SPECIAL

Scalable Policy-awarE Linked Data arChitecture for prIvacy, trAnsparency and compLiance

Deliverable D1.1

Use case scenarios V1

V1.0

SPECIAL DELIVERABLE

Name, title and organisation of the scientific representative of the project's coordinator:

Mr Philippe Rohou t: +33 4 97 15 53 06 f: +33 4 92 38 78 22 e: philippe.rohou@ercim.eu

GEIE ERCIM, 2004, route des Lucioles, Sophia Antipolis, 06410 Biot, France

Project website address: http://www.specialprivacy.eu/

Project	
Grant Agreement number	731601
Project acronym:	SPECIAL
Project title:	Scalable Policy-awarE Linked Data arChitecture for prIvacy, trAnsparency and compLiance
Funding Scheme:	Research & Innovation Action (RIA)
Date of latest version of DoW against which the assessment will be made:	17/10/2016
Document	
Period covered:	M1-M14
Deliverable number:	D1.1
Deliverable title	Use case scenarios V1
Contractual Date of Delivery:	31/05/2017
Actual Date of Delivery:	31/05/2017
Editor (s):	Piero Bonati, Sabrina Kirrane, Rigo Wenning
Reviewer (s):	Piero Bonati, Sabrina Kirrane, Rigo Wenning
Participant(s):	P.A. Bonatti, J. Colbeck, F. De Meersman, R. Jacob, S. Kirrane, M. Kurze, M. Piekarska, R. Wenning, B. Whittam-Smith, H. Zwingelberg, E. Schlehahn
Work package no.:	1
Work package title:	Use Cases and Requirements
Work package leader:	CeRICT
Distribution:	PU
Version/Revision:	1.0
Draft/Final:	Final
Total number of pages (including cover):	48

Disclaimer

This document contains description of the SPECIAL project work and findings.

The authors of this document have taken any available measure in order for its content to be accurate, consistent and lawful. However, neither the project consortium as a whole nor the individual partners that implicitly or explicitly participated in the creation and publication of this document hold any responsibility for actions that might occur as a result of using its content.

This publication has been produced with the assistance of the European Union. The content of this publication is the sole responsibility of the SPECIAL consortium and can in no way be taken to reflect the views of the European Union.

The European Union is established in accordance with the Treaty on European Union (Maastricht). There are currently 28 Member States of the Union. It is based on the European Communities and the Member States cooperation in the fields of Common Foreign and Security Policy and Justice and Home Affairs. The five main institutions of the European Union are the European Parliament, the Council of Ministers, the European Commission, the Court of Justice and the Court of Auditors (http://europa.eu/).

SPECIAL has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731601.

Contents

I Recommendation system (PROX) 8
1 General pilot description . . . 9
1.1 Personal interest profiles . . . 9
1.2 The event database . . . 11
1.3 Customer communication . . . 11
2 Data usage policy and consent requests . . . 12
3 Main goals of the pilot . . . 15
3.1 Assumptions . . . 16
3.2 Open Questions . . . 16

II Reusing telephony data for traffic alerts, optimising road layout, and targeted marketing (TLABS) 17
4 General pilot description . . . 18
4.1 Location Based Services . . . 18
4.2 Use case scenarios . . . 20
5 Main goals of the investigation . . . 20
6 Data usage policy and consent requests . . . 21

III Know-your-customer reports (TR) 24
7 General pilot description . . . 25
7.1 KYC data collection . . . 26
7.2 KYC Data Storage . . . 28
7.3 KYC Data Processing . . . 29
7.4 KYC Data Sharing . . . 30
8 Data usage policy and consent requests . . . 30
8.1 Purpose of data collection and processing . . . 30
8.2 Collected data . . . 31
8.3 Where are collected data stored . . . 31
8.4 How long for . . . 31
8.5 How is data processed and shared . . . 31
8.6 Consent requests . . . 32
8.7 Open questions . . . 32
9 Main goals of the pilot . . . 32

IV Discussion **33**
10 A unified view of the location based services use cases 34
10.1 A simple location based service 34
10.2 Consent for the LBS case 35
10.3 Service delivery 36
11 Policy features occurring in the pilots 37
11.1 Personal information categories 37
11.2 Formal policy statements and related challenges 38
11.3 Enabling consent, transparency and compliance 39
12 Legal challenges 40
12.1 Challenges regarding anonymisation 41
12.2 Consent and its inherent challenges 43

List of Figures

1 Technical diagram 14
2 KPIs 15
3 KYC Process Flow 25
4 A simple location based service 35
5 Location based scenario with consent/control 36

Introduction

This document describes the three pilots of SPECIAL, that have been conceived by partners PROX, TLABS, and TR. The three pilots are one of the means by which SPECIAL keeps in touch with real problems and concrete industrial needs; as such they are a source of requirements and research questions, encompassing both technical and legal aspects. Each pilot has been refined and analysed during three face to face meetings with the respective pilot leader – that took place between February 28 and March 21 2017 – and during the face to face meeting in Nice (May 8-10 2017), plus several ad hoc teleconferences occured in the first five months of the project. This first version of the use case description is primarily aimed at a general description of the pilots and a more detailed account of their data usage policies and informed consent request, that are needed for collecting legal requirements and designing the policy language (deliverables D1.2 and D1.3). The other relevant details, needed for eliciting technological requirements, identifying the interfaces between SPECIAL's horizontal components and the pilot-specific components, and designing the user interfaces, will be the subject of the second version of the use case descriptions (D1.5). D1.1 is organised as follows: In each of the next three sections, we describe a pilot together with its data usage policy. In a final section we collect and summarise the main features of the policies introduced before, which will constitute the pilot-related input to D1.3. Furthermore, we will point out some legal challenges stemming from the pilots' data usage policies and related consent requests. This will constitute one of the inputs for D1.2.

