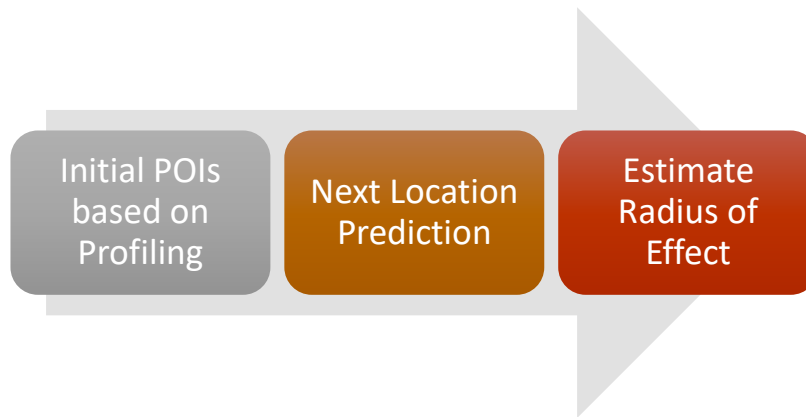


Figure 3-9 Route estimation process

On the bright side, machine learning techniques, as well as recommendation systems, possess enough power to tackle these issues successfully and produce useful indicators on the child’s general whereabouts, as long as the data deriving from these multiple data sources is of sufficient quality. In other words, this is the task where the data sources come to play the most significant role.

3.3.3.1 Possible sources

The available data sources can be divided into two groups: static ones, like the past cases with their respective spatiotemporal events, open data, or social network preferences, and dynamic ones, which mostly have to do with the incoming geolocated evidence in real time. The former is analysed during the PROFILING phase with the sole purpose to create a number of initial predictions for POIs or possible destinations. During the ACTION phase, the spatiotemporal information collected, can be utilised to recommend next possible locations and form route trajectories based on stochastic methods.

The data sources that will be used as input for the recommended methods are summarised below.

Table 3-17 Data sources that will be used as input for the recommended methods

Type	Data sources
Mandatory	Past cases events data, place the child was last seen, evidence suggested locations
Optional	Open data (Weather, social events, transportation routes & infrastructure), social media personal preferences, social media status posts

The events recorded in a previous case, as already explained, should follow the past cases Event Template, which includes, among other information, the date and time of the event, the location, and any used transport. Apparently, for the training of any algorithm we consider only the verified (to be true) events.

3.3.3.2 Methods and algorithms

The objectives of this task can be viewed as common undertakings in the research community that deals with spatiotemporal data. POI recommendations and trajectories predictions have been a subject of study for numerous publications. However, in the ChildRescue case, the surrounding conditions are not that favourable.

Recommending a point of interest or predicting a route, is quite challenging when the primary dataset you base your analysis on, does not possess this kind of information. In that case, the influence of external data sources becomes crucial and making the most of state-of-the-art methods becomes a necessity.

Under these circumstances, the proposed algorithms attempt to cover all possible objectives. The location recommendations, either initial or later, should employ various techniques and select the most efficient. Even ensembles can be of use. For the trajectory prediction, the utilisation of probabilistic models, combined with clustering methods, seems to be the only route.

Ultimately, depending on the resulted routes or POIs, machine learning algorithms should estimate the centre and radius of influence that will infuse geofencing techniques with input attributes.

The proposed sets of algorithms, divided by objective, are summarised in the next table.

Table 3-18 Methods and algorithms for route/destination estimation

Objective	Related Family of algorithms	Popular algorithms
Initial and next POI recommendations	Recommenders, User mobility modelling, Clustering, Classification, Deep Learning	Collaborative Filtering Markov models and decision process Non-negative Matric factorisation Decision tree, Random Forest, ANNs, Naive Bayes, SVM, Fuzzy expert systems RNNs (e.g. LSTM)
Route (trajectory) prediction	Probabilistic modelling, Clustering	Markov models Density-based clustering algorithms (e.g. OPTICS, DBSCAN), Voronoi partitions
Radius estimation	Regression	Decision tree, Random Forest, ANNs, SVM

Geofencing

The widespread use of smartphones equipped with several modern sensors allows for the detection of the user's current physical activity or the user's presence in a designated area. The latter is often referred to as Geofencing [177]. Geofencing technology is a location-based service, mostly encountered in mobile phone applications, that allows the sending of notifications to users who enter or exit a specified geographical area. As such, this service has become very popular among today's mobile marketing strategies and smart urban environments [178].

In order to specify a geographical area for this mobile sensing service, one has to designate a centre and a radius. The resulted circle functions, more or less, like a fence, triggering an event when a mobile user (with the appropriate application) enters or exits the area. This is the reason it is required by our methodology to be able to infer the circle's attributes.

Geofencing technology, or similar techniques, will be adopted by the ChildRescue platform in order to send location-based notifications to registered users, citizens or volunteers. Since it is a well-founded technology, more details on its technical aspects and how this is going to be implemented in the framework of the communication functions of ChildRescue will be examined in the related WP3 Tasks.

3.3.3.3 Prediction Algorithms for ChildRescue

In the context of ChildRescue, there are two cases where predictive analytics can be applied, both involving mostly runaway children which is the most common category of missing children. The first case has to do with the prediction of initial locations a child has decided to visit based on his or her personality, social status and habits. The second case is related to the route a child might follow based on dynamic spatiotemporal information that the system collects during the investigation.

For the prediction of initial points of interest, an analysis of past cases of children with similar characteristics will take place. Social media activity, if available, could also add more information about the child's preferences and places he or she likes to visit according to his or her behavioural profile. During the analysis, the most significant features that are related to location selection need to be found, and then, a classification algorithm or a recommendation system can be employed to deliver a type of PoI (e.g. parent or friend's house, social event, etc.) most suitable for the given personality.

For the route estimation, it is expected that spatiotemporal information is going to be scarce, since mass gps and mobile phone data will not be available. The primary source of information will be the log of events occurring during the investigation and ChildRescue's ACTION phase. Usually, this log includes a small number of events (1 to 10), which, in addition, cannot be verified as true or false. However, ChildRescue's intention is not to replace the police force. ChildRescue's main purpose is the dissemination of information (i.e. alerts) in one or more specific geographical areas. Therefore, the predictive abilities of this second case focus on estimating dynamically the appropriate centre and radius wherein the message needs to be broadcasted at the given time. From a computational learning point of view, this is a regression task and the most popular algorithms that will be tested are: linear regression, decision trees, random forest and neural networks.

Algorithm	Inputs	Tasks	Output
Predictive #1	Past cases details, Social media activity	Correlation analysis, Classification, Recommenders	Possible initial PoI types
Predictive #2	Past cases events, usage of transportation infrastructure from open data	Regression	Radius value (e.g. from point last seen)

In the following months, extensive experiments will be performed, using all available information from past cases and relevant datasets. The comparative results and the produced models are to be presented in the next iteration of this deliverable in D2.5 [M24].

3.3.4 Discussion and Limitations

In this section, we proposed a multi-source analytics methodology based on the knowledge extracted from the state-of-the-art review, properly adapted to facilitate the special data properties and characteristics of ChildRescue. A step-by-step procedure was established first, to explicitly describe all the required actions that can lead to a successful data analytics strategy. The data analytics are, then, broken down into three separate approaches, each one dealing with multiple sources of data.

All three of them derive from the examined literature domains, but with focus on predictive analytics based on human behavioural profiling, evidence evaluation and validation, and the geospatial data analysis required to estimate points of interest and possible routes. These three fields of analytics compose the core of the mechanism that we call ChildRescue multi-source data analytics. A common data source in all three methods is the past cases archive of the pilot partners, and this is why a specific data template design was clearly defined from an early stage, so that we can collect these datasets in a structured and uniform fashion.

All the methods and algorithms presented in sections 3.3.1, 3.3.2 and 3.3.3 will be tested and comparatively assessed on the basis of the data to be provided by the pilot partners. The most suitable ones will serve as the algorithmic basis in the relevant components of the ChildRescue platform.

The major challenges expected to emerge during the ChildRescue project, in respect to data collection and analysis, can be summarised as follows:

- Heterogeneity of data sources available at the different pilot partners (data types, naming conventions, different languages, etc.)
- Currently, most information regarding the psychological profiles is maintained in large text files. In some cases, the full content is not even digitised.
- Past cases may have missing details in several aspects of the data profile.
- The spatiotemporal data analysis, as estimated given the availability of data, will be hardly able to predict real trajectories and routes efficiently.
- Data analysis privacy issues, since ChildRescue is expected to handle personal, and sometimes, sensitive, data.
- Crucial GDPR issues with social media. At the time of this writing, the EU regulation concerning data privacy has been applied to all data owners/providers in a global scale. The most popular social media platforms (e.g. Facebook) have decided to be very strict about enforcing the privacy regulations and access to their data is now very limited.

For each of these limitations, a contingency plan was considered. The table below presents the limitations, characterised by their respective probability of occurrence, along with the recommended contingency plan.

Table 3-19 Summary of data collection and analysis limitations

Challenge / Limit	Probability	Contingency Plan
Heterogeneity of data sources available at the different pilot-partners legacy systems (data types, naming conventions, language, etc.)	High	Generic data model representation of the different data sources;
Most historical information is unstructured in text format (e.g. word documents)	High	Manual transformations; descriptive analytics.
Complex spatiotemporal data processes that may lead to low performance trajectory prediction	High	Focus the analysis on estimating approximate values for centre and radius of a circle, instead of estimating the exact route followed by a missing child.
Privacy issues	Medium	Use of anonymisation techniques on data. Use of algorithms and methods able to cope with anonymised data.
Cases with missing information	Medium	Social media analysis approach to retrieve missing values based on profiling; Methods for handling missing values.
Limited social media data	Low	Examine many different social platforms and combine data.

The ChildRescue architectural design and implementation should reflect on these limitations, as well as the suggested workarounds, and ensure that the respective components can successfully cope with any, data-related, challenge.

4 Conclusions & Next Steps

The focus of this deliverable has been to lay the background for the methodological aspects of profiling, analytics and privacy that will be applied in ChildRescue. The most prominent theories related to activity and behaviour profiling were presented from the disciplines of social and computer science, followed by a thorough state-of-the-art research study on the multi-source analysis of human behavioural patterns, social media and spatiotemporal data. These three data analytics domains were regarded as the ones which cover most aspects of a missing child investigation. The last part of the research analysis was devoted to privacy and data protection, which define the most significant challenges in terms of data collection and processing for ChildRescue.

The research analysis on the activity and behaviour profiling led to the identification of five distinct profiles that are related to the five categories of missing children cases. These are:

- Runaways
- Third-person abductions
- Parental Abductions
- Missing unaccompanied migrant minors
- Lost, injured or otherwise missing

Each such profile derives from the corresponding category and describes not only the case, but attempts also to explain the child's psychological and social motivations using different theories. Consequently, it has distinct characteristics and can be treated in a different manner.

Additionally, a set of interviews with different stakeholders and domain experts was conducted to complement the research findings on activity and behavior profiling concerning missing children. The information identified as crucial from both the interviews and the literature review provides the necessary recommendations for building a complete profile of a missing child case. Such a profile, should contain situational details of the case (such as the place the child was last seen) as well as more stationary data concerning the child (e.g. demographics, psychological reports, social status, etc). Applying the suggested indicators to the case, enables in the first place a more efficient data analytics process, and ultimately a timelier and effective recovery of the child.

As the data that is needed to create informative profiles of the children's situation and their activity or behavioural patterns is highly sensitive in nature (i.e. family issues, health status, personal data etc.), ChildRescue will take strict safekeeping measures to minimise the risk of breaching the data security and protect data privacy. The recommended mechanisms for this purpose are related to encryption and anonymisation techniques, and are presented in this document. The algorithms and modules to be implemented were carefully selected so as to be in line with the technical requirements that derive from the need for GDPR conformity. Since these modules will be used any time personal data is going to be stored, communicated and processed, ChildRescue should be GDPR compliant concerning data privacy requirements. Apart from pseudonymisation and anonymisation techniques, ChildRescue will also implement consent monitoring functionalities that will lead to data erasure upon consent revocation. This process will offer ChildRescue the provision of the required data control to data subjects.

Regarding the analysis of the data, an extensive state-of-play analysis on the relevant fields of interest identified useful methods and technologies for processing behavioural profiles, mobility patterns, and social networking data. Several academic papers were reviewed and compared, presenting the most significant aspects of the three selected areas.

The step following the research analysis was to define the appropriate procedures that can lead to useful data analytics able to assist the ChildRescue cause. A robust and uniform way to collect data from past cases maintained by the pilot partners was initially proposed, and based on research findings, the algorithms and tools that can manipulate and analyse the available data were identified. Three axes were specified where the power of analytics could be applied: the knowledge extraction and predictions deriving from profiling data, the evaluation and validation of incoming information, and the estimation of routes and points of interests related to the missing children investigation. Using all of these tools successfully, in the framework of ChildRescue, will certainly lead to a more efficient investigation process of a missing child or the tracing of an unaccompanied migrant minor.

The results reported in this deliverable, will play an important role in the tasks of "WP3- ChildRescue Platform Architecture Definition and Implementation". Especially in Task 3.1, during the design of the platform's architecture, the proposed methods and algorithms for data analytics as well as for data protection offer several avenues of well-documented good practices and design patterns which can be considered and, eventually, adopted in ChildRescue implementation. Furthermore, the evaluation of the methodology by the end-users during the "WP4- Missing Persons Cases Piloting and Evaluation" phase, will provide possible modifications and updates for optimising ChildRescue mechanisms and data operations.

The final, incorporated, methodology, along with the initial input and provided feedback or updates from pilots' experiences will be presented in deliverable "D2.5 - Profiling, Analytics and Privacy Methodological Foundations, Release II" on M24.

Annex I: References

- [1] Cancedda, Alessandra; Day, Laurie; Dimitrova, Dafina; Gosset, Martin (2013): Missing children in the European Union: Mapping, data collection and statistics. ECORYS Nederland BVCrosland, Kimberly; Joseph, Ruby; Slattery, Lindsey; Hodges, Sharon; Dunlap, Glen (2018): Why youth run: Assessing run function to stabilize foster care placement. In: *Children and Youth Services Review* 85, 35–42. DOI: 10.1016/j.childyouth.2017.12.002.
- [2] Payne, Malcolm (1995): Understanding 'Going Missing': issues for social work and social services. In: *British Journal of Social Work* 25, 333–348.
- [3] Warren, Janet I.; Wellbeloved-Stone, James M.; Hilts, Mark A.; Donaldson, William H.; Muirhead, Yvonne E.; Craun, Sarah W. et al. (2016): An investigative analysis of 463 incidents of single-victim child abductions identified through Federal Law Enforcement. In: *Aggression and Violent Behavior* 30, 59–67. DOI: 10.1016/j.avb.2016.07.006
- [4] Beasley, James Oliver; Hayne, Anita S.; Beyer, Kristen; Cramer, Gary L.; Berson, Sarah Bradley; Muirhead, Yvonne; Warren, Janet I. (2009): Patterns of prior offending by child abductors. A comparison of fatal and non-fatal outcomes. In: *International journal of law and psychiatry* 32 (5), 273–280. DOI: 10.1016/j.ijlp.2009.06.009.
- [5] Kaptelinin, Victor (1993): Activity Theory: Implications for Human Computer Interaction. In: M. D. Janse und T. L. Harrington (Hg.): *Human-machine communication for educational systems design. Part 1: Fundamentals of Human Perception and Reasoning*. Eindhoven (IPO manuscript), 15-16.
- [6] Er, Michael (2014): Activity Theory. In: David Coghlan und Mary Brydon-Miller (Hg.): *The SAGE encyclopedia of action research*. Los Angeles, Calif.: SAGE Publ (SAGE reference).
- [7] Crosland, Kimberly; Joseph, Ruby; Slattery, Lindsey; Hodges, Sharon; Dunlap, Glen (2018): Why youth run: Assessing run function to stabilize foster care placement. In: *Children and Youth Services Review* 85, 35–42. DOI: 10.1016/j.childyouth.2017.12.002.
- [8] White, Leroy; Burger, Katharina; Yearworth, Mike (2016): Understanding behaviour in problem structuring methods interventions with activity theory. In: *European Journal of Operational Research* 249 (3), 983–1004. DOI: 10.1016/j.ejor.2015.07.044.
- [9] Park, Robert (1927): Human Nature and collective behaviour. In: *American Journal of Sociology* 1927 (32), 733-745.
- [10] Pang, N. ; Goh, D. (2016): Can blogs function as rhetorical publics in Asian democracies? In: *Telematics and Informatics* (33).
- [11] Beierle, Sarah; Hoch, Carolin (2017): *Straßenjugendliche in Deutschland. Forschungsergebnisse und Empfehlungen*. DJi: München.
- [12] Scott, J. (2013) *Social Network Analysis*. 3rd ed. London: SAGE.
- [13] Bilecen, Başak (2013): *Analyzing Informal Social Protection Across Borders: Synthesizing Social Network Analysis with Qualitative Interviews* (SFB 882 Working Paper Series, 19).

- [14] Borkert, Maren; Fisher, Karen E.; Yafi, Eiad (2018): The Best, the Worst, and the Hardest to Find: How People, Mobiles, and Social Media Connect Migrants In(to) Europe. In: *Social Media + Society* 4 (1), 1-11. DOI: 10.1177/2056305118764428.
- [15] Frith, Emily (2017): Social media and children's mental health: a review of the evidence. Education Policy Institute.
- [16] Young, Sean D. (2014): Behavioral insights on big data: using social media for predicting biomedical outcomes. In: *Trends in microbiology* 22 (11), 601–602. DOI: 10.1016/j.tim.2014.08.004.
- [17] Burke, Roger (2009): *Introduction to Criminological Theory*. Willan Publishing.
- [18] Cohen, Albert (1972): Social Control and Subcultural Change. In: *Youth & Society* March 1972.
- [19] Schneider, Hans Joachim (2010): Neue Erkenntnisse der kriminologischen Verbrechensopferforschung – ihre Auswirkungen auf die Opferpolitik. In: *Juristische Rundschau* 2010 (9).
- [20] ICMEC (2016): *Missing Children Assessment and Recommendations Best Practices Guide*. Belarus, Canada, Finland, Kazakhstan, Russia, and the United States. Hg. v. ICMEC.
- [21] Jones Johnson, Regina; Rew, Lynn; Kouzekanani, Kamiar (2006): Gender differences in victimised homeless adolescents. In: *ADOLESCENCE* 41 (161), 39-53.
- [22] Warren, Janet; Wellbeloved-Stone, James; Hilts, Mark; Donaldson, William; Muirhead, Yvonne; Craun, Sarah; Burnette, Anna grace; Millspaugh, Sara (2016): An investigative analysis of 463 incidents of single-victim child abductions identified through Federal Law Enforcement. In: *Agression and Violent Behavior* 30, 59-67.
- [23] O. Osonde and D. Paul, "An Artificial Intelligence/Machine Learning Perspective on Social Simulation: New Data and New Challenges," RAND Corp., 2018.
- [24] E. Kolaczyk, *Statistical analysis of network data*. 2009.
- [25] A. Papachristos, "The coming of a networked criminology," *Meas. crime Crim.*, 2011.
- [26] O. Gallupe, *Network Analysis*. 2016.
- [27] G. Oatley and B. Ewart, "Data mining and crime analysis," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 1, no. 2, pp. 147–153, 2011.
- [28] A. Kriegler, "Using Social Network Analysis to Profile Organised Crime," *ISS Secur. Stud.*, 2014.
- [29] A. Rattinger, G. Wallner, A. Drachen, J. Pirker, and R. Sifa, Integrating and inspecting combined behavioral profiling and social network models in *Destiny*, vol. 9926 LNCS. 2016.
- [30] G. U. Vasanthakumar, P. Deepa Shenoy, and V. K. R., "An Overview on User Profiling in Online Social Networks," *Int. J. Appl. Inf. Syst.*, 2017.
- [31] E. Cambria, B. Schuller, Y. Xia, and C. Havasi, "New avenues in opinion mining and sentiment analysis," *IEEE Intell. Syst.*, vol. 28, no. 2, pp. 15–21, 2013.
- [32] R. PLUTCHIK, "A GENERAL PSYCHOEVOLUTIONARY THEORY OF EMOTION," in *Theories of*

Emotion, 1980.

- [33] H. S. Ibrahim, S. M. Abdou, and M. Gheith, "Sentiment Analysis for Modern Standard Arabic and Colloquial," *Int. J. Nat. Lang. Comput.*, 2015.
- [34] A. Banerjee, T. Bandyopadhyay, and P. Acharya, "Data Analytics: Hyped Up Aspirations or True Potential?," *Vikalpa*, vol. 38, no. 4, pp. 1–12, 2013.
- [35] A. McAfee and E. Brynjolfsson, "Big data: the management revolution.," *Harv. Bus. Rev.*, vol. 90, no. 10, pp. 62–68, 2012.
- [36] A. L. Samuel, "Some Studies in Machine Learning Using the Game of Checkers," *IBM J. Res. Dev.*, 1959.
- [37] T. M. Mitchell, *Machine Learning*. 1997.
- [38] W. Bleidorn and C. J. Hopwood, "Using Machine Learning to Advance Personality Assessment and Theory," *Personal. Soc. Psychol. Rev.*, 2018.
- [39] M. Kosinski, D. Stillwell, and T. Graepel, "Private traits and attributes are predictable from digital records of human behavior," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 110, no. 15, pp. 5802–5805, 2013.
- [40] J. S. Hartford, J. R. Wright, and K. Leyton-Brown, "Deep Learning for Predicting Human Strategic Behavior," *Adv. Neural Inf. Process. Syst.* 29, 2016.
- [41] D. Silver et al., "Mastering the game of Go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, 2016.
- [42] K. Blackmore, T. Bossomaier, S. Foy, and D. Thomson, "Data Mining of Missing Persons Data," Springer, Berlin, Heidelberg, 2005, pp. 305–314.
- [43] Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*, 15(5), 662-679.
- [44] Deka, G. C. (2016). Big data predictive and prescriptive analytics. In *Big Data: Concepts, Methodologies, Tools, and Applications* (pp. 30-55). IGI Global.
- [45] Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35(2), 137-144.
- [46] McCue, C. (2014). *Data mining and predictive analysis: Intelligence gathering and crime analysis*. Butterworth-Heinemann.
- [47] Jonas, J., & Harper, J. (2006). *Effective counterterrorism and the limited role of predictive data mining*. Washington DC: Cato Institute.
- [48] Perry, W. L. (2013). *Predictive policing: The role of crime forecasting in law enforcement operations*. Rand Corporation.
- [49] Pramanik, M. I., Lau, R. Y., Yue, W. T., Ye, Y., & Li, C. (2017). Big data analytics for security and criminal investigations. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 7(4), e1208.

- [50] Valiant, L. G. (1984). A theory of the learnable. *Communications of the ACM*, 27(11), 1134-1142.
- [51] Angluin, D. (1992, July). Computational learning theory: survey and selected bibliography. In *Proceedings of the twenty-fourth annual ACM symposium on Theory of computing* (pp. 351-369). ACM.
- [52] Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press.
- [53] Afsar, P., Cortez, P., & Santos, H. (2015). Automatic visual detection of human behavior: a review from 2000 to 2014. *Expert Systems with Applications*, 42(20), 6935-6956.
- [54] Atallah, L., & Yang, G. Z. (2009). The use of pervasive sensing for behaviour profiling—a survey. *Pervasive and Mobile Computing*, 5(5), 447-464.
- [55] Pantic, M., Pentland, A., Nijholt, A., & Huang, T. S. (2007). Human computing and machine understanding of human behavior: a survey. In *Artificial Intelligence for Human Computing* (pp. 47-71). Springer, Berlin, Heidelberg
- [56] Jin, L., Chen, Y., Wang, T., Hui, P., & Vasilakos, A. V. (2013). Understanding user behavior in online social networks: A survey. *IEEE Communications Magazine*, 51(9), 144-150.
- [57] Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge & Data Engineering*, (6), 734-749.
- [58] Brooks, R. A. (1991). Intelligence without representation. *Artificial intelligence*, 47(1-3), 139-159.
- [59] Langley, P., Laird, J. E., & Rogers, S. (2009). Cognitive architectures: Research issues and challenges. *Cognitive Systems Research*, 10(2), 141-160.
- [60] Søraker, J. H., & Brey, P. (2007). Ambient intelligence and problems with inferring desires from behaviour. *International review of information ethics*, 8(1), 7-12.
- [61] Hildebrandt, M., & Gutwirth, S. (2008). *Profiling the European citizen*. Dordrecht: Springer.
- [62] Anrig, B., Browne, W., & Gasson, M. (2008). The role of algorithms in profiling. In *Profiling the European Citizen* (pp. 65-87). Springer, Dordrecht.
- [63] Hildebrandt, M. (2008). Defining profiling: a new type of knowledge?. In *Profiling the European citizen* (pp. 17-45). Springer, Dordrecht.
- [64] Srividya, M., Mohanavalli, S., & Bhalaji, N. (2018). Behavioral Modeling for Mental Health using Machine Learning Algorithms. *Journal of medical systems*, 42(5), 88.
- [65] Harley, J. M., Trevors, G. J., & Azevedo, R. (2013). Clustering and profiling students according to their interactions with an intelligent tutoring system fostering self-regulated learning. *JEDM| Journal of Educational Data Mining*, 5(1), 104-146.
- [66] Han, X., Wang, L., & Huang, H. (2017). Deep Investment Behavior Profiling by Recurrent Neural Network in P2P Lending.
- [67] Chen, Y., Pavlov, D., & Canny, J. F. (2009, June). Large-scale behavioral targeting. In *Proceedings*

- of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 209-218). ACM.
- [68] Rattinger, A., Wallner, G., Drachen, A., Pirker, J., & Sifa, R. (2016, September). Integrating and inspecting combined behavioral profiling and social network models in destiny. In *International Conference on Entertainment Computing* (pp. 77-89). Springer, Cham.
- [69] Yeung, D. Y., & Ding, Y. (2003). Host-based intrusion detection using dynamic and static behavioral models. *Pattern recognition*, 36(1), 229-243.
- [70] Fawcett, T., & Provost, F. J. (1996, December). Combining Data Mining and Machine Learning for Effective User Profiling. In KDD (pp. 8-13).
- [71] de Montjoye, Y. A., Quoidbach, J., Robic, F., & Pentland, A. S. (2013, April). Predicting personality using novel mobile phone-based metrics. In *International conference on social computing, behavioral-cultural modeling, and prediction* (pp. 48-55). Springer, Berlin, Heidelberg.
- [72] Chittaranjan, G., Blom, J., & Gatica-Perez, D. (2011, June). Who's who with big-five: Analyzing and classifying personality traits with smartphones. In *Wearable Computers (ISWC), 2011 15th Annual International Symposium on* (pp. 29-36). IEEE.
- [73] Nath, S. V. (2006, December). Crime pattern detection using data mining. In *Web intelligence and intelligent agent technology workshops, 2006. wi-iat 2006 workshops. 2006 ieee/wic/acm international conference on* (pp. 41-44). IEEE.
- [74] Blackmore, K. L., & Bossomaier, T. R. J. (2003). Soft computing methodologies for mining missing person data. *INTERNATIONAL JOURNAL OF KNOWLEDGE BASED INTELLIGENT ENGINEERING SYSTEMS*, 7(3), 132-138.
- [75] Bachrach, Y., Kosinski, M., Graepel, T., Kohli, P., & Stillwell, D. (2012, June). Personality and patterns of Facebook usage. In *Proceedings of the 4th annual ACM web science conference* (pp. 24-32). ACM.
- [76] Oberlander, J., & Nowson, S. (2006, July). Whose thumb is it anyway?: classifying author personality from weblog text. In *Proceedings of the COLING/ACL on Main conference poster sessions* (pp. 627-634). Association for Computational Linguistics.
- [77] Borges, J., & Levene, M. (1999, August). Data mining of user navigation patterns. In *International Workshop on Web Usage Analysis and User Profiling* (pp. 92-112). Springer, Berlin, Heidelberg.
- [78] Turaga, P., Chellappa, R., Subrahmanian, V. S., & Udrea, O. (2008). Machine recognition of human activities: A survey. *IEEE Transactions on Circuits and Systems for Video technology*, 18(11), 1473.
- [79] Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist*, 48, 26-34.
- [80] Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, 12(6), 1100-1122.
- [81] Bleidorn, W., & Hopwood, C. J. (2018). Using Machine Learning to Advance Personality Assessment and Theory. *Personality and Social Psychology Review*, 1088868318772990.

- [82] Agrawal, R., & Srikant, R. (2000). *Privacy-preserving data mining* (Vol. 29, No. 2, pp. 439-450). ACM.
- [83] Aggarwal, C. C., & Philip, S. Y. (2008). A general survey of privacy-preserving data mining models and algorithms. In *Privacy-preserving data mining* (pp. 11-52). Springer, Boston, MA.
- [84] Brody, H., Rip, M. R., Vinten-Johansen, P., Paneth, N., & Rachman, S. (2000). Map-making and myth-making in Broad Street: the London cholera epidemic, 1854. *The Lancet*, *356*(9223), 64-68.
- [85] RF, Tomlinson. (1969). A geographic information system for regional planning. *Journal of Geography (Chigaku Zasshi)*, *78*(1), 45-48.
- [86] Lee, E. S. (1966). A theory of migration. *Demography*, *3*(1), 47-57.
- [87] Toch, E., Lerner, B., Ben-Zion, E., & Ben-Gal, I. (2018). Analyzing large-scale human mobility data: a survey of machine learning methods and applications. *Knowledge and Information Systems*, 1-23.
- Shen, L., & Stopher, P. R. (2014). Review of GPS travel survey and GPS data-processing methods. *Transport Reviews*, *34*(3), 316-334.
- [88] Gonzalez, M. C., Hidalgo, C. A., & Barabasi, A. L. (2008). Understanding individual human mobility patterns. *nature*, *453*(7196), 779.
- [89] Bengtsson, L., Lu, X., Thorson, A., Garfield, R., & Von Schreeb, J. (2011). Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: a post-earthquake geospatial study in Haiti. *PLoS medicine*, *8*(8), e1001083.
- [90] Massimo, D., & Ricci, F. (2018, September). Harnessing a generalised user behaviour model for next-POI recommendation. In *Proceedings of the 12th ACM Conference on Recommender Systems* (pp. 402-406). ACM.
- [91] Cho, E., Myers, S. A., & Leskovec, J. (2011, August). Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1082-1090). ACM.
- [92] Chen, N. C., Xie, W., Welsch, R. E., Larson, K., & Xie, J. (2017, June). Comprehensive Predictions of Tourists' Next Visit Location Based on Call Detail Records Using Machine Learning and Deep Learning Methods. In *Big Data (BigData Congress), 2017 IEEE International Congress on* (pp. 1-6). IEEE.
- [93] Yuan, J., Zheng, Y., & Xie, X. (2012, August). Discovering regions of different functions in a city using human mobility and POIs. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 186-194). ACM.
- [94] Ashbrook, D., & Starner, T. (2003). Using GPS to learn significant locations and predict movement across multiple users. *Personal and Ubiquitous computing*, *7*(5), 275-286.
- [95] Mohibullah, W., & Julie, S. J. (2013, October). Developing an agent model of a missing person in the wilderness. In *Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on* (pp. 4462-4469). IEEE.
- [96] Biljecki, F., Ledoux, H., & Van Oosterom, P. (2013). Transportation mode-based segmentation and

- classification of movement trajectories. *International Journal of Geographical Information Science*, 27(2), 385-407.
- [97] Liu, Q., Wu, S., Wang, L., & Tan, T. (2016, February). Predicting the Next Location: A Recurrent Model with Spatial and Temporal Contexts. In *AAAI* (pp. 194-200).
- [98] Horozov, T., Narasimhan, N., & Vasudevan, V. (2006, January). Using location for personalized POI recommendations in mobile environments. In *Applications and the internet, 2006. SAINT 2006. International symposium on* (pp. 6-pp). IEEE.
- [99] Zheng, Y., Zhang, L., Xie, X., & Ma, W. Y. (2009, April). Mining interesting locations and travel sequences from GPS trajectories. In *Proceedings of the 18th international conference on World wide web* (pp. 791-800). ACM.
- [100] Lin, L., & Goodrich, M. A. (2010). A Bayesian approach to modeling lost person behaviors based on terrain features in wilderness search and rescue. *Computational and Mathematical Organization Theory*, 16(3), 300-323
- [101] Traag, V., Browet, A., Calabrese, F., & Morlot, F. (2011). Social event detection in massive mobile phone data using probabilistic location inference.
- [102] Hasan, S., Schneider, C. M., Ukkusuri, S. V., & González, M. C. (2013). Spatiotemporal patterns of urban human mobility. *Journal of Statistical Physics*, 151(1-2), 304-318.
- [103] Noulas, A., Scellato, S., Lambiotte, R., Pontil, M., & Mascolo, C. (2012). A tale of many cities: universal patterns in human urban mobility. *PLoS one*, 7(5), e37027.
- [104] Preotjuc-Pietro, D., & Cohn, T. (2013, May). Mining user behaviours: a study of check-in patterns in location based social networks. In *Proceedings of the 5th Annual ACM Web Science Conference* (pp. 306-315). ACM.
- [105] Calabrese, F., Di Lorenzo, G., & Ratti, C. (2010). Human mobility prediction based on individual and collective geographical preferences.
- [106] Laube, P., van Kreveld, M., & Imfeld, S. (2005). Finding REMO—detecting relative motion patterns in geospatial lifelines. In *Developments in spatial data handling* (pp. 201-215). Springer, Berlin, Heidelberg.
- [107] Bohte, W., & Maat, K. (2009). Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: A large-scale application in the Netherlands. *Transportation Research Part C: Emerging Technologies*, 17(3), 285-297.
- [108] Koperski, K., Adhikary, J., & Han, J. (1996, June). Spatial data mining: progress and challenges survey paper. In *Proc. ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery, Montreal, Canada* (pp. 1-10).
- [109] Rao, K. V., Govardhan, A., & Rao, K. C. (2012). Spatiotemporal data mining: Issues, tasks and applications. *International Journal of Computer Science and Engineering Survey*, 3(1), 39.
- [110] Gedik, B., & Liu, L. (2005, June). Location privacy in mobile systems: A personalized anonymisation model. In *Distributed Computing Systems, 2005. ICDCS 2005. Proceedings. 25th IEEE International*

Conference on (pp. 620-629). IEEE.

- [111] Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business horizons*, 53(1), 59-68.
- [112] Fan, W., & Gordon, M. D. (2014). The power of social media analytics. *Communications of the ACM*, 57(6), 74-81.
- [113] Zeng, D., Chen, H., Lusch, R., & Li, S. H. (2010). Social media analytics and intelligence. *IEEE Intelligent Systems*, 25(6), 13-16.
- [114] Batrinca, B., & Treleaven, P. C. (2015). Social media analytics: a survey of techniques, tools and platforms. *Ai & Society*, 30(1), 89-116.
- [115] Salathe, M., Bengtsson, L., Bodnar, T. J., Brewer, D. D., Brownstein, J. S., Buckee, C., ... & Vespignani, A. (2012). Digital epidemiology. *PLoS computational biology*, 8(7), e1002616.
- [116] Asur, S., & Huberman, B. A. (2010, August). Predicting the future with social media. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 01* (pp. 492-499). IEEE Computer Society.
- [117] Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., ... & Ungar, L. H. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9), e73791.
- [118] Lima, A. C. E., & De Castro, L. N. (2014). A multi-label, semi-supervised classification approach applied to personality prediction in social media. *Neural Networks*, 58, 122-130.
- [119] Golbeck, J., Robles, C., & Turner, K. (2011, May). Predicting personality with social media. In *CHI'11 extended abstracts on human factors in computing systems* (pp. 253-262). ACM.
- [120] Skowron, M., Tkalčič, M., Ferwerda, B., & Schedl, M. (2016, April). Fusing social media cues: personality prediction from twitter and instagram. In *Proceedings of the 25th international conference companion on world wide web* (pp. 107-108). International World Wide Web Conferences Steering Committee.
- [121] Farnadi, G., Sitaraman, G., Sushmita, S., Celli, F., Kosinski, M., Stillwell, D., ... & De Cock, M. (2016). Computational personality recognition in social media. *User modeling and user-adapted interaction*, 26(2-3), 109-142.
- [122] De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2013). Predicting depression via social media. *ICWSM*, 13, 1-10.
- [123] De Choudhury, M., Kiciman, E., Dredze, M., Coppersmith, G., & Kumar, M. (2016, May). Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI conference on human factors in computing systems* (pp. 2098-2110). ACM.
- [124] Berjani, B., & Strufe, T. (2011, April). A recommendation system for spots in location-based online social networks. In *Proceedings of the 4th Workshop on Social Network Systems* (p. 4). ACM.
- [125] Gao, H., Tang, J., Hu, X., & Liu, H. (2015, January). Content-Aware Point of Interest Recommendation on Location-Based Social Networks. In *AAAI* (pp. 1721-1727).

- [126] Ye, M., Yin, P., & Lee, W. C. (2010, November). Location recommendation for location-based social networks. *In Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems* (pp. 458-461). ACM.
- [127] Bao, J., Zheng, Y., & Mokbel, M. F. (2012, November). Location-based and preference-aware recommendation using sparse geo-social networking data. *In Proceedings of the 20th international conference on advances in geographic information systems* (pp. 199-208). ACM.
- [128] Yang, D., Zhang, D., Yu, Z., & Wang, Z. (2013, May). A sentiment-enhanced personalized location recommendation system. *In Proceedings of the 24th ACM Conference on Hypertext and Social Media* (pp. 119-128). ACM.
- [129] Lu, Z., Wang, H., Mamoulis, N., Tu, W., & Cheung, D. W. (2017). Personalized location recommendation by aggregating multiple recommenders in diversity. *GeoInformatica*, 21(3), 459-484.
- [130] Yao, L., Sheng, Q. Z., Wang, X., Zhang, W. E., & Qin, Y. (2018). Collaborative Location Recommendation by Integrating Multi-dimensional Contextual Information. *ACM Transactions on Internet Technology (TOIT)*, 18(3), 32.
- [131] Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4), 1093-1113.
- [132] Ravi, K., & Ravi, V. (2015). A survey on opinion mining and sentiment analysis: tasks, approaches and applications. *Knowledge-Based Systems*, 89, 14-46.
- [133] Paltoglou, G., & Thelwall, M. (2012). Twitter, MySpace, Digg: Unsupervised sentiment analysis in social media. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(4), 66.
- [134] dos Santos, C., & Gatti, M. (2014). Deep convolutional neural networks for sentiment analysis of short texts. *In Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers* (pp. 69-78).
- [135] Bouazizi, M., & Ohtsuki, T. (2017). A pattern-based approach for multi-class sentiment analysis in twitter. *IEEE Access*, 5, 20617-20639.
- [136] Pak, A., & Paroubek, P. (2010, May). Twitter as a corpus for sentiment analysis and opinion mining. *In LREc* (Vol. 10, No. 2010, pp. 1320-1326).
- [137] Ortigosa, A., Martín, J. M., & Carro, R. M. (2014). Sentiment analysis in Facebook and its application to e-learning. *Computers in human behavior*, 31, 527-541.
- [138] You, Q., Luo, J., Jin, H., & Yang, J. (2015, January). Robust Image Sentiment Analysis Using Progressively Trained and Domain Transferred Deep Networks. *In AAAI* (pp. 381-388).
- [139] Wang, Y., Wang, S., Tang, J., Liu, H., & Li, B. (2015, July). Unsupervised Sentiment Analysis for Social Media Images. *In IJCAI* (pp. 2378-2379).
- [140] Kalampokis, E., Tambouris, E., & Tarabanis, K. (2013). Understanding the predictive power of social media. *Internet Research*, 23(5), 544-559.
- [141] Tufekci, Z. (2014). Big Questions for Social Media Big Data: Representativeness, Validity and Other

Methodological Pitfalls. ICWSM, 14, 505-514.

- [142] Metaxas, P. T., Mustafaraj, E., & Gayo-Avello, D. (2011, October). How (not) to predict elections. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom)*, 2011 IEEE Third International Conference on (pp. 165-171). IEEE.
- [143] Jagwani, Priti & Kaushik, Saroj. (2017). *Privacy in Location Based Services: Protection Strategies, Attack Models and Open Challenges*. 12-21. 10.1007/978-981-10-4154-9_2.
- [144] Duckham M., Kulik L. (2005) A Formal Model of Obfuscation and Negotiation for Location Privacy. In: Gellersen H.W., Want R., Schmidt A. (eds) *Pervasive Computing. Pervasive 2005. Lecture Notes in Computer Science*, vol 3468. Springer, Berlin, Heidelberg
- [145] U.S. Department of Health and Human Services Office for Civil Rights. *HIPAA Administrative Simplification Regulation*, 45 CFR Parts 160, 162, and 164. 2013
- [146] Benitez K, Malin B. Evaluating re-identification risks with respect to the HIPAA Privacy Rule. *J Am Med Inform Assoc*. 2010;17(2):169–77.
- [147] Weaving technology and policy together to maintain confidentiality. Sweeney, L. *Journal of Law, Medicine and Ethics*. 1997, 25:98-110. (impact factor 1.04) Cited and discussed in the commentary of the HIPAA Privacy Rule.
- [148] Sweeney, L. (2002). k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*. 10. 10.1142/S0218488502001648.
- [149] CHATURVEDI, Dr. SETU. (2012). An Optimization of Association Rule Mining using K-Map and Genetic Algorithm for Large Database. *International Journal of Computer Applications*. 84. 10.5120/14680-2143.
- [150] Machanavajjhala, Ashwin; Kifer, Daniel; Gehrke, Johannes; Venkatasubramanian, Muthuramakrishnan (March 2007). "L-diversity: Privacy Beyond K-anonymity". *ACM Trans. Knowl. Discov. Data*.
- [151] Wong R.CW., Liu Y., Yin J., Huang Z., Fu A.WC., Pei J. (2007) (α , k)-anonymity Based Privacy Preservation by Lossy Join. In: Dong G., Lin X., Wang W., Yang Y., Yu J.X. (eds) *Advances in Data and Web Management. APWeb 2007, WAIM 2007. Lecture Notes in Computer Science*, vol 4505. Springer, Berlin, Heidelberg
- [152] Li, Ninghui & Li, Tiancheng & Venkatasubramanian, Suresh. (2007). t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. *IEEE 23rd International Conference on Data Engineering (ICDE)*. 2. 106 - 115. 10.1109/ICDE.2007.367856.
- [153] Mehmet Ercan Nergiz, Maurizio Atzori, Chris Clifton: Hiding the presence of individuals from shared databases. *SIGMOD Conference 2007*: 665-676.
- [154] Xiao, Zhen & Xu, Jianliang & Meng, Xiaofeng. (2008). p-Sensitivity: A Semantic Privacy-Protection Model for Location-based Services. 47 - 54. 10.1109/MDMW.2008.20.
- [155] Bettini C., Mascetti S., Wang X.S., Freni D., Jajodia S. (2009) Anonymity and Historical-Anonymity in Location-Based Services. In: Bettini C., Jajodia S., Samarati P., Wang X.S. (eds) *Privacy in Location-*

Based Applications. Lecture Notes in Computer Science, vol 5599. Springer, Berlin, Heidelberg.

- [156] Data Cube: A Relational Aggregation Operator Generalizing Group-By, Cross-Tab, and Sub-Totals, Jim Gray, Surajit Chaudhuri, Adam Bosworth, Andrew Layman, Don Reichart, Murali Venkatrao, Frank Pellow, Hamid Pirahesh, January 1997, Data Mining and Knowledge Discovery: Volume 1 Issue 1, 1997.
- [157] Ghinita, Gabriel & Zhao, Keliang & Papadias, Dimitris & Kalnis, Panos. (2010). A reciprocal framework for spatial K-anonymity. Cyber Center Publications. 35. 10.1016/j.is.2009.10.001.
- [158] Hayashida, Shuhei & Amagata, Daichi & Hara, Takahiro & Xie, Xing. (2018). Dummy Generation Based on User-Movement Estimation for Location Privacy Protection. IEEE Access. PP. 1-1. 10.1109/ACCESS.2018.2829898.
- [159] Tian F, Gui X, An J, Yang P, Zhao J, Zhang X. Protecting location privacy for outsourced spatial data in cloud storage. ScientificWorldJournal. 2014;2014:108072.
- [160] ARX: Data Anonymisation Tool <https://arx.deidentifier.org/>
- [161] ARX: Privacy Models <https://arx.deidentifier.org/overview/privacy-criteria/>
- [162] Dankar, Fida Kamal, Khaled El Emam, Angelica Neisa and Tyson Roffey. "Estimating the re-identification risk of clinical data sets." BMC Med. Inf. & Decision Making (2012).
- [163] Brickell, Justin & Shmatikov, Vitaly. (2008). The cost of privacy: Destruction of data-mining utility in anonymized data publishing. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 70-78. 10.1145/1401890.1401904.
- [164] Bild, Raffael & A Kuhn, Klaus & Prasser, Fabian. (2018). SafePub: A Truthful Data Anonymisation Algorithm With Strong Privacy Guarantees. Proceedings on Privacy Enhancing Technologies. 2018. 10.1515/popets-2018-0004.
- [165] Khaled El Emam, Ph.D. (University of Ottawa) and Bradley Malin, Ph.D. (Vanderbilt University): CONCEPTS AND METHODS FOR DE-IDENTIFYING CLINICAL TRIAL DATA.
- [166] Wan, Zhiyu & Vorobeychik, Yevgeniy & Xia, Weiyi & Clayton, Ellen & Kantarcioglu, Murat & Ganta, Ranjit & Heatherly, Raymond & Malin, Bradley. (2015). A Game Theoretic Framework for Analyzing Re-Identification Risk. PloS one. 10. e0120592. 10.1371/journal.pone.0120592.
- [167] G. Yuan, F. Wang and Y. Hao, "Research on Data Encryption Technology Based on Chaos Theory," Eighth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing (SNPD 2007), Qingdao, 2007, pp. 93-98. doi: 10.1109/SNPD.2007.57
- [168] Robak, Martin (2003): Profiling: Täterprofile und Fallanalysen als Unterstützung strafprozessualer Ermittlungen. Polizeiliche Methoden und deren kriminalpolitische Bedeutung. Münster.
- [169] Shankar, Ravi; Gadkar, Ravindra (2015): Family Factors and Runaway Missing Children: A Review of Theories and Research. In: *International Journal of Management Research and Social Science* 2 (4), pp. 115–119.