Part I

Recommendation system (PROX)

1 General pilot description

Proximus would like to build a geo-fenced (i.e. geographically restricted) tourist recommendation tool for: (i) Proximus customers (ii) that are Belgian residents and (iii) that visit the Belgian coastal area as tourists[1]

Although the pilot will focus on Belgian tourists in Belgium, the broader vision is to eventually offer the service to roaming tourists. In such a scenario, the initial roaming message could potentially be used by Proximus to ask for consent without additional intrusion and spam. Additionally, it is possible that even Belgians living in the area could benefit from such a tool.

The goal of the tool is to recommend tourist events that match with the users personal interests and to direct tourists to events that they might like. Such a tool also promotes tourism in the region by creating traffic to events that without the tool may go unnoticed. Nice-to-have functionality includes: the ability to request other/extra recommendations; for users to be able to add the event to their calendar; and to enable tourists to give a score on how much they liked the event upon participation.

The overarching aim is to enable Proximus to create statistical insight from the generated profiles and to investigate if Proximus clients derive value from such profiles.

Considering that the usecase requires the processing of both personal and location data, the legal requirements will be derived from both the General Data Protection Regulation and the ePrivacy Regulation.

1.1 Personal interest profiles

A personal interest profile will be created for each of the Proximus customers that participate in the pilot. This section provides details of the data sources that will be consulted in order to generate personal interest profiles and describes the transparency and control that will be provided to the customer.

1.1.1 Data Sources

The profile will be constructed by using telephone records, mobile telephony location data, television viewing data and Internet surfing behaviour[2]

Call detail records (CDRs): CDRs reveal which people communicate with each other (i.e. are making voice calls or sending SMSs to each other). CDRs will be used for discovering family relationships and for performing family-centred Social Network Analysis (SNA). The characteristics of this communication (frequency, volume and timing) allow the personal or professional nature of the communication to be derived and hence to infer whether people have family ties (and cliques). The underlying reason is that most tourist activities take place in a family or social (clique) context. Also, it is possible within each clique to derive the most *consulted* (and potentially the most influential) person, which could be the one targeted by

[1] In order to qualify as a tourist visit a number of conditions have to be fulfilled. The Belgian coastal area is not entirely a tourist area and thus the nature of the visit cannot be automatically derived from geographical data that is visited.

[2] Some of these concepts need more background information. Therefore some documents providing background information will be made available.

the recommendation tool (of course this person must be one of the customers that opted-in).

Mobile telephony location data (network data): Location data records will be used for identifying the customers that are in the recommended area, previous visits to the observed tourist areas, the *most likely living place* of the customer and potentially a *second residence*.

Television viewing (Audio-Visual Data Records (AVRs)): Television data will be used for determining interest profiles. Since not all mobile telephony customers are also Internet Protocol television (IPTV) customers, the customers that can participate in the pilot must be users of both products. AVRs will be used for determining interest profiles in the following way: One AVR is created for every two minutes that a viewer watches TV. For example, for a person that watches one program for 60 minutes (without zapping) thirty AVRs will be created, while someone who is zapping at least once in every two minute interval, will not generate any AVRs. AVRs will be linked to actual TV programs by using the Electronic Program Guide (EPG). The EPG has a summary per program that characterises a TV program (film, documentary, sports,..) using several keywords. These keywords will be used to define TV viewing profiles.

Internet surfing: Top Uniform Resource Locator (URL) records will also be used to generate customer interest profile data. Here top URL records refer to the domain names on the Internet that have been visited, not the (sub)pages themselves that have been visited and not the search keywords.

By default the profile will **NOT** include information in relation to films watched via video on demand, visits to adult websites, and sensitive information and keywords (e.g. sexual orientation, religion, political convictions).

1.1.2 Data Sharing

The question about sharing is more or less the question about the benefit for Proximus. Neither the data collected nor the profiles will be shared with others. However, Proximus could potentially use/share statistical data such as aggregated movement data with 3rd parties in order to improve tourist offerings at the Belgian coast.

1.1.3 Managing the personal interest profile

Profiles are sets of keywords and not predefined profiles. A profile can be for example *Aeroplanes, golden retriever, Champions league, fitness, DIY home improvement* and associated metadata (e.g. that a customer spends eighty percent of the time watching premium channels or that a particular keyword is linked to a particular TV channel). Metadata may also be used to infer information about education, skills and competences.

The customer will be given full control over their personal interest profile (e.g. a data subject can both view and delete keywords from the profile).

All customers participating in such a pilot will need to complete a privacy declaration. The aim of the privacy declaration is to ensure that the user is fully aware of the data that will be used by the recommendation tool and where the data comes from.

However, standard privacy declarations are lengthy documents containing page after page of legalese text that explains data processing and sharing often in a manner that the user does not understand. As such, Proximus are particularly interested in looking into the possible alternatives.

The principle of informed consent is to ensure that the customer is in the driving seat (i.e. the more open a person is regarding their profile the less self-censorship the tool will apply to the recommendations). If a person, for example, is willing to share that he likes naturist beaches, the recommendation tool will include related recommendations. The consent should be dynamic in the sense that the customer should be able to both add and revoke consent all times through an online form (authentication is therefore needed). The research challenges with respect to informed consent are described in *Section* 12.

1.2 The event database

The event database is a database of events that take place during the summer holidays season together with event keywords and metadata. Individual recommendations will be derived by matching customer profiles against event keywords, thus ensuring that only relevant recommendations are sent to the customer. It is envisaged that recommendations will be sent to data subjects one week in advance with a possible reminder three days before the event takes place. The number of recommendations sent is limited to three recommendations per week unless the customer is spending their holidays at the Belgian coast, then they get a daily recommendation.

The event database will be populated by Westtoer, a public non-commercial organisation. Events organised by commercial organisations will not be included in the event list. No information or data will be shared with external parties, all processing takes place on Proximus servers (either backend IT or cloud).

1.3 Customer communication

This section provides an overview of envisaged customer communications including an invitation to sign up for the service, the consent form that will be presented to the customer and details regarding the mode of communication. Open questions include, how do you effectively present the *consent form* on a mobile phone? How do we move from a traditional paper based approach to one that can be executed remotely.

1.3.1 Marketing of the Pilot

Broadly speaking the following template will be used to invite customers to participate in the Pilot. It is worth noting that Proximus are also considering using some sort of reward to encourage customers to trial the Pilot, if so this information will also be included in the initial communication.

Dear customer,

With your agreement and based on your location and your interests, we want to suggest cultural and other events to you. We will intelligently derive your interests as you go, but give you full control to alter or erase your profile data. Of course you can opt-out at any time.

To subscribe send yes to 3615-tourist

[Proximus signature]

1.3.2 Recommendations

Once the profile is rich enough to determine relevant recommendations, Proximus will match the profile with the event database in order to determine relevant events. Those will be communicated via the most appropriate channel. If the user is on the road, Proximus will send the information via SMS. The SMS may contain more information that allows the user to access a web interface with the recommendations. This automatic channel determination will be the default. A user interface will allow the user of the service to stop all notifications, or to select a specific channel only. In case all notifications are stopped, the service turns from a push service to a pull service where the user can find the recommendations in a specific place, but will never be notified of new additions to the list. The SMS sending will be done by the Proximus enco.io platform by use of its SMS Application Program Interface (API). The email sending will be done by an email server set up in the Proximus environment.

2 Data usage policy and consent requests

According to the General Data Protection Regulation (GDPR)[23], informed consent request shall specify clearly which data are collected and processed, and for which purpose, in order to make the data subjects' consent legally valid. The consent requests must specify (at least) the following features, which characterise also the data usage policy that shall be enforced by the company. We include also a description of how data subjects can exercise their rights to control their data, since in this pilot the control modality is particularly direct and fine-grained.

One of the primary research challenges for SPECIAL is obtaining contextualised consent. For example, we need the flexibility to add a new data source, that will be used to generate a profile, based on a suggestion that the user accepts. One approach would be to start off with something like location only and progressively ask for consent for additional data sources and/or categories of data.

Purpose of data collection and processing

The purpose of the data collection is to constitute profiles of the data subject representing their interests. Those interests are matched against an event database to determine relevant events. Additional data is collected so that the data subject can use the control

interface and edit the profile data. The data subject's personal data is used to communicate the relevant events found via an automatic determination of the most appropriate communication channel.

Collected data

The data categories collected and processed for the above purpose are:

1. *Mobile telephony location data*, collected from the antennas to which telephones connect. Note that these data are collected in the normal course of operations of the company.
2. *Internet TV programs viewed*; this information is collected from set-top boxes. Note that video on demand and adult content are *not* collected.
3. *Visited web sites*; this information is collected from the company's devices that provide the internet connection. Only domain names are collected; domains with adult content are not stored.

Items 1, 2 and 3 are personal data that can be collected only after the data subjects give their informed consent.

How is data processed

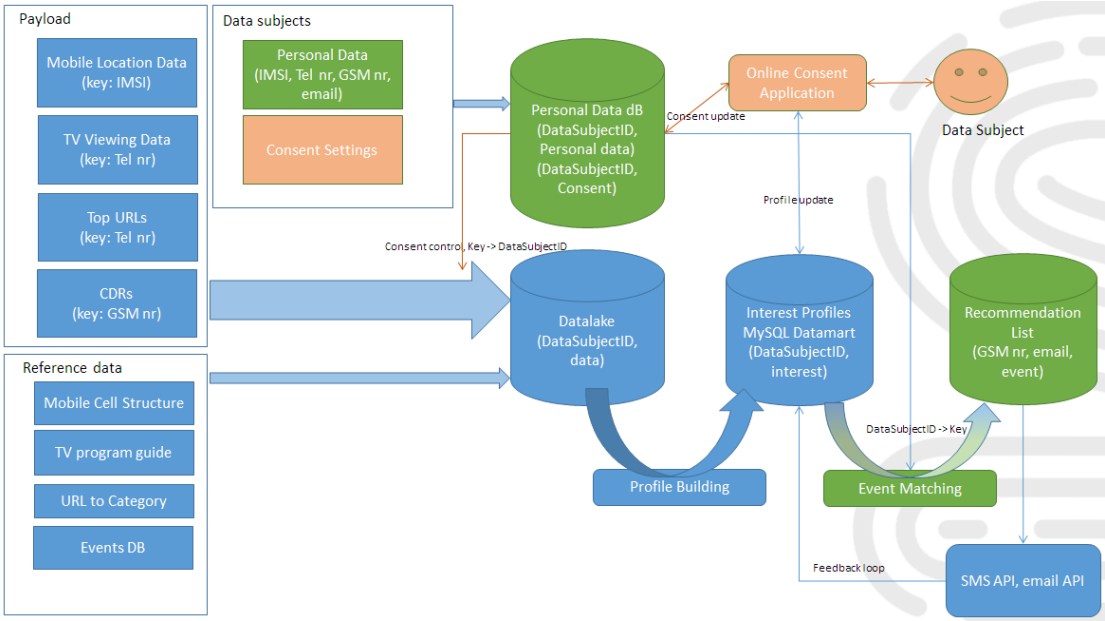
- The collected data are analysed to extract an interest profile for the data subject.
  - The interest profile is a set of keywords, not referring to sexual, political or religious orientation. The profile indicates also whether the user is interested in the coastal area.
  - The keywords are extrapolated from the visited web sites' classification and from the TV guide (i.e. the description of the selected programs).
  - Interest in the coastal area is detected from the above data *and* from location data.
- Interest profiles are matched against the events stored in the "Westtoer Datahub" database and are used to form tourist recommendations.
- Tourist recommendations are sent to the user via SMS and email.