- [170] Mislove, A., Viswanath, B., Gummadi, K. P., & Druschel, P. (2010, February). You are who you know: inferring user profiles in online social networks. In *Proceedings of the third ACM international conference on Web search and data mining* (pp. 251-260). ACM.
- [171] Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3), 15
- [172] Surowiecki, J. (2004). The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business. *Economies, Societies and Nations*, 296.
- [173] Alag, S. (2009). *Collective intelligence in action* (pp. 274-306). Greenwich, CT: Manning.
- [174] Wang, D., Abdelzaher, T., Kaplan, L., & Aggarwal, C. C. (2013, July). Recursive fact-finding: A streaming approach to truth estimation in crowdsourcing applications. In *Distributed Computing Systems (ICDCS), 2013 IEEE 33rd International Conference on* (pp. 530-539). IEEE.
- [175] Kreps, G. A. (1984). Sociological inquiry and disaster research. *Annual review of sociology*, 10(1), 309-330.
- [176] Vieweg, S., Palen, L., Liu, S. B., Hughes, A. L., & Sutton, J. N. (2008). *Collective intelligence in disaster: Examination of the phenomenon in the aftermath of the 2007 Virginia Tech shooting*. Boulder, CO: University of Colorado.
- [177] Cardone, G., Cirri, A., Corradi, A., Foschini, L., Ianniello, R., & Montanari, R. (2014). Crowdsensing in urban areas for city-scale mass gathering management: Geofencing and activity recognition. *IEEE Sensors Journal*, 14(12), 4185-4195.
- [178] Garg, A., Choudhary, S., Bajaj, P., Agrawal, S., Kedia, A., & Agrawal, S. (2017, November). Smart Geo-fencing with Location Sensitive Product Affinity. In *Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (p. 39). ACM.

Annex II: Analysis of Research Literature

II.1 Social Science theories applicable to cases of missing children

Table Annex II- 1 Social Science applicable to missing children cases

Theory	Core Assumptions	Limitations	Applicable to:
Activity Theory	<ul style="list-style-type: none"> - Behaviour as result of correlation between individual and natural environment - Consciousness is shaped by cultural and technical environment and influences unrelated situations (which tools are available) 	<ul style="list-style-type: none"> - Underestimates ability of youth to find resources outside their environment/ formerly unknown - Relies heavily on social influence, might undervalue influence of individual character 	Runaways (specifically pushaways and throwaways), unaccompanied migrant minors, parental abductions
Collective Behaviour Theory	<ul style="list-style-type: none"> - Analysis of patterns of behaviour in (loosely) organised groups - Behaviour as result of group dynamics, individual not regarded 	<ul style="list-style-type: none"> - No analysis of individual pattern of behaviour - Group behaviour requires time to fully form (limited applicability to missing children cases) 	Runaways (Street youth living in gangs), unaccompanied minor migrants (if travelling in collectives)
Social Network Analysis	<ul style="list-style-type: none"> - Analysis of patterned relationships in social networks - Behaviour as result of social relations rather than individual character 	<ul style="list-style-type: none"> - Underestimates the individual's potential to act 'unusual' in social relations - Relies on correct reports of relations within the social network 	Runaways (specifically throwaways and pushaways), unaccompanied migrant minors (also online social network), Parental abductions, Third-person abductions
Subcultural Theory	<ul style="list-style-type: none"> - Behaviour as result of norms and social rules that apply in subculture 	<ul style="list-style-type: none"> - Not useful in explaining behaviour untypical for subculture 	Runaways, unaccompanied migrant minors

	<ul style="list-style-type: none">- Focusses on interests and available alternatives for members of subculture		
Victimology	<ul style="list-style-type: none">- Analysis of characteristics of victims and victim-offender dynamic- Offers insight into risk behaviours/ vulnerability factors	<ul style="list-style-type: none">- Discredits individual agency of victims- Risk of overlooking 'atypical' victims	Third-person abductions

II.2 Creating Activity and Behaviour Profiles of Missing Children

A comparison list of the studied literature is presented in the following table. Each row describes one scientific paper and the related methodology using the following fields:

Title [Reference]: Article title and reference to related citation.

Year: The year the article was published.

Category: Depending on the focus of the article, this field contains the category of the algorithms presented in the article. The categorisation is made in respect to the tasks involved, such as classification, regression, clustering, statistical analysis, etc.

Algorithm / Method: The algorithm/method suggested by the article, along with proposed adaptations for enhancement of results or mitigation of computation complexity.

Experimentation dataset: The source and form of the data used to test and evaluate the proposed methodology, if any.

Primary focus: The general research field of the paper and any special focus.

Results: Brief summarisation of results, conclusions and key points.

Table Annex II- 2 Literature review for the creation of activity and behaviour profiles of missing children

Title [reference]	Year	Category	Algorithm/ Method	Experimentation dataset	Primary focus	Results
Behavioral Modeling for Mental Health using Machine Learning Algorithms [64]	2018	<ul style="list-style-type: none"> • Clustering • Classification 	<ul style="list-style-type: none"> • K-means, Agglomerative, K-medoids. • Logistic regression, Naïve Bayes, Support Vector Machines (SVM), Decision tree, K-nearest neighbours, ensemble bagging and random forest 	656 individuals, 20 features, 3 class labels	Identify state of mental health by creating mental health profiles	<ul style="list-style-type: none"> • Clusters individuals based on responses in questionnaire into 3 classes (mentally distressed, neutral and happy). • SVM, k-NN, ensembles and random forest outperform other classifiers. • The inclusion of physiological parameters is recommended.
Deep Investment Behavior Profiling by Recurrent Neural Network in P2P Lending [66]	2017	<ul style="list-style-type: none"> • Deep learning 	<ul style="list-style-type: none"> • Recurrent Neural Networks (RNNs) with GRU and LSTM 	Dataset of 7455 investors of Prosper platform (16 features of investor and invest details)	User Profiling and time-series prediction	<ul style="list-style-type: none"> • Profiling users' investment preferences and forecasting trends. • Considers it a time-series analysis problem. • Compares with k-NN and Bayesian structural time-series.
Integrating and inspecting combined behavioral profiling and social network models in destiny [68]	2016	<ul style="list-style-type: none"> • Clustering 	<ul style="list-style-type: none"> • Archetypal analysis 	10,000 players of Destiny game (performance and networking data)	Constructing behavioural profiles based on on-line gaming data	<ul style="list-style-type: none"> • Clusters players into five profiles based on their gaming performance, as well as their social networking inside the game. • Interesting patterns were extracted.
Clustering and Profiling Students According to their Interactions with an Intelligent Tutoring	2013	<ul style="list-style-type: none"> • Clustering 	<ul style="list-style-type: none"> • Expectation-Minimisation (EM) 	Data from 106 college students (12 variables) containing their interactions (log file) with an e-	Design adaptive systems relying on behavioural profiles	<ul style="list-style-type: none"> • Analysis of student profiles derived from clustering methods. • Additional data sources are suggested, such as gaze behaviour and emotions (from analysis of

System Fostering Self-Regulated Learning [65]				learning system		video recordings).
Private traits and attributes are predictable from digital records of human behavior [39]	2013	<ul style="list-style-type: none"> • Dimensionality Reduction • Regression • Classification 	<ul style="list-style-type: none"> • Singular-value decomposition (SVD) • Linear regression (for predicting numeric values, e.g. age) • Logistic regression (for predicting binary variables, e.g. gender) 	Around 58.500 Facebook profiles from "myPersonality" app	Psycho-demographic profile extraction from Facebook Likes	<ul style="list-style-type: none"> • Illustrates the potential of predictive analytics in today's digital society. • Predicts individual traits and attributes, such as religion, sexuality, political views, relationship status, alcohol consumption, age, personality scores, etc. • Application in marketing and recommender systems.
Predicting personality using novel mobile phone-based metrics [71]	2013	<ul style="list-style-type: none"> • Classification 	<ul style="list-style-type: none"> • Support Vector Machines (SVM) 	69 individuals with their mobile data records	Behavioural modelling from mobile-phone metrics	<ul style="list-style-type: none"> • Infers user personalities (five-factor model) based on basic information provided by mobile phone usage. • Uses mobile data indicators.
Personality and patterns of Facebook usage [75]	2012	<ul style="list-style-type: none"> • Classification • Natural Language Processing (NLP) • Dimensionality reduction 	<ul style="list-style-type: none"> • Support Vector Machines (SVM), Simple Minimal Optimisation (SMO), MultiBoostAB, AdaBoostM1 	250 user instances with activity and demographic data and ~10,000 status updates	Personality modelling from Facebook data	<ul style="list-style-type: none"> • Achieves high precision in predicting user personality based on big-five model. • A set of 725 features was used broken down into five groups: demographics, activities, status updates, networking data, word classification schemes.
Who's who with big-five: Analyzing and classifying personality traits with smartphones [72]	2011	<ul style="list-style-type: none"> • Classification 	<ul style="list-style-type: none"> • C4.5, Support Vector Machines (SVM) with RBF kernel 	83 participants with mobile usage data collected over 8 months (Call logs, sms, Bluetooth scans, app logs)	Classifying personality traits based on smartphone usage	<ul style="list-style-type: none"> • To determine personality, the authors use the big-five personality framework. • Variety of data sources • Sensorial data from mobile phones may increase performance

Large-scale behavioral targeting [67]	2009	<ul style="list-style-type: none"> • Dimensionality reduction • Regression 	<ul style="list-style-type: none"> • Inverted index • Poisson linear regression 	Yahoo user base (training data: 5-week period, 500 mil. Examples)	Behavioural targeting for ads	<ul style="list-style-type: none"> • Predicts click-through rate (CTR) from user behavioural data. • Large-scale implementation with Hadoop MapReduce framework.
Crime pattern detection using data mining [73]	2006	<ul style="list-style-type: none"> • Clustering 	<ul style="list-style-type: none"> • K-means 	Hundreds of confidential crime cases	Grouping crimes with similar attributes in a geographical region	<ul style="list-style-type: none"> • Identifies crime patterns using clustering techniques. • Assigns results to a geo spatial plot (map).
Whose thumb is it anyway?: classifying author personality from weblog text [76]	2006	<ul style="list-style-type: none"> • Classification • Dimensionality reduction • Natural Language Processing (NLP) 	<ul style="list-style-type: none"> • Support Vector Machines (SVM), Naïve Bayes classifier • N-grams 	71 authors and their blog posts (raw text)	Classification of weblog authors personality	<ul style="list-style-type: none"> • Predicting author personality based on big-five model. • Comparison of binary and multi-class classification.
Data mining of missing persons data [42]	2005	<ul style="list-style-type: none"> • Decision trees • Classification • Rule induction 	<ul style="list-style-type: none"> • C4.5 tree 	Careful selection of 357 missing persons cases	Data mining on missing person profiles	<ul style="list-style-type: none"> • The implemented algorithm produced inconsistent results. • Data used consisted of human, non-structured, judgments and estimations(suspicions) that may have deteriorated data quality.
Host-based intrusion detection using dynamic and static behavioral models [69]	2003	<ul style="list-style-type: none"> • Anomaly detection 	<ul style="list-style-type: none"> • Dynamic model: Hidden Markov models (HMM) with max. likelihood • Static model: Frequency distributions with min. cross entropy 	Unix system calls and shell commands datasets	Behavioural modelling for user profiling	<ul style="list-style-type: none"> • Builds user profiles from Unix system usage data. • Performs intrusion detection based on a dynamic HMM model which outperforms the static model.
Soft computing methodologies for mining missing person data [74]	2003	<ul style="list-style-type: none"> • Classification • Dimensionality reduction 	<ul style="list-style-type: none"> • Artificial Neural Networks (ANNs) • Isomap, PCA 	Selection of 326 missing persons cases	Data mining on missing person profiles	<ul style="list-style-type: none"> • Classify as runaway, suicide, or foulplay. • ANNs achieved high accuracy (93%) that outperforms rule-based

						systems.
Data mining of user navigation patterns [77]	1999	<ul style="list-style-type: none"> • Stochastic modelling • Dimensionality reduction 	<ul style="list-style-type: none"> • Markov chains • N-gram model, Depth-First Search graph 	2-month user navigation log data from websites	Predict users' web navigation	<ul style="list-style-type: none"> • Mining log data for user navigation patterns. • Uses probability grammar modelling to model navigation.
Combining Data Mining and Machine Learning for Effective User Profiling. [70]	1996	<ul style="list-style-type: none"> • Rule induction 	<ul style="list-style-type: none"> • BruteDL 	610 accounts comprising ~350,000 calls over 4 months	Methods for detecting fraudulent behaviour from cellular phone usage	<ul style="list-style-type: none"> • Presents methods for detecting fraudulent usage of mobile phones based on profiling customer behaviour and evidence combination.

II.3 Spatiotemporal Data Analysis

A comparison list of the studied literature is presented in the following table. Each row describes one scientific paper and the related methodology using the following fields:

Title [Reference]: Article title and reference to related citation.

Year: The year the article was published.

Category: Depending on the focus of the article, this field contains the category of the algorithms presented in the article. The categorisation is made in respect to the tasks involved, such as classification, regression, clustering, statistical analysis, etc.

Algorithm / Method: The algorithm/method suggested by the article, along with proposed adaptations for enhancement of results or mitigation of computation complexity.

Experimentation dataset: The source and form of the data used to test and evaluate the proposed methodology, if any.

Primary focus: The general research field of the paper and any special focus.

Results: Brief summarisation of results, conclusions and key points.

Table Annex II- 3 Literature review in Spatiotemporal Data Analysis

Title [reference]	Year	Category	Algorithm/ Method	Experimentation dataset	Primary focus	Results
Harnessing a generalised user behaviour model for next-POI recommendation [90]	2018	<ul style="list-style-type: none"> • Recommender • User behaviour modelling • Clustering • Reinforcement Learning 	<ul style="list-style-type: none"> • Markov decision process • Non-Negative Matrix Factorisation • Inverse Reinforcement Learning 	575 geo-localised trajectories of users' POI-visits,	Recommendations of POIs	<ul style="list-style-type: none"> • Recommendations based on user behaviour (set of trajectories formed by sequences of POIs). • The trajectory generation includes also weather data.
Comprehensive Predictions of Tourists' Next Visit Location Based on Call Detail Records Using Machine Learning and Deep Learning Methods [92]	2017	<ul style="list-style-type: none"> • Classification • Deep Learning 	<ul style="list-style-type: none"> • Decision Tree, Random Forest, Artificial Neural Networks, Naïve Bayes, SVM (with RBF kernel) • Recurrent Neural Network (LSTM) 	16,568,179 data points in January 2015 (Call Data Records)	Location prediction based on mobile phone calls	<ul style="list-style-type: none"> • Predict tourists' next stops using existing stops, other related information and POIs. • Recurrent network outperforms (by 7%) the other methods.
Predicting the Next Location: A Recurrent Model with Spatial and Temporal Contexts [97]	2016	<ul style="list-style-type: none"> • Deep Learning 	<ul style="list-style-type: none"> • Recurrent Neural Network (ST-RNN) 	Two real world datasets: Gowalla location-based social network data and Global Terrorism Database	Prediction of next location from contextual information	<ul style="list-style-type: none"> • Models time intervals in a recurrent architecture. • Incorporates distance-specific transition matrices for modelling geographical distances. • Experiments in real-world datasets show that ST-RNN outperforms other methods.
Spatiotemporal patterns of urban human mobility [102]	2013	<ul style="list-style-type: none"> • Probabilistic Modelling 	<ul style="list-style-type: none"> • Simplistic models 	1,000 users Oyster card transactions	Urban mobility patterns detection and prediction	<ul style="list-style-type: none"> • The article aims to detect spatial and temporal patterns of human mobility in a city using the data from smart subway fare card transactions. • Authors suggest that the heterogeneity of individuals and time dependence in trip

						selection should also be incorporated in the model.
Mining user behaviours: a study of check-in patterns in location based social networks [104]	2013	<ul style="list-style-type: none"> • Clustering • Movement prediction 	<ul style="list-style-type: none"> • K-means • Order-K Markov mode • Use of transition probability matrix between venue categories 	10,000 Foursquare users, each with 1-month trails	Urban mobility patterns detection and prediction	<ul style="list-style-type: none"> • The method first clusters users based on their behavior and then predicts their mobility taking into account temporal periodicities. • Splits venues into 9 categories (e.g. residence, nightlife, shop, work, etc.). • Interesting associations were discovered (e.g. very strong weekly patterns, increase of activity as the week progresses). • Incorporating explicitly the periodicity of user behavior shows to give the best result.
Developing an agent model of a missing person in the wilderness [95]	2013	<ul style="list-style-type: none"> • Stochastic modelling 	<ul style="list-style-type: none"> • Probabilistic Markov model for movement states • Topography transition table for attractive force modelling 	GPS logs of 10 users	Modelling human movement in unfamiliar environment	<ul style="list-style-type: none"> • Complex human behavioural states and interactions with environment are described through an agent-based model (UAV). • Agent model is controlled by a set of strategies and can switch between goals. • Performs better than diffusion model which was less accurate and focused.
Transportation mode-based segmentation and classification of movement trajectories [96]	2013	<ul style="list-style-type: none"> • Classification 	<ul style="list-style-type: none"> • Trajectory segmentation • Fuzzy expert system 	Data from Dutch National Travel survey (benchmark dataset)	Transportation mode estimation based on movement trajectories	<ul style="list-style-type: none"> • Selection of indicators resulted in nine values: 3 speed related (nearly maximum speed, mean speed, mean moving speed), 5 average proximities from infrastructures (railway, tram lines, roads, bus lines, metro lines) and the location of the trajectory in respect to water surfaces. • Uses OpenStreetMap for geographical data and transportation infrastructure. • Data are organised into geometry (polylines) and attributes (tags).
A tale of many cities: universal patterns in	2012	<ul style="list-style-type: none"> • Probabilistic Modelling 	<ul style="list-style-type: none"> • Probability Density Functions (for displacement 	35,289,629 Foursquare <i>check-</i>	Modelling urban	<ul style="list-style-type: none"> • The authors consider consecutive check-ins within same city as one movement.

human urban mobility [103]			probability distribution) <ul style="list-style-type: none"> • Kolmogorov-Smirnov test for distribution synthesis • Maximum Likelihood Estimation for parameter fine-tuning 	<i>ins</i> from 925,030 unique users over 4,960,496 venues	mobility	<ul style="list-style-type: none"> • Distance is not the deciding factor in human mobility, but place density is important. • Each place is described through a rank metric, which presents the #places offered between origin and destination. • The rank-based model is used to calculate human mobility in the city. • Distribution of human displacements approximated with power law.
Discovering regions of different functions in a city using human mobility and POIs [93]	2012	<ul style="list-style-type: none"> • Clustering • Topic Modelling • Regression • Semantic Annotation 	<ul style="list-style-type: none"> • K-means clustering for region functionality aggregation • TF-IDF on POIs data • Latent Dirichlet Allocation (LDA) on mobility data • Dirichlet Multinomial Regression (DMR) • Kernel Density Estimation model for functionality intensity (number of visits) 	Two Beijing POI datasets for years 2010 and 2011 Two GPS trajectory datasets from 12,000 taxicabs Beijing road networks	Mobility pattern detection in big city regions	<ul style="list-style-type: none"> • Uses Regions of different Functions (DRoF) and points of interest. • Employs topic-based inference model which regards region as document, function as topic, categories of POIs as metadata and human mobility patterns as words. • Uses raster-based model to represent road network. • Compares DRoF against TF-IDF-based and LDA-based methods.
Friendship and mobility: user movement in location-based social networks [91]	2011	<ul style="list-style-type: none"> • Probabilistic Modelling 	<ul style="list-style-type: none"> • Probability distributions • Expectation-Minimisation (EM) 	<ol style="list-style-type: none"> 1) 6.4 mil. check-in data from Gowalla social platform 2) 4.5 mil. from Brightkite social platform 3) 450 mil. phone calls dataset 	Modelling human mobility patterns and temporal dynamics	<ul style="list-style-type: none"> • Modelling human mobility that combines short range movement with travel. • Multiple sources, from social networks to phone call logs, are analysed. • Network connections (friendships) are also considered. • Gives an order of magnitude better performance than other stochastic models. • Social networks influence long distance travel more than short distance. • Periodicity also investigated.
Social event detection in	2011	<ul style="list-style-type: none"> • Probabilistic 	<ul style="list-style-type: none"> • Voronoi partition (for 	~ 900 million calls	Social events	<ul style="list-style-type: none"> • Aims to detect unusual massive gatherings

massive mobile phone data using probabilistic location inference [101]		Modelling	location inference) • Bayesian location inference model	and text messages (anonymised caller and callee, location estimation from towers)	detection and attendance prediction	(concerts, sports finals, emergencies, protests etc.) from mobile network data. • Focus on non-routine behaviour. • Non-routine behaviour of people correlates with behaviour of other people too. • The framework detected successfully the events that took place in the examined areas.
Human mobility prediction based on individual and collective geographical preference [105]	2010	• Probabilistic Modelling	• Custom model	1) Mobile phones locations: traces from 2000 users 2) Land use data from MassGIS 3) POI data from Yelp	Predict location of a person over time	• Method combines data from multiple sources, using open data and mobile phones data • Model is based on the geographical features of the area where the person moves, in terms of land use, points of interests and collectivity's habits • Can be used as recommender
A Bayesian approach to modeling lost person behaviors based on terrain features in wilderness search and rescue [100]	2010	• Probabilistic Modelling	• Bayesian model for probability distribution map • Order-1 Markov mode	Synthetic GPS data	Predict route of missing persons	• Behaviour modelling is based on three terrain features: topography, vegetation and local slope. • Past operations data, expert opinions and past statistical data can be incorporated into the model. • A temporal state transition matrix allows the generation of predictions for any given time interval.
Mining interesting locations and travel sequences from GPS trajectories [99]	2009	• Clustering • Ranking (of location interest) • Graph theory	• OPTICS: density-based clustering algorithm • nDCG, MAP	GPS trajectories of 107 users over 1 year	Discover interesting locations and travel sequences	• HITS-based inference model • Works best as a recommender for tourist attractions and best sequence to visit them • Clear advantages over <i>rank-by-count</i> and <i>rank-by-frequency</i> methods
Using location for personalized POI recommendations in	2006	• Recommender	• Collaborative filtering (based on location)	12,000 restaurant POIs	Recommend POIs	• Recommendations based on user ratings for restaurant POIs that are in the same vicinity • Ways to deal with the "Cold start" problem

mobile environments [98]						(when there are no votes yet)
Finding REMO— detecting relative motion patterns in geospatial lifelines [106]	2005	<ul style="list-style-type: none"> • Clustering 	<ul style="list-style-type: none"> • Voronoi partition • Geometric algorithms for track & encounter patterns 	Lifeline data (id, location, time) and Motion attributes (speed, change of speed, azimuth)	Cluster detection of motion patterns in space-time	<ul style="list-style-type: none"> • The method relies solely on point observations in a Euclidean space • Introduces a wide variety of motion patterns • REMO patterns are spatiotemporal • REMO patterns can be used for any object that can be depicted as point and leaves a track behind
Using GPS to learn significant locations and predict movement across multiple users [94]	2003	<ul style="list-style-type: none"> • Clustering • Predictive Modelling • Probabilistic Theory 	<ul style="list-style-type: none"> • K-means for location and sublocation definition from individual places • Markov model 	GPS logs from one user for a 4-month period	Meaningful locations modelling and extraction	<ul style="list-style-type: none"> • Single-user and multi-user applications • Significant places are traced from GPS time-gaps (due to lack of signal from inside buildings)

II.4 Social Media Analytics

A comparison list of the studied literature is presented in the following table. Each row describes one scientific paper and the related methodology using the following fields:

Reference

Title [Reference]: Article title and reference to related citation.