The process of collecting and processing as well as matching and consent interaction is displayed in the following diagram:

Where are collected data and profiles stored

All collected and derived data are stored in the company's servers, which are located in Belgium.

Figure 1: Technical diagram



For how long are the data stored

Payload data is only stored for the time needed for processing.

1. TV viewing data (AVRs): As soon as the keywords derived from a TV program that were mentioned in an AVR are extracted, the related AVR will be deleted as it no longer needed.

2. Location data: As soon as 'most likely living place', 'second residence' and other derived aggregated variables are calculated, the location data will be deleted. In this particular case, data of up to 6 weeks can be needed to calculate the derived variables.

3. Visited web site data: Similar to TV viewing data, as soon as the domain name is categorised (translated to a list of keywords), the visited domain name will be deleted.

Disclosure to third parties

All data is processed within the company and there is no disclosure outside the company.

Control mechanisms for data subjects

Data subjects can access their interest profile and modify it, by adding and deleting data sources and keywords.

3 Main goals of the pilot

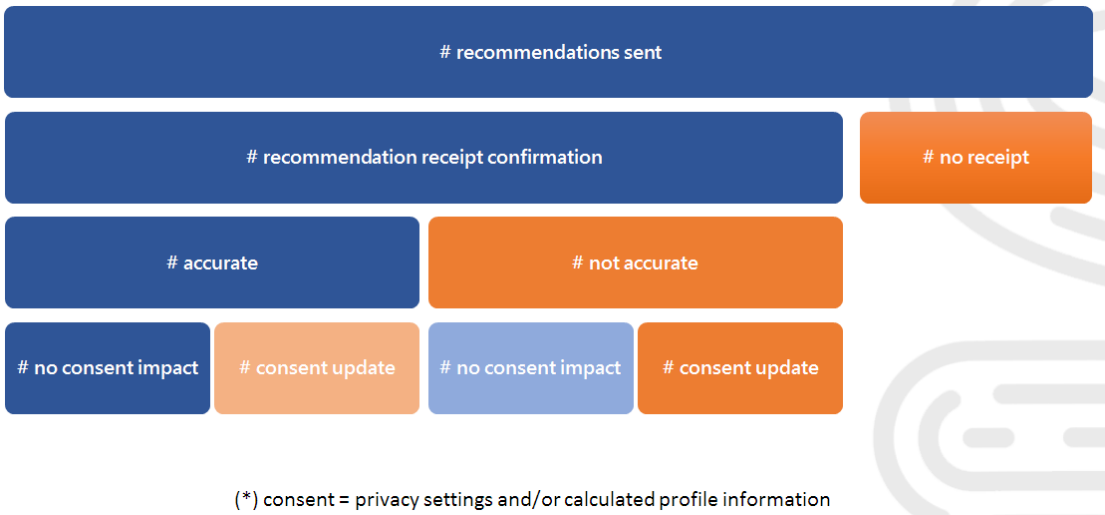
The pilot is strictly non-commercial and will be carried out with customers who voluntarily participate in the pilot. These can be Proximus employees, consultants or friends/family of the former. The target size of the test group is between 500 and 1000 participants. The required duration of the pilot period is to be determined. Proximus will first test the invitation to participate with 10 persons, then a group of 50, and finally the remaining group.

The pilot will be designed to verify the following:

- Customers received the recommendation;
- Customers acknowledge the recommendation to be accurate and useful;
- Customers act upon the recommendation, positively by attending the event, or negatively by changing privacy settings.

The following diagram gives an idea of the type of KPIs being looked for:

Figure 2: KPIs



The general objectives of the Proximus pilot are:

- To test the responsiveness of data subjects to personalised recommendations with the aim of discovering the potential for sending commercial messages based on personal data;
- To test the market acceptance of products based on data;
- To test the abilities for algorithmic discovery of specific behaviour based on data;
- To find out to what extent profiles are appealing to customers.

3.1 Assumptions

1. Appropriate Compression/Encryption techniques are expected to be provided by the Consortium in order to minimise the risk in case of hacking or stealing of payload data. This includes techniques for querying information on encrypted data.

2. Appropriate authentication / identity access management should be in place for the data subject to log in and retrieve only his/her personal consent and interest data, and allowing him/her to update/delete it.

3.2 Open Questions

1. Social Network Analysis based on CDRs collects information about the numbers that an opted-in customer is calling. The numbers called are not necessarily opted-in. Can their number be stored, or stored as a hash?

Part II

Reusing telephony data for traffic alerts, optimising road layout, and targeted marketing (TLABS)

4 General pilot description

Today T-Mobile Poland collects data 24 hours a day 365 days a year. Data is collected throughout Poland, where the area of measurement is dependent on the density of the base stations. The number of data points (i.e. single unique piece of information) totals to 45 billion. In general T-Mobile Poland collects raw data (i.e. information about the calls, who calls from where for how long etc..) from the base stations. T-Mobile Poland does not control the network, only their own base stations, so they can only get the information from there. The data is stored in Hadoop databases in a manner that support analytics and intelligence gathering. Such data intelligence could be used for individual location insight services, aggregated location insight services, individual location based services and aggregated location based services.

Inorder to exemplify the range of services and the value that could be offered, TLabs will focus on three distinct yet related scenarios. The municipality road layout scenario will investigates sharing location data with local government in order to optimise public infrastructure. The bank travel insurance scenario involves the combination of location and banking data to offer cheaper travel insurance to the client. While, the traffic condition warning scenario will provide realtime traffic alerts to clients.

As per the Proximus, legal requirements will be derived from both the General Data Protection Regulation and the ePrivacy Regulation.

4.1 Location Based Services

Today T-Mobile Poland have 5 million opt-in clients who agree to share their data for marketing purposes. They have three lists: the whitelist is composed of opt-ins, the greylist is composed of customers that have not decided one way or the other and the blacklist is composed of opt-outs. The opt-out list is no higher than 9% for postpaid and mix and only 3% for prepaid clients. In all cases, due to legal requirements in Poland the company needs to both protect and anonymise the individual data. Data protection is currently achieved via a two steps process. In the first step, they strip off the Personally Identifiable Information (e.g. they hash the client ID, an leave the location event, the age group, the gender, zip code, advertisement ID etc... as is). In the second step, the data is aggregated, so that it can be displayed on higher resolution heat map areas, and stored in a Local Data Lake with a new key.

4.1.1 What data is collected

Data collected from the base stations falls into three main categories:

1. Demographic
   (a) Gender
   (b) Age
   (c) Size of the city
   (d) City
   (e) Voivodship/ County
   (f) Real Geolocation according to the base stations

   (g) Movement information - switching between cells

2. Behavioral
   (a) Contract type (business/postpaid/prepaid)
   (b) Average Bill
   (c) Average Roaming Bill
   (d) Number of SMS and MMS sent
   (e) Data Transmission Size
   (f) Operating System Used
   (g) Handset Used

3. Billing - derived from invoice activities
   (a) Drivers - does one use paid navigation apps
   (b) Parents - does one use parental control, find my kid type of apps
   (c) Music activity - what music is being downloaded, ringtones, play while wait gadgets
   (d) Movie activity - VOD
   (e) Language learning - apps that help learning foreign languages
   (f) Games

4.1.2 Obtaining consent

From a consent perspective, the status quo is to ensure that consent is as general as possible e.g. "I agree for my data to be shared with third parties for marketing and business purposes". Under the General Data Protection Regulation (GDPR) this is not sufficient as it is not specific enough and as such does not constitute as informed consent.

According to T-Mobile Poland, obtaining consent in a non-intrusive fashion is particularly challenging, as users will need to explicitly opt-in to any additional processing, and it is difficult to explain to customers what the company would like to do and the potential benefits for the customer.

Another major challenge faced by T-Mobile Poland is the fact that in addition to data protection regulations they are also bound by telecommunications legislation. In contrast, Over-the-top Content (OTT) companies are able to work with highly granular data that does not protect the user anonymity and can easily be traced back to the individuals.

As part of the SPECIAL project T-Mobile Poland would like to investigate how they can provide better/enhanced services to their customers and/or improve local services, by using more precise data, while as the same time protecting their customers personal data. For example, users are flooded with telemarketing information based on the aggregated information about gender, age bracket and phone number. They are annoyed and discouraged. On the contrary, if T-Mobile Poland had access directly to individual roaming data and phone number (no need for any other details) the user could be provided with improved insurance when traveling.

4.2 Use case scenarios

From the data that they have T-Mobile Poland can derive tremendous intelligence and improve customer service. Imagine you could ask questions like *"Where are my clients?"*, *"What's the value of my real estate?"*, *"How many people attend my mass events?"*, *"How can I secure them?"*, *"How can I best access the clients?"*. Today based on the measurements that we have, based on the login information to the T-Mobile's base stations, analysing the data traffic and combining that with the sociodemographic data we can actually answer any of these questions and more. Use cases like immediate information about dangers (fires, floods etc), optimisation of traffic, optimisation of placement of roads, hospitals, public venues so that they are based on where users actually move and work, reducing the redundancy in marketing communication by targeting the right client groups. As part of the SPECIAL project T-Mobile Poland have select three distinct yet related usecases:

Scenario 1: Municipality road layout optimisation T-Mobile Poland would like to give information to local government concerning how often people commute to which regions, what paths they are taking etc, in order to lay out local roads in the optimal way. This requires linking age, usage, time of day and geolocation information and sharing it with a third party.

Scenario 2: Sending bank travel insurance A Bank would like to offer better travel insurance to clients that travel often. They share their phone number database with T-Mobile Poland and would like to know which of these phone numbers often use roaming. This scenario could be problematic from a sharing perspective as both T-Mobile Poland and the bank will need to get the relevant consent for both processing and sharing.

Scenario 3: Sending traffic condition warning T-Mobile Poland would like to send out push information suggesting that users avoid particularly busy roads or public transportation lines based on their location. For example, information like «*It is getting busy on the road, you better call your cab now or you will be late*».

Considering that both scenario 1 and 3 are based on aggregated data, data protection may not be a major consideration, however these pilots could be used to provide some clarity around the boundaries of both the GDPR and the ePrivacy Regulation.

In order to realise these usecases there is a need for a consent and transparency framework that would allow users to understand how their data is being used and what is the actual benefit for them as customers. According to the new 3rd Generation Partnership Project (3GPP) standard we could switch the geolocation on the mobile devices remotely which could be used for optimisation of the network cell coverage (tilting of the base stations). However that needs SPECIAL framework to communicate it properly to the users.