Year: The year the article was published.

Category: Depending on the focus of the article, this field contains the category of the algorithms presented in the article. The categorisation is made in respect to the tasks involved, such as classification, regression, clustering, statistical analysis, etc.

Algorithm / Method: The algorithm/method suggested by the article, along with proposed adaptations for enhancement of results or mitigation of computation complexity.

Experimentation dataset: The source and form of the data used to test and evaluate the proposed methodology, if any.

Primary focus: The general field of the paper and any special focus.

Results: Brief summarisation of results, conclusions and key points.

Table Annex II- 4 Literature review in Social Media Analytics

Title [reference]	Year	Category	Algorithm/ Method	Experimentation dataset	Primary focus	Results
Discovering shifts to suicidal ideation from mental health content in social media [123]	2016	<ul style="list-style-type: none"> Linguistic Analysis Classification 	<ul style="list-style-type: none"> Variable extraction Logistic regression 	Discussion data from 880 users in Reddit (approx. 13,000 posts & 100,000 comments)	Intention discovery	<ul style="list-style-type: none"> Mostly statistical approach The approach considers 5 models: linguistic structure, inter-polar awareness, interaction, content, & full. Method can be easily adapted to other social media and for other vulnerable groups of people
Fusing social media cues: personality prediction from twitter and Instagram [120]	2016	<ul style="list-style-type: none"> Linguistic Analysis Feature extraction Regression 	<ul style="list-style-type: none"> Linguistic Inquiry and Word Count (LIWC), ANEW lexicon F-statistic subsampling for feature extraction Random forest Big Five model 	Instagram and Twitter (images, texts -tweets and image captions-, #followers & #followees)	Personality prediction based on various social media data	<ul style="list-style-type: none"> Best results when using input from both social networks Method can be easily adapted to other social media The overall best regressor was the one using the complete feature set (linguistic, image, meta)
Computational personality recognition in social media [121]	2016	<ul style="list-style-type: none"> Linguistic Analysis Regression 	<ul style="list-style-type: none"> Linguistic Inquiry and Word Count (LIWC) Univariate regression (SVM with radial kernel and decision tree), 	<ul style="list-style-type: none"> Facebook, Twitter, (demographics, texts, activities) YouTube vlogs 	Personality prediction based on various social media platforms and data	<ul style="list-style-type: none"> Age and gender showed high correlation with all personality traits Decision tree models generally outperformed

			<ul style="list-style-type: none"> • Multivariate regression algorithms (SVM, decision tree) • Big Five model 	(audio-video features)		<p>SVM</p> <ul style="list-style-type: none"> • Overall best performance: MTSC and ERCC with decision tree base learner
A multi-label, semi-supervised classification approach applied to personality prediction in social media [118]	2014	<ul style="list-style-type: none"> • Semi-supervised Classification 	<ul style="list-style-type: none"> • Naïve Bayes, Support Vector Machine, MLP • Big Five Model 	Tweets with grammatical meta-attributes	Personality prediction based only on tweets meta-attributes	<ul style="list-style-type: none"> • Extroversion, agreeableness and neuroticism are accurately predicted • Openness and conscientiousness were more difficult to predict with the suggested meta-attributes • Less dependent of language in comparison to grammar-based approaches
Predicting depression via social media [122]	2013	<ul style="list-style-type: none"> • Linguistic Analysis • Feature extraction • Regression 	<ul style="list-style-type: none"> • Statistical approach • PCA • SVM (with RBF kernel) 	Tweets and attributes (188 features vector)	Depression prediction based on various social media data	<ul style="list-style-type: none"> • Depression prediction ahead of onset yields accuracy ~70% • Model that uses only linguistic features performs better
Personality, gender, and age in the language of social media: The open-vocabulary approach [117]	2013	<ul style="list-style-type: none"> • Linguistic Analysis • Feature extraction • Correlation analysis 	<ul style="list-style-type: none"> • Linguistic Inquiry and Word Count (LIWC) • Latent Dirichlet Allocation (LDA) • Ordinary Least Squares regression 	Analysed 700 mil. words of 75,000 Facebook users	Open vocabulary method for personality prediction	<ul style="list-style-type: none"> • Open-vocabulary technique • Largest study to date • Predictions on gender, age and personality • A word cloud visualisation is presented
Predicting personality with social media [119]	2011	<ul style="list-style-type: none"> • Linguistic Analysis • Feature weighting • Regression 	<ul style="list-style-type: none"> • Linguistic Inquiry and Word Count (LIWC) • Multiple linear 	Facebook profiles (personal info, preferences and	Personality prediction based on various social media	<ul style="list-style-type: none"> • Achieved personality prediction within 11% of actual values

			<p>regression to predict feature weights for each personality trait (M5' and Gaussian processes)</p> <ul style="list-style-type: none"> • Big Five Model 	activities, posts, network characteristics)	data	<ul style="list-style-type: none"> • Short texts, like fb posts, can be insufficient for linguistic analysis, so individual texts were unified before analysis
Collaborative Location Recommendation by Integrating Multi-dimensional Contextual Information [130]	2018	<ul style="list-style-type: none"> • Recommender 	<ul style="list-style-type: none"> • Model-based Collaborative Filtering (Tensor Factorisation) • Non-negative Matrix Factorisation • Stochastic gradient descent (SGD) 	<ol style="list-style-type: none"> 1)Brightkite user check-ins 2) Gowalla user check-ins 	Location (POI) recommendation	<ul style="list-style-type: none"> • Incorporates friends' influence • Outperforms other CF models
Personalized location recommendation by aggregating multiple recommenders in diversity [129]	2017	<ul style="list-style-type: none"> • Recommender Ensemble 	<ul style="list-style-type: none"> • Various Collaborative Filtering (CF) methods 	<ol style="list-style-type: none"> 1) Foursquare user check-ins 2) Gowalla user check-ins 	Location recommendation with recommender ensemble	<ul style="list-style-type: none"> • Novel framework (LURWA) of recommender ensemble using various weighting strategies • Linear aggregation • Outperforms typical CF models
Content-Aware Point of Interest Recommendation on Location-Based Social Networks [125]	2015	<ul style="list-style-type: none"> • Recommender • Sentiment Analysis 	<ul style="list-style-type: none"> • Custom CF method (CAPRF) 	Foursquare user check-ins, user tweets and POIs properties	Location recommendations based on social networks	<ul style="list-style-type: none"> • Combines POI properties, user interests, and sentiment indications to recommend a location • Sparse dataset that produces very low precision results
A sentiment-enhanced personalized location recommendation system	2013	<ul style="list-style-type: none"> • Recommender • Sentiment Analysis 	<ul style="list-style-type: none"> • Collaborative Filtering (CF) • Location-based Social Matrix Factorisation 	1)Foursquare user check-ins and network connections and	Personalised location recommendation using user check-ins and contextual	<ul style="list-style-type: none"> • Social influence and venue similarity are considered. • This hybrid model

[128]			<ul style="list-style-type: none"> • Dictionary-based unsupervised sentiment analysis 	2) Foursquare venue categories & tips	information	<p>outperforms other models that only use check-ins or consider tips in a random way</p> <ul style="list-style-type: none"> • Models that incorporate social influence impact perform better
Location-based and preference-aware recommendation using sparse geo-social networking data [127]	2016	<ul style="list-style-type: none"> • Recommender 	<ul style="list-style-type: none"> • Collaborative Filtering (CF) 	Number of visits to venue	Location recommendation based on user preferences and social opinions	<ul style="list-style-type: none"> • Opinions from local experts with whom the user shares interests are included in the calculation • User preferences extracted from location history <p>This approach outperformed major recommendation methods MPC, LCF, PCF</p>
A recommendation system for spots in location-based online social networks [124]	2011	<ul style="list-style-type: none"> • Recommender 	<ul style="list-style-type: none"> • Memory-based Collaborative Filtering (CF) • Regularised Matrix Factorisation • Specific region restriction 	Gowalla check-ins	Personalised location recommendation	<ul style="list-style-type: none"> • Provides a solution to the lack of straight forward ratings • The error metrics were below the baseline • Geographic and temporal aspects are considered in the recommendation
Location recommendation for location-based social networks [126]	2010	<ul style="list-style-type: none"> • Recommender 	<ul style="list-style-type: none"> • Collective Matrix Factorisation • Friend-based CF (FCF) • Geo-Measured FCF 	Foursquare user profiles (User-location pairs & network)	Location recommendation based on social and geographical characteristics	<ul style="list-style-type: none"> • Considering friend similarity limits the user space and calculations • Distance between friends is also a factor in the geo-measured FCF

						<p>method</p> <ul style="list-style-type: none"> Both proposed techniques offer lower computational complexity
A pattern-based approach for multi-class sentiment analysis in twitter [135]	2017	<ul style="list-style-type: none"> Linguistic Analysis ML Classification 	<ul style="list-style-type: none"> Natural Language Processing (NLP) techniques Random Forest 	40,000 Tweets (manually labelled)	Multi-class sentiment analysis	<ul style="list-style-type: none"> Classify into 7 emotion-based classes instead of usual 3 Performance ~60% Software (java) called SENTA was created after this approach
Unsupervised Sentiment Analysis for Social Media Images [139]	2015	<ul style="list-style-type: none"> Unsupervised Classification 	<ul style="list-style-type: none"> Non-negative matrix factorisation SentiBank and EL frameworks with K-means instead of SVM/logistic regression 	Flickr, Instagram images with text captions	Image sentiment analysis, enhanced by contextual information	<ul style="list-style-type: none"> Proposes an Unsupervised Sentiment Analysis framework (USEA) Incorporates visual and textual information, in order to provide more accurate predictions on image sentiment content Too much textual information dominates the learning process and results in overfitting
Robust Image Sentiment Analysis Using Progressively Trained and Domain Transferred Deep Networks [138]	2015	<ul style="list-style-type: none"> Deep learning 	<ul style="list-style-type: none"> Deep Convolution Neural Networks Progressive CNN for noise reduction Transfer learning 	Flickr images, sentiment labelled Twitter images	Large scale image sentiment analysis	<ul style="list-style-type: none"> Both progressive training and transfer learning based on a small, confidently labelled subset, increased performance Leveraged large training and testing datasets for

						developing a more robust model
Deep convolutional neural networks for sentiment analysis of short texts [134]	2014	<ul style="list-style-type: none"> • Deep learning 	<ul style="list-style-type: none"> • Word-embeddings (e.g. word2vec) • Deep Convolution Neural Networks 	Movie reviews and twitter messages (SSTb and STS corpora)	Sentiment analysis with deep convolutional neural networks	<ul style="list-style-type: none"> • Performs analysis using character-level, word-level and sentence-level representations
Sentiment analysis in Facebook and its application to e-learning [137]	2014	<ul style="list-style-type: none"> • Linguistic Analysis and Lexicon-based classification • Feature selection • ML Classification 	<ul style="list-style-type: none"> • Natural Language Processing (NLP) techniques • Correlation-based feature selection • Decision trees (C4.5), Naïve Bayes, SVM 	Facebook posts	Text sentiment classification with both lexicon-based and ML techniques	<ul style="list-style-type: none"> • The approach detects emotional changes by comparing the “current” sentiment of a user with the “usual” one. • ML algorithms combined with lexicon-based techniques, achieve higher performance • Decision trees had slightly better performance among ML algorithms
Twitter, MySpace, Digg: Unsupervised sentiment analysis in social media [133]	2012	<ul style="list-style-type: none"> • Linguistic Analysis and Lexicon-based classification • ML Classification 	<ul style="list-style-type: none"> • Natural Language Processing (NLP) techniques • Naïve Bayes, Maximum Entropy, SVM 	Twitter, Digg, Myspace comments	Subjectivity and polarity classification in less domain-specific, informal texts	<ul style="list-style-type: none"> • In polarity classification, the proposed lexicon-based classifier outperformed ML techniques • In subjectivity classification the proposed lexicon-based classifier outperformed ML techniques, except for the twitter dataset • Emotional word detection and neighbourhood scanning for

Twitter as a corpus for sentiment analysis and opinion mining [136]	2010	<ul style="list-style-type: none">• ML Classification	<ul style="list-style-type: none">• Multinomial Naive Bayes• Also experimented with SVM and CRF	Tweets for corpus collection POS-tags	Sentiment classification of microblogging posts	negators/intensifiers <ul style="list-style-type: none">• The proposed method for negative/positive/neutral sentiment corpus collection from Twitter posts can be fully automated and has no volume limitation• Bigrams performed better than uni- and trigrams
---	------	---	--	--	---	---

II.5 Privacy and Anonymisation

A comparison list of the studied literature is presented in the following table. Each row describes one scientific paper and the related methodology using the following fields:

Title [Reference]: Article title and reference to related citation.

Year: The year the article was published.

Category: Depending on the focus of the article, this field contains the category of the approach and/or privacy model presented in the article.

Algorithm / Method: The algorithm/method suggested by the article.

Experimentation dataset: The source and form of the data used to test and evaluate the proposed methodology, if any.

Primary focus: The general research field of the paper and any special focus.

Results: Brief summarisation of results, conclusions and key points.

Table Annex II- 5 Literature review in Privacy and Anonymisation

Title [reference]	Year	Category	Algorithm/ Method	Primary focus	Results
Dummy Generation Based on User-Movement Estimation for Location Privacy Protection [158]	2018	<ul style="list-style-type: none"> Location-based services Location privacy, Data privacy 	<ul style="list-style-type: none"> Framework of Edge Intersection with user Intersection between Dummies Anonymous Area Enlargement 	Dummy-based approaches that generate dummies and their locations are sent along with the actual location of a user to an LBS provider.	<ul style="list-style-type: none"> proposed Edge that can lower the traceability and keep the user-required anonymous area size under the practical assumption. Edge utilises the traveling salesman problem to estimate the user-movement and designs trajectories of dummies so that they intersect with the user effectively and enlarge the anonymous area size appropriately.

Title [reference]	Year	Category	Algorithm/ Method	Primary focus	Results
					<ul style="list-style-type: none"> simulation experiments using real map information were conducted. The experimental results show that Edge can lower the traceability more compared with the existing methods.
A Truthful Data Anonymisation Algorithm With Strong Privacy Guarantees [164]	2018	<ul style="list-style-type: none"> Data Privacy Differential Privacy Anonymisation Disclosure control 	<ul style="list-style-type: none"> Data anonymisation algorithm 	Satisfy the differential privacy model. The algorithm does not perturb input data or generate synthetic output data	<ul style="list-style-type: none"> Flexible differentially private data release mechanism that produces truthful output data.
Privacy in Location Based Services: Protection Strategies, Attack Models and Open Challenges [143]	2017	<ul style="list-style-type: none"> Location-based services Location privacy Attack models Privacy protection strategies K-anonymity 	<ul style="list-style-type: none"> Cloaking algorithm to generate cloaked regions Randomisation based reconstruction algorithm for releasing anonymised trajectory data 	A state-of-art survey of privacy in location-based services containing details of all privacy protection schemes is presented. Some open challenges in the area of location privacy are also demonstrated.	<ul style="list-style-type: none"> Demonstrate various achievements and research works accomplished in the area of location-based privacy. All the mentioned attacks and measures to prevent have been integrated and also suggested future research directions.
A Game Theoretic Framework for Analyzing Re-Identification Risk [166]	2015	<ul style="list-style-type: none"> De-identification Model Identity disclosure risk 	<ul style="list-style-type: none"> De-identification algorithms 	Review de-identification and anonymisation models.	<ul style="list-style-type: none"> Contextualise research on de-identification and anonymisation models with respect to other investigations into game theory for characterising and addressing

Title [reference]	Year	Category	Algorithm/ Method	Primary focus	Results
					privacy and security concerns.
Protecting location privacy for outsourced spatial data in cloud storage [159]	2014	<ul style="list-style-type: none"> Location Privacy Spatial Data Data Transformation 	<ul style="list-style-type: none"> Index modification algorithm to improve the security of SHC, denoted as SHC Index generation algorithm for DSC 	Improve the security of standard Hilbert curve (SHC)	<ul style="list-style-type: none"> The estimation distortion shows that SHC* and DSC are more secure than SHC. The index generation time shows that DSC is more efficient than SHC and SHC*, so DSC achieves good security and efficiency performance.
Estimating the re-identification risk of clinical data sets [162]	2012	<ul style="list-style-type: none"> Location Privacy Re-identification 	<ul style="list-style-type: none"> A Monte Carlo simulation was performed to evaluate the uniqueness estimators on six clinically relevant data sets. 	Evaluate the accuracy of uniqueness estimators on clinically relevant data sets.	<ul style="list-style-type: none"> This decision rule had the best consistent median relative error across multiple conditions and data sets.
k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY [148]	2012	<ul style="list-style-type: none"> data anonymity data privacy re-identification 	<ul style="list-style-type: none"> k-anonymity protection model 	A formal protection model named k-anonymity and a set of accompanying policies for deployment	<ul style="list-style-type: none"> k-anonymity protection model, explored related attacks provided ways in which these attacks can be thwarted.
An Optimization of Association Rule Mining using K-Map and Genetic Algorithm for Large Database [149]	2012	<ul style="list-style-type: none"> Association rule mining k-map Data privacy 	<ul style="list-style-type: none"> A priori algorithm Tree based algorithm Genetic Algorithm 	A novel method for optimisation of association rule mining.	<ul style="list-style-type: none"> The proposed algorithm is a combination of k-map and genetic algorithm. The large generated rule is optimised with genetic algorithm. Local pruning achieves a reduction in the number of searched candidates and this reduction has a

Title [reference]	Year	Category	Algorithm/ Method	Primary focus	Results
Evaluating re-identification risks with respect to the HIPAA Privacy Rule [146]	2010	<ul style="list-style-type: none"> • Encryption • Re-identification 	risk metrics: <ul style="list-style-type: none"> • Encryption expected number of re-identifications; estimated proportion of a population in a group of size g or less, and monetary cost per re-identification. 	Estimate re-identification risk for data sharing policies of the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. Evaluate the risk of a specific re-identification attack using voter registration lists.	proportional impact on the reduction of exchanged messages. <ul style="list-style-type: none"> • Illustrates that blanket protection policies, such as Safe Harbor, leave different organisations vulnerable to re-identification at different rates. • It provides justification for locally performed re-identification risk estimates prior to sharing data.
A reciprocal framework for spatial K-anonymity [157]	2010	<ul style="list-style-type: none"> • Spatial Data • Location-based Services • Anonymisation 	<ul style="list-style-type: none"> • A general framework for obtaining reciprocal (i.e., secure) cloaking algorithms 	propose a general framework for implementing reciprocal algorithms using any existing spatial index on the user locations	<ul style="list-style-type: none"> • A reciprocal framework that allows the implementation of a variety of secure algorithms for spatial K-anonymity on top of a spatial index. • Extend the framework to support users with variable query. • Demonstrate the versatility of the proposed framework by using it to implement a variety of partitioning techniques on top of two popular spatial indices.

Title [reference]	Year	Category	Algorithm/ Method	Primary focus	Results
Anonymity and Historical-Anonymity in Location-Based Services [155]	2009	<ul style="list-style-type: none"> • Location-based services • Location privacy • Data privacy 	<ul style="list-style-type: none"> • Greedy algorithm for historical k-anonymity 	Investigate the issues involved in the experimental evaluation of anonymity based defence techniques	<ul style="list-style-type: none"> • Performance evaluation of techniques depending on the adversary model and on the specific service deployment model.
p-Sensitivity: A Semantic Privacy-Protection Model for Location-based Services [154]	2008	<ul style="list-style-type: none"> • Location-based services • Location privacy • Data privacy • Anonymisation 	<ul style="list-style-type: none"> • PE-Tree to model all possible anonymisations of user queries, based on which search algorithms and heuristics are developed to find the optimal anonymisation. 	A novel privacy protection model that considers query diversity and semantic information in anonymising user locations	<ul style="list-style-type: none"> • A PE-Tree based method to implement the p-sensitivity model. • Search algorithms and heuristics are developed to efficiently find the optimal p-sensitivity anonymisation in the tree. • Preliminary experiments are conducted to demonstrate that p-sensitivity provides high-quality services without compromising users' query privacy. • The algorithm achieves a much lower anonymisation cost than the existing algorithm.
The cost of privacy: Destruction of data-mining utility in anonymized data publishing [163]	2008	<ul style="list-style-type: none"> • Data privacy • Anonymisation 	<ul style="list-style-type: none"> • machine learning algorithms on both trivially sanitised versions of the database. 	Whether generalisation and suppression of quasi-identifiers offer any benefits over trivial sanitisation which simply separates quasi-identifiers from sensitive attributes.	<ul style="list-style-type: none"> • The privacy vs. utility trade-off for the respective algorithms is very poor.