5 Main goals of the investigation

The three focus areas T-Mobile would like to look into are:

1. Communicating important information for marketing purposes;
2. Enriching customer data based on the sociodemographic and behavioral data from the external sources;
3. Sending marketing message to customers who are in a particular area.

All of these elements should be done with as little user-interaction as possible. We need to obtain user-permission in a non-intrusive, friendly and transparent way. Usecases matching the above include:

1. Analysis of user data from various sources within the existing T-Mobile Services;
2. Matching data from various sources from various companies;
3. Profiling T-Mobile users and providing them with messages in real time mode.

6 Data usage policy and consent requests

According to the GDPR, informed consent requests shall specify clearly which data are collected and processed, and for which purpose, in order to make the data subjects' consent legally valid. Then consent requests must specify (at least) the following features, which characterise also the data usage policy that shall be enforced by the company.

Purpose of data collection and processing

Three different use case scenarios are available for this pilot, each corresponding to a different purpose for personal data processing:

1. *Municipality road layout optimisation*
2. *Sending travel insurance ads*
3. *Sending traffic alerts*

Collected data

The data collected is a combination of raw data and billing data. Raw data is the data used to trace the network termination point, telecommunications terminal equipment, the originating of the call and the called party. Such data can be used to infer the date and time of a call and its duration, the type of a call, and the location of telecommunications terminal equipment. Billing data includes information with regard to the paid calls made, including for each call: the called number, the date and time when the call was originated, call duration and the fee charged for this call.

The data categories processed for the above purposes are already collected for the normal operations of the company. Some of them must even be preserved by national law, namely the Polish Telecommunication act.

This pilot focusing on reusing data for different purposes, which generally requires collecting the user's consent, and as such T-Mobile Poland are particularly interested in understanding the boundaries of this consent. The personal data used for the above three purposes are:

1. *Mobile telephony location data*, collected from the antennas to which telephones connect.
2. *Telephone number(s)* provided by the company.

Where are collected data stored

In the company's servers, which are located in Poland.

For how long are the data stored

According to Article 180a of the Polish Telecommunication act, the provider of publicly available telecommunications services is obliged at their own cost to retain and store raw data generated in a telecommunications network or processed by that operator or provider, in the territory of the Republic of Poland, for the period of 12 months counted from the day of a call or an unsuccessful call attempt, and to erase the data as of the expiry of this period, excluding data protected under separate provisions. Raw data is the data used to trace the network termination point, telecommunications terminal equipment, the originating of the call and the called party.

How is data processed and shared

Next, for each of the three scenarios, we describe which of the above data are processed and how.

Municipality road layout optimisation

- The location data are analysed to extract statistics on citizen movements. More specific details are required in order to decide whether consent must be requested or not.
- T-Mobile Poland does aggregation on its own infrastructure. Detailed data never leaves company demilitarised zone.
- Such statistic data are sent to regional municipality service providers. There are different names and organisation structures in different cities.

Sending travel insurance ads The user should be contacted primarily via sms or mms. However we also want to contact through mobile web pages with internet partners – i.e. we can match the Mobile Station International Subscriber Directory Number (MSISDN) with the google advertising identifier (ADV ID) and then serve banners to this user when he/she browses web pages. It was assumed that bank will send the mobile number of their customers to T-Mobile Poland, who will in turn enrich the data with additional scoring data and will return a message such as the following "CLIENT XYZ Traveled Abroad [Frequently|Sometimes|Never]" to the bank. It is assumed that the contact with customer will be initiated by bank.

Sending traffic alerts

- The location data are analyzed to calculate velocity and identify traffic jams.
- Traffic alerts are sent to the users registered to the service via SMS, MMS and Interactive voice response (IVR).
- The analysis will be carried by T-Mobile, however 3rd parties can enrich data.

Open questions

How can users exercise their rights to access, rectify, delete information? Are there any potential conflicts between deletion requests and any laws requiring to store information for a minimum amount of time?

Part III

Know-your-customer reports (TR)

7 General pilot description

Thomson Reuters Org ID is an end-to-end client identity and verification service that provides a complete legal entity due diligence and document management service through which financial institutions and their end clients (asset managers, hedge funds, correspondent banks and corporates) can more effectively manage their response to new KYC regulatory requirements.

DEFINITIONS

- Financial Institution (FI): Thomson Reuter's customer.
- End Client (EC): Financial Institutions customer (i.e. Asset Managers, Hedge Funds, Banks, etc.)
- Related Parties (Entities and data subjects)
  - Entities related to End Client (i.e. branch or subsidiary of a bank)
  - Ultimate Beneficial Owners (UBO)
  - Controlling Owners
  - Senior Management Officials (SMO)
- Outreach: Process for corresponding with an End Client (i.e. Document requests)
- Screening: Process for retrieving risk intelligence data for due diligence from Thomson Reuters screening products (i.e. Screening Online & World Check One)

KYC process flow The KYC process flow is illustrated by the following picture:

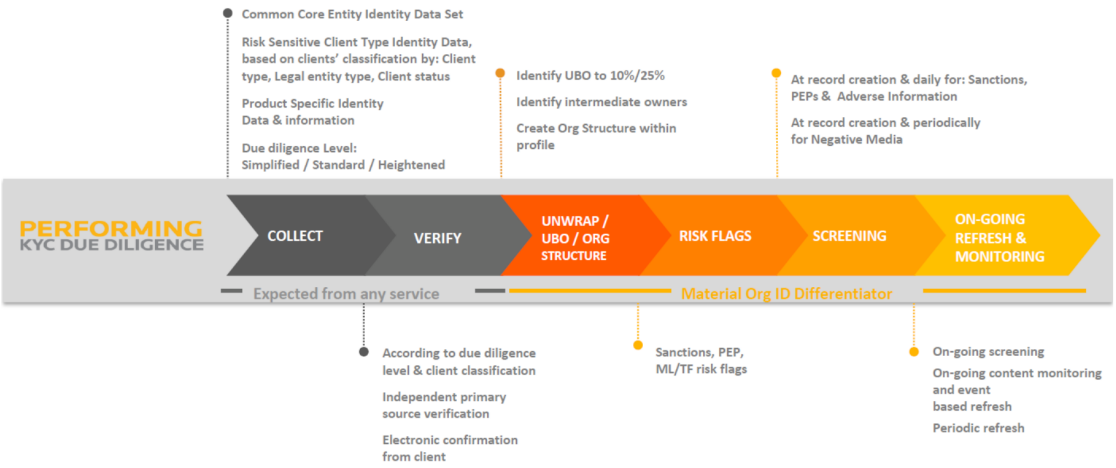


Figure 3: KYC Process Flow

7.1 KYC data collection

7.1.1 Why

Org ID only collects data as far as is required to confirm and verify the ownership and control of the legal entities we undertake KYC on. This is in line with international and national Anti-Money Laundering (AML) laws and regulations. The level and volume of documents/data collected is in line with the risk level at which the legal entity is assessed, typically Low, Medium or High.

7.1.2 What

The data collected by Org ID is used to satisfy various proof types, examples of these proof types are listed below:

- Proof of legal name
- Proof of legal address
- Proof of listing
- Proof of regulations
- Proof of formation

Note: 80-90% of the data collected by Org Id is public data, with the remaining data sourced via Outreach and Screening.

Public Data The majority of standard data required for KYC is collected via public data sources, examples of which are:

- Stock exchanges
- Regulators websites
- End client websites
- Local and international government records
- Country specific data sources
- Relevant industry sources
- Local and international government websites

Private Data As required by the laws and regulations referenced in the Why section above the standard data required is- Name, date of birth, country of residence, domestic address, there are however a number of jurisdictions that require additional data such as former name and marital status, Org ID will not specifically request these data points. Examples of the documents (copies) collected are:

- Passports

- Government issued ID cards
- Birth Certificate
- Driver's license
- Bank Statements (Note: Used for proof of residential address)
- Utility bills (Note: Used for proof of residential address)

Documentation provided to Org ID should have an issued date no older than 3 months from the current date, the only exceptions can be government document (i.e. passport).

Where copies of documents are provided they must be certified in line with Org ID requirements are in line with industry standards i.e. dated & signed by a professional person e.g. Lawyer, magistrate as being a true copy of the original document.

Legal Attestation is required when documents compiled by the end client (EC), such as details of beneficial owners or SMO's provided on the EC's headed paper are submitted rather than a copy of an official document containing the information. In order for the EC document to be treated as a primary source it must be legally attested.

Legal attestation must be by a qualified legal professional, lawyer or solicitor, and must state that they have verified that the information contained in the document prepared by the EC is correct. The attestation must be dated, signed and accompanied by the firms stamp and any relevant registration no.

EC compiled document that are not legally attested can be accepted but are treated as a secondary source.

Source of funds data Source of funds is only gathered when required by specific policies and is only collected for an End Client. Information gathered for this requirement is collected primarily via public data but if no public data is available then Outreach maybe used depending on the policy requirements.

Screening data Screening results will contain heightened risk information about individuals and organisations, examples are listed below:

- PEP status (Politically exposed person)
- Sanctions
- Adverse / Negative media
- PII and private data such as date of birth, country of birth and residence as well as employment details

Note: refer to the following Appendix sections for details of data classifications and field definitions:

- Strictly confidential data classification
- Database Schema / Data Model

7.1.3 How

The majority of public data is collected from public data sources using a mixture of automation and manual processes depending on the sources. For private data in nearly all cases this is collected via Outreach.

Outreach Outreach is a process for corresponding with an EC through various different communication methods (i.e. email, portal, etc.) Examples of the types of correspondence are listed below:

- Document request
- Consent request
- Account registration
- Account troubleshooting

Screening Screening data is provided by TR screening products which comes from highly structured database of intelligence on heightened risk individuals and organisations.

This data is populated by hundreds of research analysts across all global regions and speaking more than 60 local languages. Information is collated by these analysts comes from an extensive network of hundreds of thousands of reputable sources, including:

- 530+ sanction, watch, regulatory, and law enforcement lists
- Local and international government records
- Country specific data sources
- International adverse electronic and physical media searches
- English and foreign language data sources
- Relevant industry sources

For more information on these products please refer to the following links:
`https://www.thomsonreuters.com/en/products-services/risk-management-solutions/customer-and-third-party-risk/thomson-reuters-world-check.html`
`https://risk.thomsonreuters.com/en/products/world-check-know-your-customer.html`

7.2 KYC Data Storage

Where Org ID data is primarily stored within a centralised BPM system which is governed with strict security controls and standards for housing strictly confidential data. All strictly confidential data is hosted within data centers in the UK and can be found in the following locations:

- Database - Data fields within Database Schema / Data Model used to store End Client data.

- Document Repository - Files on file server used to store Proofs which may contain strictly confidential data.
- Export folder - Files on file server used to store Profile Reports which contain strictly confidential data.