Title [reference]	Year	Category	Algorithm/ Method	Primary focus	Results
L-diversity: Privacy Beyond K-anonymity [150]	2007	<ul style="list-style-type: none"> • Data privacy • Anonymisation 	<ul style="list-style-type: none"> • Entropy ℓ-diversity • npd recursive (c_1, c_2, ℓ)-diversity 	An experimental evaluation that ℓ -diversity is practical and can be implemented efficiently	<ul style="list-style-type: none"> • ℓ-diversity, a framework that gives stronger privacy guarantees
Hiding the presence of individuals from shared databases [153]	2007	<ul style="list-style-type: none"> • K-anonymity • Data privacy • Delta presence • medical databases 	<ul style="list-style-type: none"> • Single-Dimensional Presence Algorithm: SPALM • Multi-Dimensional Presence Algorithm: MPALM 	A metric, δ -presence, that clearly links the quality of anonymisation to the risk posed by inadequate anonymisation.	<ul style="list-style-type: none"> • Design δ-presence algorithms that guarantee bounds on optimality
(α , k)-anonymity Based Privacy Preservation by Lossy Join [151]	2007	<ul style="list-style-type: none"> • K-anonymity • Data privacy • Anonymisation 	<ul style="list-style-type: none"> • (α, k)-anonymity which generalises the QID and forms one generalised table only. • Anatomy algorithm which makes use of the lossy join for the anonymisation 	Publish the data in such a way that the privacy protection for (α , k)-anonymity can be achieved with less distortion	<ul style="list-style-type: none"> • Conducting some experiments and verified the improvement on information loss.
t-Closeness: Privacy Beyond k-Anonymity and l-Diversity [152]	2007	<ul style="list-style-type: none"> • K-anonymity • Data privacy • Anonymisation 	<ul style="list-style-type: none"> • The t-closeness Principle • Analysis of t-Closeness with Earth Mover's distance (EMD) 	A novel privacy notion called t-closeness	<ul style="list-style-type: none"> • Separate the information gain an observer can get from a released data table into two parts: the one about all population in the released data and the other about specific individuals • Use of the Earth Mover Distance measure for specific t-closeness requirement; this has the

Title [reference]	Year	Category	Algorithm/ Method	Primary focus	Results
					advantage of taking into consideration the semantic closeness of attribute values
Research on Data Encryption Technology Based on Chaos Theory [167]	2007	<ul style="list-style-type: none"> • Data privacy • Data encryption • Logistic model 	<ul style="list-style-type: none"> • Chaos dynamic logistic model 	A chaos dynamic logistic model as the data encryption algorithm	<ul style="list-style-type: none"> • Randomness, correlation, complexity of Logistic chaotic series were verified through power spectrum, Lyapunov exponent mathematical analysis and it meets the needs of cryptology
A Formal Model of Obfuscation and Negotiation for Location Privacy [144]	2005	<ul style="list-style-type: none"> • Location privacy • Data privacy • Anonymisation • Pseudo-anonymisation 	<ul style="list-style-type: none"> • Negotiation proximity query with obfuscation • Computation of the relation δ 	A formal framework within which obfuscated location-based services are defined	<ul style="list-style-type: none"> • Provide a mechanism to balance an individual's location privacy with that individual's need for a high-quality location-based service.
Data Cube: A Relational Aggregation Operator Generalizing Group-By, Cross-Tab, and Sub-Totals [156]	1997	<ul style="list-style-type: none"> • Data Analysis • Data aggregation 	<ul style="list-style-type: none"> • 2N-algorithm • order-N algorithm 	Explains the cube and roll-up operators, shows how they fit in SQL, explains how users can define new aggregate functions for cubes, and discusses efficient techniques to compute the cube	<ul style="list-style-type: none"> • The cube operator generalises and unifies several common and popular concepts: <ul style="list-style-type: none"> ○ aggregates ○ group by ○ histograms ○ roll-ups and drill-downs ○ cross tabs.

Annex III: Interviews

III.1 Insights from the Hellenic Amber Alert

III.1.1 Interviews with Hotline operator

“The Smile of the Child” operates the European Hotline for Missing Children 116000 since 2008. It operates the line 24 hours a day, 7 days a week and is toll free from both land lines and mobiles. It is staffed by social workers and psychologists (hotline operators/case managers). Its purpose is the prevention of the phenomenon of missing children, providing support to children who have gone missing and their families.

Its objectives include:

- co-operation with competent authorities
- providing support to searches for missing children
- providing counselling to the child and its family once the child returns safely at home

The Hotline offers its services to:

- Parents, children and citizens
- Police authorities for providing support in the search for missing children.

The Hotline deals with all categories of missing children:

- Runaways (National / International)
- Parental abductions
- Criminal abductions
- Missing unaccompanied migrant minors
- Lost, injured or otherwise missing children

The 116000 Hotline is particularly useful to children, parents and teachers who travel, since call receivers can refer callers to the competent authorities in those countries. 116000 Hotline provides a safety net for parents and children across Europe on holidays or for professional and other reasons.

The 116000 Hotline is interconnected with the European Emergency Number 112 of the Civil Protection Secretariat, is recognised as an Emergency Line, and is incorporated under Missing Children Europe (MCE).

The fact that the European Hotline for Missing Children 116000 operates throughout Europe makes it a valuable tool also in the process of providing support to children arriving in Greece as refugees and migrants. This need led the Organisation to establish a group of employees and volunteer interpreters / -mediators with knowledge of Arabic and Farsi. In addition, the line 116000 and related services provided were communicated through special brochures, at the points of entry and residence of children refugees and migrants in Greece.

On the institutional level the Hotline cooperates with:

- Ministry of Public Order & Citizens’ Protection
- Ministry of Shipping and the Aegean
- Prosecution
- Children’s Hospitals
- NGOs

- Volunteer Partners, including Hellenic Rescue Team, RSF Hellas Group Communication and Volunteer Rescue, Civil Olympic Village Protection, Hellenic Red Cross-Volunteer Samaritans, Rescuers and lifeguards Corps.

The particular services that the Hotline offers include:

- Assistance and consultation with law enforcement
- Activation of AMBER Alert Hellas
- Activation of the Search and Rescue Team "Thanasis Makris"
- Psychological Support to family
- Communication with mass media through radio, television and print
- Publication of photograph and information on our site (www.hamogelo.gr), on our social media and on the GMCN site (www.missingkids.com)
- Creation and distribution of posters in all publicly accessible areas
- Creation of photographic leaflets distributed through the Ministry of Citizen's Protection to all patrol cars in Greece
- Coordination of public and private agencies and businesses
- Psychological support and counselling of children once they have been found.
- Assistance and consultation in cross border cases

The following tools are at the disposal of the 116000 Hotline:

- AMBER ALERT Hellas
- IP Call Centre of Cisco
- Contact Centre of Cisco that cooperates with the Siebel CRM application of Oracle
- ECAAS (European Child Alert Automated System)
- Search and Rescue Team "Thanasis Makris" with the active role of internationally recognised dog teams²⁵
- Wide range of Logistics and Operational tools - Vehicles of Direct Intervention (jeeps, ATV, motorbikes, ambulances)
- The Mobile operational Unit of the Organisation
- "Odysseas", a Mobile Crisis Management Unit

The Hotline of the Smile of the Child handles approximately 150 cases per year. According to the adopted cases classification, the following categories emerge: runaways, alarming disappearance, disappearance of unaccompanied minors and parental abduction.

Some statistics about the activities of the Hotline:

- In 2015, 8.852 calls were made to the 116000 Hotline, and 141 cases of missing children were referred (69% runaways of teenagers, 13,5% alarming disappearances, 9% missing unaccompanied migrant minors, 8,5% parental abductions). Out of 126 cases that we finally handled, 119 minors were found (94,5%). Amber Alert Hellas was activated for 7 cases, while the "Thanasis Makris" Search and Rescue Team was activated in 5 cases.

²⁵ The Search and Rescue Team "Thanassis Makris" is staffed with volunteers specialized in search and rescue, while participatory dog teams are also staffed with volunteers certified in dog rescues, certified by the National Search and Rescue Dog Association (NSARDA) of the UK.

- In 2016, 10.065 calls were made to the 116000 Hotline, and 170 cases of missing children were referred (49,5% runaways of teenagers, 23,5% alarming disappearances, 14% missing unaccompanied migrant minors, 13% parental abductions). Out of 170 cases that we finally handled, 136 minors were found (80%). Amber Alert Hellas was activated for 9 cases, while the "Thanasis Makris" Search and Rescue Team was activated in 12 cases. Out of minors who are still missing, in 10 cases (6%) relevant communication was interrupted and in 24 (14%) there is an ongoing search (17 cases relate to unaccompanied minors and 1 was found at the start of the New Year).
- In 2017, 9.379 calls were made to the 116000 Hotline, and 128 cases of missing children were referred (54,5% runaways of teenagers, 29,5% alarming disappearances, 4% Missing unaccompanied migrant minors, 12% parental abductions). Out of 128 cases that we finally handled, 105 minors were found (90%). Amber Alert Hellas was activated for 10 cases, while the "Thanasis Makris" Search and Rescue Team was activated in 8 cases. Out of 13 minors who are still missing, in 10 cases (8%) relevant communication was interrupted and in 3 (2%) there is an ongoing search (cases relate to unaccompanied minors).

When asking about the existing practice when searching for missing children in Greece, although it can be considered to be satisfactory and effective enough, there is still room for improvement. More specifically, there is lack of information of the general public about the importance of informing timely the Hotline. In addition, there is a need for specific training of the personnel, which is involved in the search of a missing child, while it is necessary to raise the awareness of specific target groups such as teachers, parents, and students. Once the child is discovered, the procedure is as follows: it is necessary to have an official identification of the child by the police department. At this point, a clarification is required of the reason the child disappeared (was it a child abuse? etc.). Depending on the case, it may be compulsory for the prosecutor to be informed and follow the procedures foreseen. On behalf of the organisation, further support with high expertise is offered (counselling services etc.) It should be stressed that the main role of the hotline is to collect and convey information about the missing child to the authorities. The hotline does not substitute the role of police; it is supporting the authorities. There is collaboration with the authorities via MoUs with clear roles.

Some examples are presented below in order to understand the main elements of action and reaction when searching for a child.

Example of case 1:

On the 30th of November 2016 we received a call from a police officer informing us about the disappearance of a 17-year-old boy (unaccompanied minor) from a specific camp. On 24th of November 2016 he went to the Embassy for interview about his upcoming relocation. The interview finished at 16.30 and since then his whereabouts were unknown. He does not know the area; a friend of him escorted him to the embassy and then left. The missing child's report to the police was made by a psychologist of an organisation responsible for the child. The child does not have relatives or other friends (apart from this one friend). The case manager started collecting information about the case and got in contact with hospitals (in case the child was transferred to a hospital), police and prosecutor. The case manager collected photos of the

child and takes the approval from police and prosecutor for publicity. The information about the disappearance was disseminated via the Smile of the Child website, social media and mass media. We received a message by a citizen through Facebook informing us that s/he saw the child at a hospital. The hospital was contacted and the information verified. The police was informed in order to proceed with the identification. The public information on the child was removed and the public informed about the closure of the case.

Example of case 2:

Smile of the Child was informed by a social worker of an organisation working with refugees about the disappearance of a 14 years old Afghan girl. She arrived in Greece with her mother on the 28th of February 2016 and they live at an apartment in the centre of Athens. The girl attends school (either via school bus or via metro). The previous day she started to go to school but she never arrived at school. She did not arrive at home at night and she does not have her asylum card with her. She is missing, and she has not gotten in contact with anyone. The mother has reported the case to the police and has provided the girls' photo. The girl has a mobile phone on which she uses only Viber. The girl is fluent in English and knows how to move in the Athens area via public transportation. Her mother speaks only Farsi. This is the first time that the girl has not returned home. Her classmates do not have any further information. The girl has a Facebook account. The case manager collects the information from the social worker and gets in contact with the police. The case manager continues to collect information from the mother too (the characteristics of the girl, family status, any other friends or relatives, pocket money, etc.). For any new information that the mother has not testified to the police, the case manager urges the mother to report it to the police and conveys the collected info to the police. The social worker informed us that the girl got in contact with a friend of hers via Facebook asking her if her mother is ok but then she blocked her and she cannot contact her again. The mother has provided her consent for the publishing of the photo and characteristics of the missing girl. The social worker informed us that the mother learned from a friend of the missing girl that she had announced her running away and that she would commit suicide because she cannot live without her father. This was reported by the missing girl 3 days before the disappearance. In addition, the mother found that her pills are missing. She was guided to report all the info to the police and we got in contact again with the police. With the informed consent of the mother and upon approval of the police, we proceeded with the public appeal of the missing girl. After the public appeal, we received a call from a social worker working at a camp in Northern Greece that she has seen this girl there. We talked with the staff of the camp, who also verified the information and we contacted the police to visit the camp for the official identification. The girl reported that she does not want to return, that she had committed 3 suicide attempts and that she had been beaten by her mother. We informed the prosecutor for the investigation of the case. The girl was transferred to a shelter.

Lastly, in the interview, the hotline operator was asked to evaluate the significance of some sources related to the investigation process (1 to 5, with 5 being the most important):

Table Annex III- 1 Evaluation of the significance of sources related to the investigation process by the hotline operator

Social Media	5
Parents/guardians testimonies	5
Friends/relatives testimonies	5
Anonymous tips	5
Past case of the same child (if existing)	5
Past similar cases	5
Transportation data (e.g. bus/train schedules)	4
Events data (e.g. big music event in the city)	2
Weather data	4
Criminal activity data of the area	5

III.1.2 Interviews with canine search unit member

“The Smile of the Child” has developed with valuable contribution of volunteers the “Thanasis Makris” Search and Rescue Team for missing children. The aim of this team is to conduct search and rescue in urban or non-urban environment to trace missing children. Key members of the Search and Rescue Team “Thanasis Makris” are international certified dog teams, while a Mobile Command Centre and several operational tools (jeeps, ATV, motor vehicle, ambulances) are also at the disposal of the team.

The team is in direct cooperation with the competent authorities to coordinate the search and in direct connection with the European Hotline for Missing Children 116000. All participants are trained and make all their resources and expertise available to the team as the main purpose is the immediate response and activation of all actors involved, as well as the optimal use of existing means to find a missing child. Key members of Missing Children Search and Rescue Team «Thanasis Makris» are the Canine Teams participating. The Canine Teams (handler and dog) are the only ones in Greece certified by the international NSARDA body (National Search and Rescue Dog Association UK) and the only Canine Teams in Greece, seeking a person. The Search and Rescue Team for Missing Children «Thanasis Makris» was created by «The Smile of the Child» in 2012.

The Search and Rescue Team for Missing Children «Thanasis Makris» is activated once there is an extremely urgent case. Since, the Hotline considers essential the Team`s contribution, there is an exchange of information about the case and then the Team gets into action by offering the services and planning the actions accordingly. This takes place 5 – 10 times annually. The most frequent cases are missing minors who their lives are at risk independently of the category of disappearance. Consequently, the cases vary from runaways and abductions to lost, injured or otherwise missing children etc.

The most critical step when missing a child is to immediately mobilise the rescue teams. The best practices in this field are mainly the immediate availability of the specialised volunteer rescuers and the close cooperation with the respective authorities. The most important partners are the police department, the special unit of the fire brigade, Hellenic Coast Guard and local voluntary rescue teams. The biggest challenge to tackle during a missing case is the time since it is critical to discover the child as soon as possible and prevent any further implication to the condition of the child. The general public can be informed about the case via the activation of Amber Alert (to inform the general public) and via specialised volunteers that contribute to the search. The most important information needed is the age of the child, the health status (including mental or psychological issues), the family environment and finally the conditions under which the child disappeared (place and time of disappearance, timing of incident report). The most critical information that facilitates to great extent the recovery of the child is their personal smell in order to be "utilised" by the K9 SAR Teams.

Lastly, the canine search unit member was asked to evaluate the significance of some sources related to the investigation process (1 to 5, with 5 being the most important):

Table Annex III- 2 Evaluation of the significance of sources related to the investigation process by the canine search unite member

Social Media	4
Parents/guardians testimonies	4
Friends/relatives testimonies	4
Anonymous tips	2
Past case of the same child (if existing)	4
Past similar cases	4
Transportation data (e.g. bus/train schedules)	3
Events data (e.g. big music event in the city)	2
Weather data	4
Criminal activity data of the area	3

III.2 Insights from the Hellenic Red Cross

III.2.1 Interviews with the Danish Red Cross

The Danish Red Cross Asylum department has been running centres for unaccompanied minors for over 25 years. The number of centres has fluctuated in accordance with the numbers fleeing from conflict areas and poverty. At one point, the Red Cross was running eight centres, but there is only one centre left, namely reception centre Gribskov. However, Gribskov is closing too and the job of receiving unaccompanied minors will move to Denmark's main reception centre (for adults), where a protected section is being established for this job.

During this long specialisation, the job of receiving minors seeking asylum has been undertaken primarily at Gribskov (barring periods where the centre had other functions). The reception centre accommodated the minors for a period of 6-8 weeks (optimally), where after the children moved on to residential centres, some under the Red Cross flag and others run by local authority operators.

During the reception period, children experience a thorough and structured programme from the Red Cross including:

- Reception interview
- Introductory interview
- Health and medical screening
- Psychological screening
- Initial 'Life Skills' training
- Screening for trafficking
- TB screening (RC – health service)
- Information: asylum, rights, local and societal orientation
- Development plans constructed with the inclusion of the child and highlighting areas for their social, psychological and cognitive development

Additionally, case assessment by the immigration authorities starts during the reception period:

- ID interview
- Information and motivation interview
- Age assessment (not all)
- Asylum interview

Each child is appointed a primary care team of 2 contact persons. Their job is to assist the child in all aspects of everyday living at the centre, keep the child orientated on the coming schedule (the many appointments listed above) and, with the principles of self-help, support the child in navigating through a very difficult period of their lives – coping strategies. The centre staff tries to establish some continuity for the child: where do you come from, where are you now and where are you going, while focusing on the child's resources and competencies.

The Danish Red Cross asylum department does not keep statistics of the number of child disappearances. However, it is estimated that, approximately 60-70% of unaccompanied minors choose to 'move on' from the reception centres. There are different reasons for the high rate of children leaving the shelters, including, but not limited to: Denmark being a transit country for other

Nordic destinations such as Sweden, Norway or Finland; their asylum application being rejected and their deportation being imminent; the nomadic socialisation of some groups of Northern African minors, who often return regularly to Denmark, using different names/identities. The Danish Red Cross mostly handles cases of unaccompanied minors seeking protection and nomadic groups, mainly North Africans, specifically from Morocco, Tunisia, and Algeria, who live on the streets in the big cities of different European countries and survive through organised (street) crime and engaging with the system to get a little respite from street life. Occasionally young women (often of African descent), who have been picked up by the authorities in street prostitution or brothel environments of Copenhagen are also involved. With regard to unaccompanied minors who disappear, Danish Red Cross is powerless to intervene as it only has a documentation and orientation responsibility, which was identified as one of the biggest challenges in the work process of the Danish Red Cross.

In order to establish potential places of interests of the missing child, knowledge on the possible involvement of the child in street crime, an ongoing risk assessment including the testimonies of the practitioners working at the shelters and the knowledge of the location of other family members are essential. The criminal activity data of the area as well as the past case file of the child were identified as the most important sources of information for the search for missing children, which were rated with a 5 (the highest mark) on the scale.

Most important pieces of information to determine what kind of case it is (kidnapping, runaway etc.) were identified to be:

- Age
- Gender
- Nationality
- Context in which they are picked up e.g. at the border, on the street, raid on red light districts
- Mental state

In order to locate missing children, the following information is crucial:

- Knowledge of intended destination
- Time of disappearance
- Build-up of trust while they are residents at the centre: This contributes to our knowledge of the child, their motivation for being here, their intentions but most of all the building of relationships between the child and a primary adult. While respecting the child's right to privacy, anonymity and confidentiality, we also have a duty of care. Risk assessments are undertaken throughout the process.
- Time is a crucial issue, as many of these children have met unfriendly adults and authorities in their home country and/or during their flight. Trust takes time. That emphasises the necessity for trained professionals. Informing children of their rights and opportunities so they can make informed choices is essential but takes time to communicate adequately and appropriately.

III.2.2 Interviews with the British Red Cross

'Surviving to Thriving' is a partnership project between the British Red Cross, the Refugee Council, and UpRising²⁶. Based in Birmingham, Luton and Leeds, the project is enabling 500 young refugees and asylum seekers to become active and valued members of their communities through a range of support services and social action.

The project provides practical support to young refugees through helping them to develop life skills and offering advice and mental health support. The project also enables these young people to boost essential skills, such as leadership and employability, to ease integration into their new environments.

Additionally, the project works closely with several Local Authorities, building their capacity to deal with the specific needs of young refugees and to engage with - and learn from - their experiences.

The British Red Cross in the framework of the project offers one-to-one casework, dealing with the specific needs of young refugees; group sessions, designed to create social networks while increasing knowledge, skills and confidence; and help with access to services, including legal representation. Since July 2017 British Red Cross has directly supported just over 380 children and young people.

By experience, the main issues for children and young people going missing from care were due to concerns over being re-trafficked in the UK and then going into hiding when they have no immigration status and are terrified of being detained or removed from the UK. A big challenge is the lack of understanding amongst professionals about the vulnerabilities of unaccompanied children particularly to be re-trafficked and exploited. By experience, often social workers and police do not know what trafficking is, how to spot indicators or respond including not having heard of the NRM (process by which someone is recognised as a victim of trafficking in the UK). This means the crucial time period when trafficked children go into care the appropriate steps are not put in place to protect them which can contribute to increased risk of them going missing and returning to situations of exploitation.

Most important pieces of information to determine what kind of case it is (kidnapping, runaway etc.) are the following:

- Testimony of the missing child to their social worker, foster care, voluntary sector staff etc. before going missing
- Immigration status can be a relevant factor particularly if they have received a refusal on their asylum claim and are fearful of detention or removal (once over 18)
- Country of origin can be relevant and underlying risk factors or vulnerabilities as well as known history such as whether they were trafficked to the UK or have been exploited since arriving in the UK.

The Young Refugee Service of British Red Cross works with unaccompanied refugees and asylum-seeking young people, aged 15-21, supporting them with the asylum process and the social care system, as they transition into the adulthood.

The Young Refugee Service offers support for young refugees in a few places in the UK – Birmingham, Glasgow, Hampshire, Kent, Leicester, Leeds, London and Luton. The British Red Cross

²⁶ <https://www.redcross.org.uk/about-us/what-we-do/how-we-support-refugees/surviving-to-thriving>

has supported 110 UASC cases and handled 10 missing children cases during the last year. Most of the cases handled were categorised as runaways. In order to recover missing children, the process of understanding the child's motives for disappearing, identifying possible places of interest to the child, and contacting all stakeholders as well as the police and the social services are crucial. However, the limited time available to each case and the culture of disbelief in the Home Office were named as the biggest challenges in the effective and timely recovery of missing children. Children that go missing commonly suffer from trauma or mental health issues, have been removed or transferred from their accommodation against their own wishes, and have a marked mistrust of authorities. The social media accounts as well as the testimonies of parents or guardians were identified to be the most significant sources of information for the investigation process, both rated with the highest mark on the scale.

Depending on what services are available in each area, this could include help with understanding the asylum process, preparing documents, supporting children in health, education or social care crises, or informing the children about their possibility of getting support.

Most important pieces of information to determine what kind of case it is (kidnapping, runaway etc.) are the following:

- Case notes prior to child going missing e.g. Change of accommodation, any complaints from child, any relevant incidents.
- Mindset of child before went missing – provided by those working closely with them.
- Any history of trafficking
- Any indicators of trafficking

Important pieces of information to determine possible places of interest for the missing children:

- Social circles and networks
- Accommodation history
- Trafficking risk

III.3 Insights from expert interviews in Germany

III.3.1 State actors – German Youth Institutes

Two interviews were conducted with members of the German Youth Institutes, one of which works in German child protective services and one from the youth services. The interviewee from the child protective services is mainly in charge of runaway cases, while the representative of the youth services mostly deals with unaccompanied minor migrants. As state actors, the child protective services as well as youth services for unaccompanied migrant minors have custody of the children in their care. While every child has a specific appointed guardian, living in an institution leads to stricter rules on processes in case children go missing.

The first interview was undertaken with a staff member of the child protective services, whose department does not work in a foster care facility, but rather organises the cases in the institution itself. Thus, the interviewee is not involved in the search for missing children unless directly requested by the police in cases of potential trauma to the children. The second interview was conducted with a member of the youth services, who supervises the shelters for unaccompanied minor migrants and is thus involved directly in reporting unaccompanied minor migrants who are in the care of the state and aids more directly in the search process for missing unaccompanied migrant

minors. The child protective services, in contrast, are not directly involved in the search for missing children unless the child contacts them or the police asks them to accompany them in case the child might be traumatised, but rather act as an intermediary between the child and the legal guardians after they have been recovered. The interviews with different state actors involved in handling missing children cases on different levels enlightened different aspects of relevance to the ChildRescue project as they both pointed out indicators for a classification of missing children cases in different areas and shed light on different procedures in the search for missing children.

The child protective services are especially involved in cases of runaways, which pose the majority of their case load with missing children, and, to a lesser degree, with cases of parental abductions or kidnappings by strangers. A crucial factor in deciding how to classify the case is the age of the child, as running away rarely occurs in young children, which fosters the suspicion of the child being taken rather than having left. If necessary, the child protective services will accompany the legal guardian to the police to file a missing person report. Another factor that was identified as being helpful in deciding if the child had run away was the timing of the incidence as many adolescents run away on the last day of school due to bad grades that they are afraid to show at home. While it was stressed that every case should be judged on an individual basis, the common theme that emerges in cases of children running away from home is the subjective feeling like the family situation is unbearable. This family situation can however differ vastly and spans from temporary, pubescent conflicts like not being allowed to stay out late for parties over difficult divorces that burden the child's mind up to physical and sexual abuse. The interviewee at the child protective services told that 1-2 incidents of missing children are reported to her per week, most of those are resolved quickly. Depending on the conflict, the resolutions and reactions of the child after the initial incidence of running away can also differ. In some cases, a reunification with the parental home is not desirable, especially when sexual or physical abuse has occurred in the past. In cases of parental abductions, the decision is especially complicated as it is not always apparent what serves the best interest of the child, which has to be the main priority. In any case, if a child has gone missing, the police will be alerted as they are the main state actor in the search for missing children with the authorisation to initiate search procedures.

Youth services for unaccompanied minor migrants are in charge of finding a suitable shelter for these children, which rarely involves foster families but rather state-run youth shelters, which have strictly enforced rules and regulations. This leads to a more direct involvement in the search for missing children as they are the legal guardians of those missing children. The shelters enforce curfews for the children in their care, depending on the age of the child. Consequently, any teenager who misses their curfew by more than two hours will be reported to the police as a missing person. In cases involving younger children this procedure will be sped up accordingly. However, as approximately half of the teenagers miss their curfews by more than two hours, the active search will often not be initiated by the police until the next morning, unless it is a case of a high priority, such as young children, a child with a disability or an unknown location (who cannot be placed at a party or other gathering by peers). If the teenager does not return by the next morning or the case has been declared high priority from the start, the staff members will interview peers, family members and staff members at other shelters or youth agencies the missing child might have been in contact with. They also search for the social media profiles of the missing child and analyse their content in search for

clues to the location or activity of the child. If available, the current phone number of the missing child and its parents will also be utilised in the search for the child, although the new General Data Protection Law complicates this process. Data of current events, such as concerts in the area, as well as the case file of the child and the profile on dating apps were identified as the most important information sources in looking for the missing child, followed by transportation data such as timetables, interviews with family or legal guardians and social media accounts. The crime statistics of the area were deemed the least helpful. The risk of victimisation was estimated to be especially high for young girls who have gone missing after arranging a date utilising a dating app. In comparison to Germany, Australia and the UK were named as positive examples of handling missing children cases, because the data protection law in Germany was viewed as too strict. Rather the child's welfare, both mental and physical, should be prioritised. However, some instances of children going missing are announced by the child prior to their disappearance. In cases of unaccompanied minor migrants this is often due to frustration about the lengthy bureaucratic process of obtaining legal citizenship or being allowed to proceed to the desired country of destination. Despite efforts of the shelters to stop the disappearance through measures such as taking away of passports, unaccompanied minors still leave regularly and put themselves at further risk of victimisation.

III.3.2 Specialised NGOs

Two interviews with specialised NGOs were conducted. The first interview was conducted with a staff member of a NGO targeting homeless youth called 'OffStreet Kids'. The second interview was conducted with an employee of the NGO 'Zora', which focusses on girls and young women in difficult situations.

In Germany, street youth includes mostly children over the age of 15 who are mostly throwaways or pushaways. Runaways, who leave spontaneously after a fight, less frequently experience homelessness as they often return home after a relatively short period of time. There are various non-governmental organisations (NGOs) who target the children on the streets and provide services which are often anonymous. Due to the anonymity children who were reported missing are rarely recovered through these organisations as they are not identified as a missing person, but they do receive counselling on the issues that led to the incident of leaving the home. Consequently, they offer insights into the most extreme cases of children that are absent from home over long periods of time, despite not necessarily being reported as missing from their parents.

Children, who lived in institutions such as foster homes before going missing, are regularly reported to the police, whereas children who leave parental homes are underreported. The NGOs offer different services such as anonymous face-to-face and online counselling, parent-child mediation and support in family court cases. The most crucial information was identified as anything health related as issues in that area might need prompt solutions. Further issues identified were the existence of a warrant of arrest issued in the name of the child as that can also influence the behaviour of the child as they might go into hiding from authorities and are thus even less likely to seek help in case of their victimisation. The most important source of information is the children themselves, who will only come forward if a foundation of trust can be established, thus an anonymous counselling is of vital importance to open possibilities of safe return from a life on the street for these children, which was stressed by both NGOs represented in this sample. The reasons for running away were located in the

family situation and often involved mental overload on the side of the parents or physical and sexual abuse of the child at home.

The NGO named Zora specifically targets girls and young women and offers them possibilities to shower in and a second-hand closet for fresh clothes – thus, the organisation is exclusively open to women and does not allow access by males. The counselling covers many topics, ranging from issues at school, violence at home or romantic relationship over substance abuse, unwanted pregnancies to mental health problems and forced marriages – although counsellors recommend also specialised counselling centres, many of the girls who have been absent from home for a longer period of time and live on the street will not reappear at another counselling centre due to a lack of trust in unknown people. Depending on the age and the mental health status of the girls, shelter will be arranged, which will lead to a closing of missing person file if they are reported. However, less than 10 cases were identified as missing persons in 2017. Social media analysis and interviews with friends or family members were identified as the most important sources for information. Additionally, the experience and knowledge of street workers in the field of street youth were deemed vital, as they know where popular 'hang-outs' of these children are and can thus check them promptly. Ideally, the current phone number and the full name would be needed in the search for missing children.

III.3.3 Police (Germany/UK)

During the course of this study, two interviews with members of different police forces were conducted. One interview was conducted with a police officer from the German federal police force called "Bundeskriminalamt" or BKA. The BKA police officer works in the taskforce on missing or unidentified deceased people ('vermisste und unidentifizierte Tote' or 'vermi/utot'). An additional interview was conducted with a member of the police force in the United Kingdom, who is in charge of implementing different software solutions in the work of the missing person taskforces in the UK. These software solutions can cause synergy effects with the ChildRescue objectives.

The police are the primary actor in charge of the recovery of missing person, who are in charge of overseeing the collaboration with other actors as well as execute the search missions. The German federal police (BKA), is in charge of co-ordinating international missing persons cases with the federal police forces in the other country involved in the case, but is not directly involved in the search process. The task of the BKA lies mostly in co-ordinating with foreign police officers to recover missing foreign minors within Germany or missing German minors in foreign countries. The United Kingdom, in turn, already implements specific software products in cases of missing persons that allow for geographical predictions of likely recovery sites based on past similar cases and that allow to contact missing persons in order to ensure their safety.

The BKA has a taskforce on missing and unidentified deceased people ("vermi/utot"), whose responsibility lies in aiding local police forces in cases involving foreign nations. This includes both cases of missing persons, who are German but went missing abroad and foreigners who went missing in Germany as well as unidentified bodies with presumed ties to other countries. Their involvement in missing person cases contains missing children, including missing minor migrants and cases of runaways. If faced with minor runaways who are threatened by forced marriages, the police still have to inform the parents or other legal guardians of the recovery. In cases of unaccompanied minor migrants, the police force of the last known country of residency will be contacted by the BKA when the child or the bodily remains are found. Cases of (international) parental abductions, however, are

not handled within the taskforce despite a missing person's report being issued for the child and parent. The cases are handled by the international legal aid instead. The vermi/utot taskforce of the BKA is informed by the local police force, which conducts the investigation into the cases, and receives all details gathered in standardised forms in the beginning of the investigations. The BKA taskforce in turn belongs to and cooperates with the Interpol affiliate in the other involved country, resulting in a differing quality of collaboration and response-time depending on the structure of the local Interpol office. This was identified as one of the major issues in the work process, as it can slow the recovery process down. However, the existing liaison officers were named as an example of a good practice to increase the speed of collaboration that should be broadened. The vermi/utot taskforce is thus mostly in charge of communicating with the local German police force and the foreign country's affiliate and acts as a mediator in between the two. They deal with few cases of third-person abductions but are heavily involved in youth who are threatened or victimised by forced marriages as well as runaways, who are romantically involved with a person in a different country or, recently, German youth joining ISIS. Thus, the group of minors who run away cannot be universally characterised as the reasons for running away vary greatly from problems at school over family conflicts to outsider-influence, such as romantic internet-relationships. The internet is used in the search process to identify key persons in the social network of the missing person through social media accounts as well as gather clues of potential locations of the missing child by reviewing places the child has been to in the past. The past case history of the minor, if available, was identified as the most helpful instrument to solve these cases, with statements from parents, friends and family members, past similar cases and flight schedules also being pointed out as vital, while bus and train schedules were named as the least helpful. Ideally, the person reporting a child missing should give indicators about their character, such as if the child is trusting toward strangers, potential locations or safe spaces of the minor in order to improve the search process.

In the United Kingdom police force, there are several different software solutions available which aid in missing person cases. The most widespread one is called COMPACT and has been in place since 2000. COMPACT started as a database for past cases of missing persons that was turned into a case management system that is used in 21 districts. The case management system eases the transition between different case workers as it collects essential information on the case and offers the option to create to-do-lists that highlight past actions and future tasks. Additionally, the iFind App was created for the police force, which can be used throughout the whole UK and aggregates data on past recovery locations of persons in similar cases, which can be used to inform the search process. It further offers statistical probability analyses of finding the missing person in a certain location based on the characteristics of the person. The App started on the basis of 1.500 analysed cases and improves over time as further use enriches its data base and resulting accuracy. The App was launched on the 22.08.2018 and is consequently still in the very early stages, making it impossible to predict its practical applicability or success rate. Additionally, Textsafe and Suicide Textsafe are utilised by the police to contact missing persons in order to reach people in need anonymously and ensure their physical and emotional safety if they decide to remain missing. The police force classifies missing person's cases into categories of 'high', 'medium', 'low' and 'no apparent' risk, which influences the further steps. In cases with no apparent risk, the police action is paused until a deadline, which leads to a review of the classification. Minors can also be classified as no apparent risk, leading to no police search for them in the beginning of the missing period. This can lead to

dangerous situations for the missing child, especially if the police force is enticed to classify cases as no apparent risk due to their limited resources. The majority of cases handled by the police are runaways with a high percentage of repeat missing instances and an elevated number of children from care or hospitals who run away. The cases are classified on the basis of a risk assessment form, which include information on the character of the child such as whether running away is atypical to the character of the child, previous instances of going missing. Cases classified as high-risk can additionally be searched for with the aid of drones that are being sent to areas that are considered to be high-probability locations due to past similar cases and the statements of friends and family members. However, the classification is not based on one single item in the risk assessment but rather a triangulation of the aggregated data. In repeat cases of missing minors, the child protective services and the police will be involved in creating preventive strategies upon the recovery of the child in order to obstruct further instances of going missing. The most important information sources for missing children aged 12 to 18 were identified to be similar past cases as well as the risk assessment form. Ideally, a national database to compare recovery sites from similar past cases should be established to improve the search process in order to cover unusual cases as well, which can as of yet not necessarily be achieved by the relatively new iFind App.

III.3.4 Current researchers on related areas (Paedophilia, homeless youth)

Two current research projects revolving around topics correlated to the aim of ChildRescue were also included in the sample of interviews to gain better insight into the data, in addition to including research reports in the literature review, namely MiKADO²⁷ and the Streetyouth study in Germany by the German Youth Institute²⁸ (DJI).

Specifically, MiKADO (abuse of children: aetiology, darkfield, victims, *in German: Missbrauch von Kindern: Aetiologie, Dunkelfeld, Opfer*) is a study that was conducted over 3.5 years in Germany and Finland and was financed by the German Ministry of Family, Senior Citizens, Women and Youth. It was chosen due to its focus point on sexual misconduct towards children online and related grooming processes, which could potentially be helpful in identifying the location of a child that was targeted for offline sexual exploitation. The MiKADO project was organised in three sub-studies on sexual abuse of children, reasons for sexual interest in children and the darkfield of victims, which were all considered for this deliverable although the focus was placed on the darkfield as this study explored the internet as a danger for sexual abuse of children. The research project conducted a series of surveys with adults and children to include both perspectives. 28.000 adults completed the survey as well as over 2.000 minors. It aimed at developing preventive strategies to obstruct the further sexual abuse of children off- and online.

Additionally, the German Youth Institute (DJI) was included due to their study on the life situation of children on the streets of Germany, which was published in a comprehensive report on the different situations and needs of streetyouth in 2017. Prior to the publication of the report, the DJI had supervised four model projects and interviewed both the leaders of the model projects as well as

²⁷ <http://www.mikado-studie.de/index.php/101.htm>.

²⁸ www.dji.de/strassenjugendliche.

youth participants. Additionally, 300 in-depth interviews with street youth under the age of 25 were conducted.

MiKado discovered that 97% of teenagers are online daily or multiple times a week, making the internet an important environment for children and placing crucial parts of their social networks online. However, this also makes them vulnerable to being approached inappropriately by adults and indeed approaching adults themselves. The choice of the individual child is dependent on the preferences of the adults who engage with the child. Adults will target children specifically by their preferred age and gender. Online communication is often explicit in nature before meeting offline does happen. Offline meetings occur in 25% of all cases in which the child only knows the adult from online interactions. The grooming process can vary in form and time. The analysis of private messages could theoretically be most informative as much of the sexual misconduct happens online, however publicly available data from social media apps, such as likes on photos, can already serve as an indicator for a close relationship with an adult.

The study conducted by the German Youth Institute, which focusses on adolescents living on the street supports the interview results from the NGOs. Children living on the streets, who are mostly teenagers, are usually "de-coupled" from the system, meaning they no longer attend school or any other educational institution or work. They mostly sleep rough at times and "couch-hop" at other instances. Girls especially often stay at older acquaintances homes, which put them at risk of sexual exploitation or abuse. Reasons for leaving their homes are mostly family issues ranging from abuse over neglect to mental health issues. Homeless youth often have irregular access to social media, but do make use of Facebook groups to gain access to help, which is critical due to data protection issues. Online counselling, as offered by many NGOs, is especially helpful for children who are couch-hopping and thus have access to a computer.

Annex IV: Past cases list of reference

Table Annex IV- 1 List of reference for past cases

#	Pilot Name	Case ID	Date Created	Date Closed	Notes
1	Child Focus	CM: 12345	31-6-2009 11:45	27-08-2009	
2	Child Focus	CM: 580	30-12-2009 11:15	03-01-2010	
3	Child Focus	CM: 1271	18-04-2010 16:36	24-05-2010	
4	Child Focus	CM: 4173	22-05-2011 18:36	25-05-2011	
5	Child Focus	CM: 5070	6-10-2011 20:18	16-10-2011	
6	Child Focus	CM: 5318	15-11-2011 13:43	17-11-2011	
7	Child Focus	CM: 6044	7-03-2012 12:18	07-06-2012	
8	Child Focus	CM: 6498	20-05-2012 19:00	03-06-2012	
9	Child Focus	CM: 7054	11-08-2012 16:21	30-10-2012	
10	Child Focus	CM: 7790	12-12-2012 12:03	17-12-2012	
11	Child Focus	CM: 8797	13-06-2013 23:06	25-06-2013	
12	Child Focus	CM: 11637	20-11-2014 22:05	09-12-2014	
13	Child Focus	CM: 13376	8-11-2015 19:17	09-11-2015	
14	Child Focus	CM: 13823	15-02-2016 10:06	28-02-2016	
15	Child Focus	CM: 14196	26-04-2016 19:18	09-05-2016	
16	Child Focus	CM: 14286	16-05-2016 19:27	17-05-2016	
17	Child Focus	CM: 15868	18-03-2017 14:13	22-03-2017	
18	Child Focus	CM: 16214	26-05-2017 23:38	10-01-2018	
19	Child Focus	CM: 16563	25-07-2017 0:42	25-07-2017	
20	Child Focus	CM: 16931	30-09-2017 19:51	01-11-2017	
21	Child Focus	CM: 123456	6-04-2009 10:56	19-04-2009	
22	Child Focus	CM: 2676	24-10-2010 10:40	26-10-2010	

23	Child Focus	CM: 4183	23-05-2011 19:03	26-05-2011	
24	Child Focus	CM: 4395	23-06-2011 12:53	07-07-2011	
25	Child Focus	CM: 4643	31-07-2011 0:50	04-08-2011	
26	Child Focus	CM: 5009	25-09-2011 18:27	28-09-2011	
27	Child Focus	CM: 6805	5-07-2012 12:11	09-07-2012	
28	Child Focus	CM: 7501	16-10-2012 18:49	23-10-2012	
29	Child Focus	CM: 7643	14-11-2012 7:32	18-11-2012	
30	Child Focus	SM: 8196	21-02-2013 8:17	16-09-2013	
31	Child Focus	CM: 8719	30-05-2013 21:37	06-06-2013	
32	Child Focus	CM: 9359	14-09-2013 18:33	15-09-2013	
33	Child Focus	CM: 9828	16-12-2013 13:35	13-01-2014	
34	Child Focus	CM: 10511	3-05-2014 12:16	04-05-2014	
35	Child Focus	CM: 11207	7-09-2014 10:36	11-09-2014	
36	Child Focus	CM: 11286	16-09-2014 15:17	08-10-2014	
37	Child Focus	CM: 11479	20-10-2014 14:36	21-10-2014	
38	Child Focus	CM: 11886	17-01-2015 10:39	27-01-2015	
39	Child Focus	CM: 15181	25-10-2016 3:16	02-11-2016	
40	Child Focus	CM: 17019	16-10-2017 15:51	18-10-2017	
41	Smile of the Child	SOCB1	21-05-2018	02-06-2018	Alarming disappearance 1
42	Smile of the Child	SOCB2	17-09-2016	17-09-2016	Alarming disappearance 2
43	Smile of the Child	SOCB3	23-01-2018	27-01-2018	Alarming disappearance 3
44	Smile of the Child	SOCB4	12-10-2017	12-10-2017	Alarming disappearance 4
45	Smile of the Child	SOCB5	07-07-2017	09-07-2017	Alarming disappearance 5
46	Smile of the Child	SOCB6	08-03-2017	10-03-2017	Alarming disappearance 6
47	Smile of the Child	SOCB7	24-06-2018	26-06-2018	Alarming disappearance 7
48	Smile of the Child	SOCB8	15-09-2017	16-09-2017	Alarming disappearance 8
49	Smile of the Child	SOCB9	22-03-2017	23-03-2017	Alarming disappearance 9
50	Smile of the Child	SOCB10	02-10-2018	03-10-2018	Alarming disappearance 10

51	Smile of the Child	SOCC1	09-11-2016	23-11-2016	Parental abduction 1
52	Smile of the Child	SOCC2	10-09-2015	11-10-2015	Parental abduction 2
53	Smile of the Child	SOCC3	07-03-2015	Open Case	Parental abduction 3
54	Smile of the Child	SOCC4	30-03-2016	01-04-2016	Parental abduction 4
55	Smile of the Child	SOCC5	08-01-2017	09-06-2017	Parental abduction 5
56	Smile of the Child	SOCD1	24-05-2016	30-05-2016	Missing unaccompanied migrant minors 1
57	Smile of the Child	SOCD2	23-09-2015	Open Case	Missing unaccompanied migrant minors 2
58	Smile of the Child	SOCD3	24-02-2017	25-02-2017	Missing unaccompanied migrant minors 3
59	Smile of the Child	SOCD4	22-08-2016	Open Case	Missing unaccompanied migrant minors 4
60	Smile of the Child	SOCA1	29-04-2016	01-05-2016	Runaways of teenagers 1
61	Smile of the Child	SOCA2	22-11-2015	22-11-2015	Runaways of teenagers 2
62	Smile of the Child	SOCA3	15-04-2015	10-06-2015	Runaways of teenagers 3
63	Smile of the Child	SOCA4	04-05-2015	06-05-2015	Runaways of teenagers 4
64	Smile of the Child	SOCA5	10-09-2015	12-09-2016	Runaways of teenagers 5
65	Smile of the Child	SOCA6	30-09-2016	03-10-2016	Runaways of teenagers 6
66	Smile of the Child	SOCA7	14-10-2016	01-11-2016	Runaways of teenagers 7
67	Smile of the Child	SOCA8	07-04-2016	29-04-2016	Runaways of teenagers 8
68	Smile of the Child	SOCA9	07-04-2018	10-04-2018	Runaways of teenagers 9
69	Smile of the Child	SOCA10	04-05-2017	07-05-2017	Runaways of teenagers 10
70	Smile of the Child	SOCA11	18-03-2017	20-03-2017	Runaways of teenagers 11
71	Smile of the Child	SOCA12	04-04-2018	13-04-2018	Runaways of teenagers 12
72	Smile of the Child	SOCA13	19-03-2018	23-03-2018	Runaways of teenagers 13
73	Smile of the Child	SOCA14	12-05-2017	03-06-2017	Runaways of teenagers 14
74	Smile of the Child	SOCA15	11-01-2018	14-01-2018	Runaways of teenagers 15
75	Smile of the Child	SOCA16	30-01-2016	01-09-2016	Runaways of teenagers 16
76	Smile of the Child	SOCA17	29-03-2018	20-04-2018	Runaways of teenagers 17

77	Smile of the Child	SOCA18	04-06-2018	07-06-2018	Runaways of teenagers 18
78	Smile of the Child	SOCA19	23-02-2018	26-02-2018	Runaways of teenagers 19
79	Smile of the Child	SOCA20	18-05-2018	08-06-2018	Runaways of teenagers 20
80	Smile of the Child	SOCA21	17-04-2017	17-04-2017	Runaways of teenagers 21
81	Smile of the Child	SOCA22	11-02-2018	16-02-2018	Runaways of teenagers 22
82	Smile of the Child	SOCA23	29-10-2015	06-11-2015	Runaways of teenagers 23
83	Hellenic RedCross Tracing Division	GRC-001854	07-12-2016	23-01-2017	
84	Hellenic RedCross Tracing Division	GRC-001854	07-12-2016	23-01-2017	
85	Hellenic RedCross Tracing Division	GRC-001175	25-11-2015	26-11-2015	
86	Hellenic RedCross Tracing Division	GRC-001175	25-11-2015	26-11-2015	
87	Hellenic RedCross Tracing Division	GRC-001175	25-11-2015	26-11-2015	
88	Hellenic RedCross Tracing Division	GRC-002221	30-06-2017	24-07-2015	
89	Hellenic RedCross Tracing Division	GRC-002121	05-05-2017	12-06-2017	(still open-lost after located)
90	Hellenic RedCross Tracing Division	GRC-002121	05-05-2017	12-06-2017	(still open)
91	Hellenic RedCross Tracing Division	GRC-001372	07-03-2016	OPEN	(still open)
92	Hellenic RedCross Tracing Division	GRC-001404	30-10-2015	21-03-2016	
93	Hellenic RedCross Tracing Division	GRC-002216	22-06-2017	04-07-2017	
94	Hellenic RedCross Tracing Division	GRC-001974	25-01-2017	20-04-2017	
95	Hellenic RedCross Tracing Division	GRC-001700	07-01-2016	25-02-2017	
96	Hellenic RedCross Tracing Division	GRC-002384	25-10-2017	04-06-2018	
97	Hellenic RedCross Tracing Division	GRC-002294	24-08-2017	OPEN	(still open)
98	Hellenic RedCross Tracing Division	GRC-001380	12-03-2016	OPEN	(still open)
99	Hellenic RedCross Tracing Division	GRC-002438	06-11-2017	OPEN	(still open)
100	Hellenic RedCross Tracing Division	GRC-001378	10-03-2016	30-03-2016	
101	Hellenic RedCross Tracing Division	GRC-001424	10-03-2016	OPEN	(still open)
102	Hellenic RedCross Tracing Division	GRC-001056	02-10-2015	OPEN	(still open)
103	HRC Kalavryta Reception Center for UMC	2	28-07-2017	15-09-2017	
104	HRC Kalavryta Reception Center for UMC	4	28-07-2017	14-01-2018	

105	HRC Kalavryta Reception Center for UMC	5	28-07-2017	14-01-2018	
106	HRC Kalavryta Reception Center for UMC	6	28-07-2017	15-09-2017	
107	HRC Kalavryta Reception Center for UMC	7	31-07-2017	17-03-2018	
108	HRC Kalavryta Reception Center for UMC	8	31-07-2017	14-01-2018	
109	HRC Kalavryta Reception Center for UMC	12	31-07-2017	07-02-2018	
110	HRC Volos Reception Center for UMC	1383	28-08-2018	13-09-2018	
111	HRC Volos Reception Center for UMC	1379	27-08-2018	13-09-2018	
112	HRC Volos Reception Center for UMC	1378	23-08-2018	03-09-2018	
113	HRC Volos Reception Center for UMC	1380	28-08-2018	30-08-2018	
114	HRC Volos Reception Center for UMC	1381	28-08-2018	29-08-2018	
115	HRC Volos Reception Center for UMC	1330	30-06-2017	17-06-2018	
116	HRC Volos Reception Center for UMC	1366	30-03-2018	15-06-2018	
117	HRC Safe Zone program in Ritsona	358484	26-06-2018	01-08-2018	
118	HRC Safe Zone program in Ritsona	97665	04-05-2018	07-05-2018	
119	HRC Safe Zone program in Ritsona	81159	04-05-2018	07-05-2018	
120	HRC Safe Zone program in Ritsona	101067	02-03-2018	18-03-2018	
121	HRC Safe Zone program in Ritsona	66046	08-06-2017	09-08-2017	
122	HRC Safe Zone program in Ritsona	59481	12-05-2017	09-08-2017	

Typical examples of full cases

The table below lists 6 full cases (2 from each pilot partner) in a vertical form. More specifically, cases SOCA3 and SOCB10 from SoC, CM13376 and CM16931 from CF, and GRC-001404 and 66046 from HRC, are hereby presented. These representative cases showcase the data available, as well as a few semantical inconsistencies among pilots on the ways and formats that data are collected. A fact that is going to be altered with ChildRescue.

Table Annex IV- 2 List of 6 full cases of missing children (2 per pilot partner)

Name of field	Category	SOCA3	SOCB10	CM13376	CM16931	GRC-001404	66046
Currying Mobile phone?	Case data	Yes	Yes	Yes	Yes	No	Yes
Conditions of disappearance	Case data	While going to school with sister, the child said that will return home because of a body pain.	In the metro, Athens, Greece	On way to school	Was at home	Shipwreck	On trip to another city
Possible Reasons of disappearance	Case data	Relation with the opposite sex	Other	unknown	Left after an argument with father	irregular	in search for relatives
State of child when found	Case data	Married	Lost	normal	normal	drown at sea	unknown
Have money or credit card on her	Case data	N/A	Yes	Yes	Yes	N/A	Yes
Had area knowledge?	Case data	Yes	No	Yes	Yes	No	Yes
Rescue teams utilised	Case data	No	No	Yes	No	Yes	Yes
Transit country/-ies reached or intended to be reached	Case data	Portugal; UK	N/A	None	None	Turkey; Greece	None
Volunteers utilised	Case data	No	No	No	No	No	No
Organisations cooperated	Case data	5	2	N/A	N/A	N/A	N/A

Date of Arrival at hosting facility	Case data	N/A	N/A	N/A	N/A	N/A	8/6/2017 12:00
Type of disappearance (Category)	Case data	Runaway	Alarming Disappearance	Runaway	Runaway	Missing Unaccompanied Migrant Minor	Runaway
Area/Location child was last seen	Case data	On a way to school	In the metro, Athens, Greece	<i>Classified</i>	<i>Classified</i>	Turkey	<i>Classified</i>
Date and time child was found	Case data	10-06-15	03-10-18 5:30	09-11-15	02/10/2017 03h00	2016	unknown
Date and time of disappearance	Case data	15-04-15 8:00	02-10-18 0:00	08-11-15 7:45	29/09/2017 23h00	28.10.2015	9/8/2017 14:00
Multiple-times case	Case data	1	2	1	1	1	1
Case ID	Case data	SOCA3	SOCB10	CM13376	CM16931	GRC-001404	66046
Family members	Case data	5	6	4	3	N/A	27
Probable destinations (location/city/country)	Case data	Unknown	Unknown	None	friend's house in Charleroi	Greece-Lesvos	Germany
Clothing with Scent (for dogs)	Case data	No	No	No	Yes / No	N/A	N/A
Home / Facility Address (area only)	Demographics	<i>Classified</i>	<i>Classified</i>	<i>Classified</i>	<i>Classified</i>	<i>Classified</i>	<i>Classified</i>
Education level & current participation in educative activities	Demographics	Secondary School	N/A	BuSo	secondary	N/A	first grade

Languages spoken (number of)	Demographics	2	2	1	1	1	Arabic, English
Nationality	Demographics	Pakistani	French	Belgian	Belgian	Kurdish-Iraqi	Syria
Place of birth / Country of origin	Demographics	Pakistani	France	Belgium	Belgium	Iraq	Syria
Age (or Birthday)	Demographics	15	14	16-04-00	10-10-01	17.03.2014	05-01-01
Sex	Demographics	Female	Male	Female	Female	Male	Male
Home / Facility post code	Demographics	N/A	N/A	<i>Classified</i>	<i>Classified</i>	N/A	<i>Classified</i>
Health issues (current)	Medical Profile	No	No	No	drug user	N/A	N/A
Medical treatment required?	Medical Profile	No	N/A	No	No	N/A	No
Social media accounts	Personality/Social Profile	Facebook	Facebook	unknown	snapchat	N/A	No
Protection concerns/ Vulnerabilities	Personality/Social Profile	No particular concern	Probably	in foster care	drug use	N/A	other, Unaccompanied minor
Specific personality characteristics/psychological disorders	Personality/Social Profile	No particular disorder	Yes	autism	suicidal	N/A	depressive
Family status	Personality/Social Profile	Living with both biological parents	Living with both biological parents	living in foster family	living with biological father and stepmother	N/A	(unaccompanied child)
Parents' (Tracing enquirer) profile evaluation	Personality/Social Profile	Sufficient	Good	Good	Father good, mother very bad	N/A	Sufficient
School grades, Absences	Personality/Social Profile	N/A	N/A	Student with normal absences	unknown	N/A	A' student with normal absences

Interests/Hobbies	Personality/Social Profile	N/A	N/A	unknown	Select multiple from list	N/A	music, football, dancing, painting.
Relationship status (single, in a relationship, etc)	Personality/Social Profile	N/A	N/A	unknown	in a relationship	N/A	single
Religion	Personality/Social Profile	N/A	N/A	unknown	unknown	N/A	<i>Classified</i>
Weight (Kg)	Physical data	<i>Classified</i>	<i>Classified</i>	<i>Classified</i>	<i>Classified</i>	<i>Classified</i>	<i>Classified</i>
Height (cm)	Physical data	<i>Classified</i>	<i>Classified</i>	<i>Classified</i>	<i>Classified</i>	<i>Classified</i>	<i>Classified</i>
Distinguishing features	Physical data	<i>Classified</i>	<i>Classified</i>	<i>Classified</i>	<i>Classified</i>	<i>Classified</i>	<i>Classified</i>

Annex V: Sample Consent Form

We provide a sample Consent Form that users will have to accept before using ChildRescue mobile app. Many of the details of this form may be modified through the project, however the sample contains all the basic Sections that inform the Users regarding their rights and obligations when using ChildRescue. *ChildRescue Entity* refers to any organisation that deploys and uses ChildRescue for its operations. The list consists of the Smile of the Child, the Hellenic Red Cross and ChildFocus, but may include any sanctioned organism in the future. In this sense, the term *ChildRescue Entity* that appears in the sample consent form is meant to be replaced by the name of each organisation which uses ChildRescue and needs to distribute the consent form.

The *ChildRescue Entity*, identified as *We*, with which *You*, enters in agreement when using *ChildRescue*, processes Personal Data both as a Controller as defined in the and the EU Data Protection Directive GDPR.

Data Collected

You can use *ChildRescue* either anonymously or by registration. In both cases, and to ensure the correct operation of the ChildRescue platform, *ChildRescue* will collect data regarding Your:

- Current Location
- Network

When You register, you will be able to be more actively take part to ongoing missing children investigations. For this reason, a verification of Your email address or mobile phone number may take place before disclosing to You data regarding ongoing investigations. The registration and identification procedure will collect data regarding Your:

- Identification number
- Name
- Year of birth / age
- Gender

ChildRescue reserves the right to ask You to prove your identity in case You register.

Usage of Your Data by the Authorities

We inform You that when using *ChildRescue*, you will become a member of investigation by providing feedback to ongoing missing children cases. As such you agree that:

- The authorities can use Your data to come into contact with You, if the need arises.
- Data collected by Us, and which are relevant to an ongoing investigation may be handled to the authorities if requested.
- When You give feedback to an ongoing investigation that is officially carried out by the Police. Since feedback given by You can be used for present or future legal investigations, Your identification number, feedback and locations from every feedback You have given will be not deleted from Our databases even if You request so. Please refer to Article 23 of the GDPR.

Usage of Your Data by the Third Parties

We inform you that we may share Your data with other Humanitarian Organisations dealing with missing children so that to increase the efficiency of cross border investigations. Data will be shared in the same format as stored in our databases so that the same mechanisms of archiving and deletion can be applied to them by the third-party entities. We will ensure that the third party:

- Is based on a country that has adopted GDPR
- Is explicitly required by contract that it will enforce our data policy regarding Your personal data
- In case that You revoke Your consent, it will be notified within 2 working days regarding Your revocation and will be required to delete all of Your data from its premises within 5 working days after notification has been sent.

Your rights concerning your data

ChildRescue performs pseudonymisation and anonymisation on Your personal data as these are recommended by the GDPR. What this means is that:

- Your personal data are stored in a way that is unreadable except from the *ChildRescue Entity* technical personnel and the ChildRescue Application.
- The above process is implemented by following best practices in encryption and anonymisation to ensure that the data are securely stored
- Special data that can lead to the re-identification of Your personal data, are kept in separate database, which is referred to as the *Identification Database*, and can only be accessed by the *ChildRescue Entity* technical personnel.

You have the right to:

- At any time, request and retrieve information regarding how Your data are used and/or processed.
- Rectification of Your personal data in case You find inaccuracies and/or omissions.
- Erasure of Your data that are already being stored. Upon request, Your identification entries will be removed within 24 hours upon receipt of request from the *Identification Database* and You will not be able to be identified by the platform or by any *ChildRescue* operator any more. An exception is that of your identification number, feedback You provided and the locations from which You provided it as this is justified in Section *Usage of Your Data by the Authorities*. You can request that these Personal Data become *offline*, meaning that they will be stored in facilities without network access.
- Restrict the processing of your data. This right involves only Us; if authorities need Your data for processing, the *ChildRescue Entity* will provide the data.
- Download all your stored data in portable formats

Please note that any processing, profiling and decision making performed by ChildRescue, will generate suggestions that will have impact on ongoing investigations. You will not be the subject of any such process.

Why We process Your data

Collection and processing of Your data is necessary for *ChildRescue Entity* to perform its tasks that are necessary for the public interest. More specifically, collection of Your data is necessary for registering and/or identifying you when You use ChildRescue to view alerts and/or provide feedback/information (see also the *Data Collected* Section). Procession of Your data is needed to perform analytics on past and current cases; these analytics will provide prognostic functionality to the Entity and they will aid the Entity to future cases. In case You register with identification, Your personal data will also be processed in order to match You against the feedback/information You provided in order to evaluate Your reliability. Your personal data will be stored to the Entities' storage infrastructure. The Entity may share Your personal data with the authorities upon a legal request (please refer to the *Usage of Your Data by the Authorities Section*). For the process of deleting Your personal data upon Your request, please refer to *Your rights concerning Your data* Section. In any case, Your Personal Data will be anonymised, after 5 years of inactivity. Anonymisation means that the Entity will have no way of identifying You from the platform's data. The Entity however will still keep a minimum of Your Personal information; please refer to Section *Usage of Your Data by the Authorities*.

Cookie details

We inform You, that ChildRescue uses essential cookies for maintaining session information, performance cookies that collect non-identifiable user experience information and functional cookies to store Your personal preferences. The list of cookies that we use can be found in the following table:

Your obligations concerning the data of the ChildRescue application

- You are not allowed to keep any copy of the data of the closed cases. Any records related to the data or copy of the data of a case that You by Your own have exported in any format, should, also, be destroyed as soon as the case is closed and removed from the ChildRescue application.
- You are liable for any misuse of the application, even if You use ChildRescue anonymously. Deliberately providing false or misdirecting evidence, hindering ongoing investigations and disclosing sensitive information are a violation of terms and also constitute a legal offence. Performing the above actions, or any other criminal activity, will lead to your persecution to the full extent of the law.

[TABLE OF COOKIES SAMPLE ROW]

Provider	Name	Type	Expiration	Purpose
----------	------	------	------------	---------

ChildRescue	Session_id	Session information	5 days	Maintain information that keeps Your session valid through transactions with the platform
-------------	------------	---------------------	--------	---

Contact Details

For any feedback you can contact us at:

[CHILDRESUCE ENTITY ADDRESS DETAILS]

The *ChildRescue Entity* has a Data Protection Officer who is responsible for issues relating to privacy and data protection. You can reach the Data Protection Officer at:

[DATA PROTECTION OFFICER ADDRESS DETAILS]

Annex VI: Addressing the Ethics Requirements

The table below, lists the ethics requirements as derived from DoA, along with the conclusions from the EAB meetings, and with the current view of the ChildRescue on how these can be addressed.

Table Annex VI- 1 Ethics requirements list and how they will be addressed in ChildRescue

Ethics Requirements: D7.1 to D7.12. Points of review with subsequent adjustment sign off	EAB suggestions/comments and drawn conclusions from the Meetings with EAB	How it is addressed in D2.1, D2.2, D2.3	How it is addressed in D2.4
D7.1: GEN - Requirement No. 1 [12]			
A report by the Ethics Board must be submitted with the financial reports.	Meeting notes: Not yet finalised	Not relevant	Not relevant
D7.2: H - Requirement No. 2 [3]			
Templates of the informed consent forms and information sheet must be submitted.	Regarding Requirement No. 2, the language used to specify, correlates with the specified use of the need for the data collection and the use of the data obtained. The clarification of a recipient's role will make clear what their function and legal requirements are. This addition will draw together the project's role, expectations incumbent on recipient's as reflected in the relevant sections (2.2 and 2.3) of D2.3. Meeting notes: The reasons why ChildRescue processes	This is addressed in D2.3 section 3.3.5, as well as in points added in the sample Consent Form, in D2.3 "Annex II – Sample Consent Form".	This is addressed in section 2.3.3.5, as well as in points added in the sample Consent Form, in "Annex V – Sample Consent Form".

the acquired data are very generic. A couple of sentences need to be added in the consent form. ChildRescue needs to be clear of the categories of the recipients that it could send the data. Check the section where ChildRescue is described as a processor and/or controller of the data.

D7.3: H - Requirement No. 3 [3]

The applicant must clarify how consent/assent will be ensured in case children and/or adults unable to give informed consent are involved.

Regarding Requirement No. 3, very strong efforts (reflected on policy and a procedure) would need to be made in order to obtain the consent of a child or of the parent who holds legal parental authority or of their legal guardian or, where the legal framework is structured in this way, of the statutory child protection agency responsible for the care of the child. A humanitarian agency may fall outside this process.

- In extremis e.g. child protection issue/imminent risk of harm, data may be sought and shared.
- The threshold for 'public interest' is very broad and different across states. Further thought and discussion and advice within and from the EAB be initiated to look specifically at a mechanism, which is GDPR and protection compliant.
- The existing Missing children procedure, as detailed in the minutes, is strong, however may not be readily transferrable.

Clarified in D2.3 section 3.3.5

Clarified in D2.4 section 2.3.3.5

Meeting notes:

Some members of the EAB thought that it is not enough to get the consent from the district attorney instead of directly the child. This can only be done if there are legal grounds such as public interest, or humanitarian urgency, etc.

D7.4: H - Requirement No. 4 [3]

Details must be provided about the measures taken to prevent the risk of enhancing vulnerability/stigmatisation of individuals/groups.

Regarding Requirement No. 4, the risks of non-compliance should be included (examples to be specified e.g. 'screenshot' with sentence reflecting the list is not exhaustive) with a clear statement, requiring all data should be destroyed following the appropriate secure destruction method. It would be responsibility of ChildRescue to ensure this is clearly stated, however the responsibility to ensure destruction in all the situations outside ChildRescue core data collection, retention and destruction protocols, sits with those partners and or participating organisations who hold the data. This is also related to Requirement No. 7.

Meeting notes:

As long as ChildRescue can take this information back and there will not be any digital footprints, ChildRescue is fine. The notifications will be securely deleted from any device that were sent after they are not needed. Of course, the possibility of

All these are addressed in D2.3 sections 3.3.1, 3.3.3, & 3.3.4,

Also, more information has been added to the Consent Form, in D2.3 "Annex II – Sample Consent Form". The user is informed about the obligations he/she has to fulfil as soon as a case is closed and the relevant data are removed.

All these are addressed in D2.4 sections 2.3.3.1, 2.3.3.3, & 2.3.3.4,

Also, more information has been added to the sample Consent Form, in "Annex V – Sample Consent Form". The user is informed about the obligations he/she has to fulfil as soon as a case is closed and the relevant data are removed.

someone keeping a screenshot cannot be avoided.

D7.5: H - Requirement No. 5 [3]

Details on incidental findings policy must be provided.

Meeting notes:
It is not yet finalised

See "**Error! Not a valid result for table.**"

D7.6: H - Requirement No. 6 [3]

Copies of ethics approvals for the research with humans must be submitted.

Meeting notes:
It is not yet finalised

ChildRescue has contacted the responsible authorities and is awaiting a reply. The outcome of this communication will be reported in the interim report under WP6.

D7.7: POPD - Requirement No. 7 [3]

The applicant must explicitly confirm that the data used are publicly available. In case of data not publicly available, relevant authorisations must be provided.

Meeting notes:
ChildRescue must inform the recipients to whom ChildRescue has sent the data, in order they delete the data too. The only exemption to that is when this is impossible or it requires disproportionate effort (article 19 GDPR). ChildRescue should be ready to explain why it is impossible to do that.

This is addressed in D2.3 section 3.3.3. Also, relevant information has been added to the sample Consent Form, in D2.3 "Annex II – Sample Consent Form".

This is addressed in D2.4 section 2.3.3.3. Also, relevant information has been added to the sample Consent Form, in D2.4 "Annex V – Sample Consent Form".

D7.8: POPD - Requirement No. 8 [3]

Copies of opinion or the confirmation by the

Meeting notes:
It is not yet finalised

ChildRescue has contacted the responsible authorities and is

competent Institutional Data Protection Officer and/or authorisation or notification by the National Data Protection Authority must be submitted (which ever applies according to the Data Protection Directive (EC Directive 95/46, currently under revision, and the national law).

awaiting a reply. The outcome of this communication will be reported in the interim report under WP6.

D7.9: POPD - Requirement No. 9 [3]

Detailed information must be provided on the procedures that will be implemented for data collection, storage, protection, retention and destruction and confirmation that they comply with national and EU legislation.

Regarding Requirement No. 9, it is required to have a clear indication that the data subject can any time revoke their consent and that a relevant section will be added in the consent form. The process of what ChildRescue is going to do if someone asks for certain data to be erased, should be simply explained using terminology reflected in 2.2 with an additional chapter added to 2. Landscape analysis in privacy and anonymisation.

Information about the procedures that will be implemented for data collection, storage, protection, retention and destruction and confirmation (in alignment with EU legislation) are addressed in D2.3 section 3.3.3 & section 3.3 intro. All the handling of the data is also GDPR compliant.

Information about the procedures that will be implemented for data collection, storage, protection, retention and destruction and confirmation (in alignment with EU legislation) are addressed in D2.4 section 2.3.3.3 & section 2.3.3 intro. All the handling of the data is also GDPR compliant.

The project intends to develop a mobile application. Details on the privacy policy and permissions must be provided (in this case, it is important to make sure that app users are aware

Meeting notes:

- A clear indication that the data subject can any time revoke their consent will be added in the consent form.
- ChildRescue should have a good reason why specific data should be kept and it should be clearer about the data retention time period.

Addressed in D2.3 section 3.3.5.

Relevant section added in the Consent Form, in D2.3 "Annex II – Sample Consent Form".

Addressed in D2.4 section 2.3.3.5.

Relevant section added in the Consent Form, in D2.4 "Annex V – Sample Consent Form".

When we have a case reopen, it was agreed that for the ChildRescue platform it will be

that their data will be accessible by police authorities).

- ChildRescue should keep the data for as long it needs to in order to achieve its purpose.
- ChildRescue will describe a scenario where information will be kept available and used only if the case reopens.

regarded as a new case with a historical (parental) relationship with the old one.

The app also allows for real-time messaging among police forces and members of the community.

In order to avoid wrongful information to be circulated, the dynamics of real-time messaging must be explained.

In particular, it must be ensured that all information shared with the community is part of an ongoing investigation and has clearance to be disseminated among the public by a police or judicial body.

Explained in D2.3 section 3.1 & section 3.2.3

Any technical details of real-time messaging will be provided in WP3.

Explained in D2.4 section 2.3.1 & section 2.3.2.3

Any technical details of real-time messaging will be provided in WP3.

However, it should be made clear that there is no real-time messaging between the police forces and the community through the ChildRescue platform or app. Police force is not an eligible role for the ChildRescue platform.

Users will be informed (see "Annex V – Sample Consent Form" paragraph: "Usage of Your Data by the Authorities") that the data they provide could be disseminated among the public by a police or judicial body.

The applicant has provided draft terms of use of the platform that indicate, for instance, that "You [app user] are solely responsible for your conduct and any

The terms of use will be updated to reflect that the burden of the potential harmful uses of the platform must fall on the platform owner, and not on the user (who can be warned, but not be held

The terms of use will be updated to reflect that the burden of the potential harmful uses of the platform must fall on the platform owner, and not on the user (who can be warned, but not be held

<p>data, text, information, screen names, graphics, photos, profiles, audio and video clips, links ("Content") that you submit, post, and display on the ChildRescue service." This is unacceptable as all content would need to be reviewed before being shared. The burden of the potential harmful uses of the platform must fall on the developer/owner, not the user (who can be warned, but not be made responsible).</p>	<p>responsible). This will be implemented during WP3.</p>	<p>responsible). This will be implemented during WP3.</p>
<p>The app will allow for users to register anonymously. While their public identity can be anonymous, police forces should always be able to identify who is providing evidence related to an open case. Many countries forbid anonymous police reports, for instance. The legality and ethics of this point must be clarified.</p>	<p>Partially addressed in D2.3 in section 3.3.4.</p>	<p>Addressed in D2.4 in section 2.3.3.4 by adding the following 2 paragraphs: All data that are pseudonymised can be reverted to their original form. So, in case there is a legal police request, ChildRescue is able to provide the identity of the registered user. In order to identify an unregistered user, the police forces should contact the relevant mobile network providers since ChildRescue adheres to data minimization (Art. 5(c) GDPR Principles relating to processing of personal data). In case there is this</p>

requirement, ChildRescue could also gather further information, which will be clarified in WP3.

In any case, according to ChildRescue policy, police forces will always be allowed access to the platform in order to trace down criminal activities.

The legality and ethics of this point will be clarified in D1.2 "Regulatory framework for data protection, privacy, and ethical issues" in its updated versions.

Details on robust data deletion mechanisms (including meta data) once a user decided to leave the platform need to be provided.

Meeting notes:

- It is not clear what ChildRescue is going to do if someone asks for certain data to be erased. ChildRescue should be careful about the pseudonymisation and anonymisation. Pseudonymisation takes place when a case is closed. Anonymisation is equal to deletion and it happens after ChildRescue is certain that the case won't be reopened.
- ChildRescue needs to be clear towards the people giving the consent that they have the right to have the collected data deleted. Section 2.3.4 and everything that is related to the article 17 GDPR i.e. the right to erasure needs to be finetuned a little bit more.
- If the data are already in

Addressed in D2.3 in sections 3.3.1, 3.3.3 and 3.3.5.

In addition, it will be further elaborated in technical terms in WP3 where the implementation of the platform and mobile app are to be presented.

Addressed in D2.4 in sections 2.3.3.1, 2.3.3.3 and 2.3.3.5.

In addition, it will be further elaborated in technical terms in WP3 where the implementation of the platform and mobile app are to be presented.

pseudonymised form, and the data subject asks the deletion of the data, ChildRescue must either delete or anonymise the data.

D7.10: POPD - Requirement No. 10 [6]

The project intends to develop predictive algorithms. A detailed description of the data these will use and the assumptions they will make need to be independently assessed.

Meeting notes:

It is not possible to be covered at this phase of the project.

The detailed description of the data to be used can be found in section 3 of deliverable D2.2 in Table 3-1 and Table 3-2. Some important notes: First or Last Names and relevant identifiers are omitted, while address data contain only a general area at a zip code level. From recent discussions during the last plenary meeting, Social media accounts will not be provided explicitly, but instead the organisations will add manually any public info available concerning child's preferences and activities. Further anonymisation techniques maybe used to anonymise all the case data, if required.

The assumptions to be made from the analysis of this type of data have to do with the personality assessment of the child, the preliminary categorisation of the case, as well as the behavioural predictions associated to Points of Interest, through aggregated profiling (see section 2.1 of D2.2).

The detailed description of the data to be used can be found in section 3.3 of deliverable D2.4 in Table 3.3-1 and Table 3.3-2. Some important notes: First or Last Names and relevant identifiers are omitted, while address data contain only a general area at a zip code level. From recent discussions during the last plenary meeting, Social media accounts will not be provided explicitly, but instead the organisations will add manually any public info available concerning child's preferences and activities. Further anonymisation techniques maybe used to anonymise all the case data, if required.

The assumptions to be made from the analysis of this type of data have to do with the personality assessment of the child, the preliminary categorisation of the case, as well as the behavioural predictions associated to Points of Interest, through aggregated

profiling (see section 2.2 of D2.4).

D7.11: OEI - Requirement No. 11 [3]

The project will encourage citizens to contribute to an ongoing investigation, and thus to act as members of the police force. This has led to cases of mob justice or individuals taking matters into their own hands. The steps taken to ensure that those using the app are aware of what their role is, where it begins and where it ends, need to be provided.

Regarding Requirement No. 11, the process needs to be clearly detailed.

Meeting notes:

ChildRescue will provide extended information in several layers that start from the anonymous users, to the registered eponymous users, to the validated volunteers, and the members of the administrative team of the NGOs that handle the cases. There will be several levels of access.

It is fine to provide different information depending on the level of identification. The volunteer organisations will validate the users through a know your customer procedure.

Although citizens provide one-way information and thus contribute to an ongoing investigation, they are not actively participating in the search and rescue procedures, nor in the actual investigation itself. This should be made clear to them by having the app provide this information on first usage (e.g. through a tutorial).

On the other hand, volunteers or rescue team members who participate in the investigation procedures are verified by the organizations in person.

Guidelines with instructions on how to act in situations when encountering a missing child or some piece of evidence will be also provided in the mobile app.

The user roles and related processes will be described in WP3 where all the specifications of the ChildRescue Platform will be defined.

D7.12: OEI - Requirement No. 12 [3]

The app may be misused by people who intend to locate missing children to abuse them. Details on how this will be avoided must be provided.

Meeting notes:

ChildRescue won't provide more information than it is already disseminated with for instance, Amber Alert, but there is always the possibility of misuse of the platform, like it may happen with an Amber Alert. ChildRescue will follow the common practice.

Even though this will be explicitly defined in WP3, it is by now clear to the ChildRescue partners that the ChildRescue platform will share less information for each case than the information already disseminated from the organisations' websites and related social media or mass media.

In addition, several points on the Consent form and the terms of use of the app will make potential abusers more reluctant to use it.

Annex VII: Incidental Findings Policy

Incidental findings are defined as results that arise that are outside the original purpose for which the test or procedure was conducted. In the ChildRescue project, in an unlikely event of any incidental finding raising ethical concerns discovered throughout the execution of the project, the following policy will apply:

(1) In a first place, this incidental finding will be immediately referred to:

- the Project’s Coordinator
- the Ethics Advisory Board (EAB) and
- the European Commission, in its funding agency capacity, via the project officer

with the view to evaluate their ethical implications and to reach a decision on further action. Here the incidental findings policy reflects on the ethical complications stemming from new forms of risk, threats and vulnerabilities and the multiple meanings and normative implications of emerging surveillance technologies.

(2) Secondly, the following rules will govern any incidental findings:

- Deletion of any incidental findings will be considered by the bodies mentioned in (1);
- In case of incidental findings that include recording an illegal activity, the consortium will comply with all relevant local and international laws;
- In case of incidental findings that include any information of public interest, the bodies mentioned in (1) would decide on the need, means and timing of their communication to relevant stakeholders;

This policy might be revised and adjusted throughout the lifetime of the ChildRescue project, should these prove necessary.

Annex VIII: Ethics Advisory Board report for D2.4

The current section provides the outcomes that derived from the Ethics Advisory Board teleconference.

VIII.1 The ChildRescue Ethics Advisory Board D2.3 Teleconference

Table Annex VIII- 1 The ChildRescue Ethics Advisory Board (names and position).

Name	Position
Philip Ishola (EAB Chair) (PI)	LOVE-146 international in human rights and technology
Karen Shalev Greene (KSG)	Director of the Centre for the Study of Missing Persons
Eva Lievens (EL)	Professor Specialised in Human rights and technology
Spiros Salamastrakis (SS)	Lawyer
Prof. Dr. Andreas Jud (AJ)	Professor
Peter Van Dyck (PVD)	Lawyer
Thanasis Giannetsos (TG)	Lecturer in Secure Systems
Antoine Bon (AB)	Legal advisor

The Board was sent, before the meeting, the Table Annex VI- 1 without the last column, that lists the ethics requirements as derived from DoA, along with the conclusions from the previous EAB meetings, and with the current view of the ChildRescue on how these can be addressed, in order to review it, and make their comments. The EAB meeting took place on Thursday 20th of December 2018, at 12.00 to 14.00 CEST. Members of the Board that were attending were: Philip Ishola, Spiros Salamastrakis, Thanasis Giannetsos. From NTUA, the meeting was hosted by Christos Ntanos (CN), Ariadni Michalitsi, and Dimitris Varoutas who kept the minutes and took notes during the meeting. Minas Pertselakis (MP) from S5, Danae Vergeti and Dimitris Ntalaperas from UBITECH, Panagiotis Dragatis from REDCROSS also attended the teleconference. In this report the meeting minutes and the conclusions from the EAB teleconference are given.

VIII.2 Meeting minutes

In this section, the meeting minutes are quoted.

At first, CN made a brief introduction for the scope of the teleconference. Based on the schedule of the project, ChildRescue has to address till the end of 2018, the ethics requirements that can be addressed till then. The deliverable that will be based on the report of this teleconference, will describe if ChildRescue has managed to do what was required by the EC on the subject of ethics. The list of the requirements is the same as in the previous teleconferences, but the way they are addressed is updated. The teleconference's goal is to see if with the work and the comments provided by the EAB, the ethics are addressed.

Then, PI asked CN to give a brief update about the Frankfurt meeting.

CN gave a brief update about the Frankfurt meeting. He said that the main scope of the meeting in Frankfurt was to decide which will be the design requirements and the architecture of the system. The technical requirements were put in mock-ups. Additionally, various screens of the application were shown. There was also a discussion about the language that will be used, the information that will be shown, how to respect the restrictions and suggestions from the EAB, etc. Part of the discussion was if the application could be used to send other types of messages other than alerts (it will not), or if the area where a missing child is being searched will be shown (no). ChildRescue will send the alert to the users without revealing the area that the alert has been sent. Volunteers won't be registered through the ChildRescue platform, but the registrations will be made only through the authorised organisations.

Based on the previous reports from the EAB and the results from D2.1, D2.2, D2.3, rather than just aggregating these to D2.4, ChildRescue consortium took the chance to augment them and take care of the remaining issues on the ethics requirements.

Requirement No. 2

PI read the requirement. He asked if there were any comments on the requirement. There weren't any, so the discussion continued to the next requirement.

Requirement No. 3

PI read the requirement. He then read the comments made by AJ: "that the consent from the district attorney instead of the child directly, can only be done if there are legal grounds such as public interest, or humanitarian urgency, etc". Then he asked if there were any more comments on the requirement. There weren't any, so PI said that the requirement is addressed in D2.4 section 2.3.3.5.

Requirement No. 4

PI said that this requirement is related to the incidental finding policy. Then, he read the requirement. He asked CN about how the requirement is related with the incidental finding policy.

CN responded that the incidental finding policy is a policy, where ChildRescue describes what happens if something is discovered by accident during the project, through the data of the project.

Then, PI said that the process is pretty robust and he asked if there were any comments. There weren't any, so the discussion continued to the next requirement.

Requirement No. 5

It is addressed in

Annex VII: Incidental Findings Policy.

Requirement No. 6

PI read the requirement and he asked CN to give some more information.

CN said that there is a European board of research ethics. There is one member of the board per country. He contacted with the members in Greece and Belgium. He asked them, based on the requirements and the procedures ChildRescue follows, what is that ChildRescue has to do, whom ChildRescue has to contact, what process ChildRescue needs to follow. The Greek member said that there is not such a board for this kind of projects in Greece. The Belgian member said that he doesn't know such boards, and that would ask others. CN is waiting for his response. It became apparent that at least for non-medical research on humans, there is not such approval procedure in these two countries. If there is going to be any kind of expansion of the research side of the project to other countries, ChildRescue should contact the member of these countries.

PI said that this was very useful. He asked if there were any comments. There weren't any, so the discussion continued to the next requirement.

Requirement No. 7

PI read the requirement and asked if there were any disagreements on how the requirement is being addressed. There weren't any, so the discussion continued to the next requirement.

Requirement No. 8

PI said that it is in the same state like the requirement No. 6.

Requirement No. 9

First part

PI read the requirement and how the requirement is addressed. There were no comments from the rest of the board. He said that this was also discussed at the last meeting. He also said that the requirement is addressed and that is reflected in the particular sections that are mentioned.

Second part

PI read the requirement and the sections that it is addressed. He said that some amendments were made to the consent form, so he is confident that this requirement is also addressed as indicated.

Third part

PI read the requirement and where it is addressed. He asked CN to clarify what "no real-time messaging between the police forces and the community" means, and what is the relevance with the requirement.

CN said that the perception of the people who made requirements was that there would be a mechanism by which the general public would coordinate among themselves to search for the missing child. That perception led the ChildRescue consortium to take action against the probability of mob justice and issues like sharing personal identifiable information, especially from children who are in danger. If you have everyone take part in the investigation, then the general public could take pictures or share addresses and information among themselves. But in ChildRescue implementation, there is no such coordination. There is only coordination among authorised search and rescue teams and volunteer networks that are already registered in the existing organisations. Furthermore, the messages can be sent among the groups and moderated by the team leaders. On the other hand, the

kind of messaging that exists for the general public is only one way. Citizens cannot communicate with other citizens. There is no same level messaging for the general public. They can only send messages to the person responsible for handling these messages. The case manager will be able to broadcast an alert to a lot of people, but the people that receive the alert can only communicate with a person higher in the hierarchy.

Then, PI said that this was a very important aspect to focus on. There weren't any questions or observations on the requirement. So, the explanation in the described sections is acceptable and addresses that particular requirement.

Fourth part

PI said that this is related to WP3, so it is not relevant at this point. MP said that the Terms of Use (TU) will be implemented during WP3.

CN added that the TU are being done right before publishing an application. They refer to specific actions and events that happen within the application. The TU will reach a certain level of detail and will be complete when the application will be available to the general public. So, the TU will be published right before the application will be publicly available.

MP, said that ChildRescue needs to acknowledge that there was a mistake in the DoA, as the comment of the EAB referred to. So, in this requirement, this mistake is being addressed. The exact text for the TU will be implemented in WP3. He also added that the burden of potentially harmful uses will be on the platform owner, not the user.

CN said that it was written that the users of an application could not be held responsible for their own actions on the platform, something which is clearly wrong. He said that people are responsible for their actions, but the platform has some obligations not to endanger people by giving tools that may harm other people. CN also added that the changes in the template of the TU, have already been done before the project started and the EAB has already accepted them. The final TU will be complete when the platform gets published.

PI asked if everyone agreed with that scenario.

SS agreed.

Fifth part

PI read the requirement and he said that there was a lot of discussion at the EAB meetings. He said that it is addressed based on the sections that are mentioned. He asked his colleagues if they wanted to add something or give some advice.

SS asked if the police should need the permission of a prosecutor in order to have access to the person who has anonymously provided information.

PI answered affirmatively. That would need to happen. If there is an active police investigation, the prosecutor or the judge should initiate some kind of investigation which would allow for the approach to individuals to provide their information.

CN said that since ChildRescue operates within a legal framework of a country in the EU, it has to recognise the existence of the authorities. That is stated on how the requirement is addressed. ChildRescue cannot deny any kind of information request by the authorities, because it doesn't have the authority to deny it. ChildRescue is not in a position to describe the way the authorities choose to

make the requests and the check for legality for the actions made by the authorities fall within those organisations that enforce it.

PI said that it was quite clear. So, the information needs to be shared within the legal framework of the request.

Sixth part

At first, PI read the requirement. This has already been discussed in the previous EAB meetings. It is indicated that it is addressed in the mentioned sections, but it will be further elaborated in technical terms during WP3 when the implementation of the platform and the app are to be presented.

SS agreed on how this is currently addressed.

Requirement No. 10

PI said that AJ commented on this. He suggested that the usage (not only find an individual child but also to develop algorithms) should be referenced in the consent form. PI said that the consent form should mention the potential usage of data for data analytics.

MP made clear that the data that are mentioned in this requirement are children's data, not the users' (of the platform or the application) data.

PI asked if the consent form links to children's or the responsible adults on behalf of children.

CN said that the children fall within the category of the people unable to give consent. A missing child is not someone who would fill in a consent form. With that in mind, there are consent forms to share information with the organisations that already use that data. The consent form in this case, is for the use of personal identifiable information for research. If the case has been resolved with the child being found in any way, the people who can give consent for the use of the data for research are the parents of that child. So, after the case has ended, ChildRescue could request the parents to allow ChildRescue, as a research project and not as a platform that aids in the investigation of children, to take that data and use them for research purposes. He said that ChildRescue removed the personal identifiable information. So, after the data is anonymised (in order to be compliant with GDPR), ChildRescue can use data for research purposes. Thus, ChildRescue doesn't require a consent form as a research consortium for the use of that data in research.

MP added that ChildRescue as a platform won't have direct contact with the parents. It will only have direct contact with the organisations.

CN said that ChildRescue will never request personal identifiable information as a consortium from the pilots' members of the consortium or anyone else.

Then, PI said that this description clarified many things. He wondered if there is an option about the usage of the data. He asked if it is likely that the information provided to the platform and the organisations would be used for a different purpose than initially agreed.

CN said that this exists in the consortium agreement. It states several things about the information sharing, such as the non-disclosure information managed by a specific organisation in the consortium. ChildRescue cannot take data which are directly managed by an organisation and used for other purposes. All the terms about data sharing and information sharing within the consortium, which includes the pilot organisations, is already included in the consortium agreement.

PI said that it was helpful to knowing that and the discussion moved on to the next requirement.

Requirement No. 11

PI read the requirement. He also talked about how it is addressed and that the first paragraph clarifies the position about the level of risk. The second paragraph clearly identifies the organisations and their responsibilities in the process for verification of the individuals. PI asked CN to give clarifications about the third paragraph.

CN said that it is already agreed in the design of the application. The instructions will be provided in the application when the person either selects the specific link from the menu or he provides feedback through the application. The text that will be displayed will be the one which is selected by each of the organisations that are operating within Missing Children Europe Network.

Then, PI asked CN about the data recording forms and if they are almost complete.

CN answered that they are almost complete.

CN said to MP that PI refers to the Excel file that will be used for the profiling part. He asked MP where is that table.

MP responded that the table is in D2.2 and in D2.4 (section 3.3).

CN said that ChildRescue developed this template, which includes the importance of each field and the category of usage per field. ChildRescue will not collect anything else that is not within the accepted usage in profiling as it has already been described in the previous deliverables. ChildRescue will only gather data that it considers potentially useful within the methodologies that it developed.

MP added that out of 120 attributes, ChildRescue selected 40, based on several criteria, e.g. relevance with the data analytics algorithms, which attributes are already collected from the pilots, etc.

PI said that he is satisfied with the description, so the discussion moved on to the next requirement.

Requirement No. 12

PI read the requirement and how it is addressed. He wondered how the consent form and the terms of use will make potential abusers more reluctant to use ChildRescue.

MP, answered that a paragraph was added in the consent form. It is about obligations concerning the data of the ChildRescue application. He read this paragraph: "You are liable for any misuse of the application even if you use ChildRescue anonymously. Deliberately providing false or misdirecting evidence, hindering ongoing investigations and disclosing sensitive information are a violation of terms and also constitute a legal offense. Performing the above actions or any other criminal activity will lead to your persecution to the full extent of the law".

SS said that the above description doesn't explain the "less information" that is has been written in the first paragraph.

CN added that it explained only the second paragraph. Concerning the first paragraph, ChildRescue, when designing the platform, restricted any kind of information which would give any further clues specific to the case. For example, when Amber Alert is issued, it is broadcasted to train stations, highways, posters, flyers, internet posts, social media, etc. Today, there is the possibility of someone keeping the flyer, etc. due to the way the dissemination of material is being handled. ChildRescue, by restricting the area in which the alert is sent, it restricts the amount of information which reaches the

general public. By paying close attention not to include more information than the original Amber Alert, ChildRescue will restrict the flow of information geographically.

SS said that this description covers his question.

CN also added that ChildRescue can withdraw an alert the moment it wants to withdraw it, but it cannot track any further use of any kind of information, that could be misused, in the sense that it cannot restrict someone takes a photograph or a screenshot. That is a police matter. ChildRescue cannot either track it and enforce.

PI said that it was very clear.

VIII.3 Conclusions

PI said that EAB reached the final point of consideration. Then, he asked CN if ChildRescue has everything it needs from the EAB in order to form this report.

CN said that the next EAB teleconference will take place asked in the first 10 days of June 2019 and it will be linked to the interim review. He said that he wants to close everything that has to do with ethics requirements, up to this point, except the open issues (authorisations and the terms of use). So, he asked within the requirements and the timeframe of the first year of the project, if there is a consideration from EAB members that are pending actions that have to be done by the end of this year.

PI said that except some references regarding WP3, ChildRescue covered the areas needing addressing. The references to specific requirements have been reflected in D2.4, throughout the document.