How long for Org ID data retention principles are in line with FAFT requirements: http://www.fatf-gafi.org/media/fatf/documents/reports/AML_CFT_Measures_and_-Financial_Inclusion_2013.pdf See section 'RECORD-KEEPING REQUIREMENTS'

7.3 KYC Data Processing

In terms of what we computed in Org ID:

- Our policy operates a comprehensive covers 41 capital markets
- We unwraps the legal entity and identifies the ultimate beneficial owners (UBO)
- We maintains and updates legal entity records on a continuous basis
- We screen entities and identify risk flags
- We source legal entity documents in over 168 countries and 60 languages
- We provide on-going monitoring and refresh of end client profiles.

We link data in the following way. If we are screening a Senior Management Official and we gather from public sources his or her age / location, our analyst will use that information to fine tune the results of the screening application, to help add accuracy. We do not do ad-hoc research and only use approved sources when collecting data for processing.

Data contained within Org ID is segregated from other internal sources/systems and is only tagged with identifiers for content sources and matching (i.e. Thomson Reuters Perm ID.)

ON-GOING MONITORING As part of the ongoing monitoring we will monitor for change via the following data sources:

- Public data (i.e. Change of address)
- Screening alerts (i.e. New sanction or PEP)

Notes:

- If data cannot be verified then in some cases we will request confirmation via Outreach from the End client.
- If a significant change is identified then a Refresh is triggered, examples of significant changes are listed below:
  - Change of jurisdiction
  - New PEP or sanction
  - Change to due diligence level

REFRESH End client profile reports will be refreshed as part of a periodical process or on demand to ensure data is current and will trigger the complete KYC process again.

7.4 KYC Data Sharing

WHO We share the following types of information with Financial Institutions which are part of a Profile Report:

- Public information about an end client without getting consent.
- Screening results about an end client without getting consent
- Non-public information with consent.

We share the following types of information with End Clients:

- Attestation report to confirm validity of data collection (Note: only applicable to South Africa at present)

Note: We do not share the due diligence level, the results of screening or any risk flags to the End Client.

Exceptions There are a few exceptions to the standard sharing processes defined above which have been listed below:

- Regulatory requests (i.e. FCA, etc.)
- Law enforcement requests (i.e. National Crime Agency of UK, etc.)
- Related party requests (i.e. data subject)

Note: The data provided for these exception cases will vary depending on the circumstances.

HOW End Clients agrees to the customer portal terms and conditions. Sharing is via a permission to share to the Financial Institution, so it's explicit and informed.

8 Data usage policy and consent requests

Here we focus on the data that are subject to the GDPR, therefore fall in the scope of SPECIAL. In other words we summarise how the above pilot collects and manipulates the personal data referring to the individuals that are subject to the KYC procedure, i.e. UBOs, SMOs, Controlling Owners.

8.1 Purpose of data collection and processing

The main purpose is carrying out the KYC process in accordance with the Anti-Money Laundering laws and regulations. The KYC process is meant to produce a risk assessment report about a legal entity that may be supported by evidence regarding related individuals (UBOs, SMOs, Controlling owners).

8.2 Collected data

The data referring to individuals are:

- demographic data (name, date of birth, country of residence, domestic address), associated to evidence such as (copies of) passports, government issued ID cards, birth certificates, driver's license, bank statements, and utility bills, that may carry further personal information (place of birth, parents' names, etc.);
- negative/adverse media about UBOs, SMOs, Controlling Owners;
- source of funds and employment details for the same individuals.

The possile sources of demographic data are:

- outreach (that results in the data subject's disclosing some of the above documents and giving consent to produce the screening reports and share them with the specified FI)
- public sources (e.g. public administration, media).

Negative/adverse media are clearly gathered from public sources. Sources of funds are normally gathered from public sources, but they may be collected via outreach, if needed.

It should be verified whether the GDPR requires the explicit consent from the data subjects also for collecting and processing such public information.

8.3 Where are collected data stored

On TR's data centers located in the UK.

8.4 How long for

The policy specifies minimum and maximum storage time in accordance to the aforementioned applicable regulations.

8.5 How is data processed and shared

The collected data are used for identifying risk flags and assembling a risk report about an End Client. The report may contain as evidence (a subset of) the private information referring to the data subjects. So personal data may be shared with Financial Institutions (FI).

If such data are not collected from public sources, then a consent request specifying the FI is sent to the data subject. Otherwise, no consent is currently requested, and it should be verified whether informed consent will become necessary when the GDPR will come into force.

The data concerning individuals are also shared with the data subject to confirm their validity (South Africa only).

Data may be further shared as required by applicable regulations. Moreover data subjects may exercise their data access rights (note that the due diligence level adopted, risk flags, and the results of screening are *not* shared with the data subjects).

8.6 Consent requests

During the outreach phase, individuals may be asked for documents and the consent to use them for the KYC process. The data subjects have a few options here. First, they may choose which documents to disclose (from the aforementioned list). Second, they may deny consent to using some or all of their personal information. In this case, the final risk assessment report may be incomplete.

8.7 Open questions

1. Does the current workflow support thoroughly the data subjects' rights to access, rectify and delete their personal information?
2. Is there any conflict between deletion rights and applicable regulations requiring that data be stored for a minimum given period?
3. Can the data subjects object to using public information related to themselves? Is this a right guaranteed by the GDPR?

9 Main goals of the pilot

Thomson Reuters main goals for the GDPR research pilot have been outlined below:

- To validate options for automating compliance to the GDPR using machine-readable policy-based approaches to regulatory enforcement within the Org ID product.
- To investigate consent and transparency-based approaches to sharing risk-related data between financial institutions (within and beyond the Org ID product).

STRICTLY CONFIDENTIAL DATA CLASSIFICATION Org ID's definition of Strictly Confidential (SC) data has been defined into the following data types:

- Sensitive Thomson Reuters Data: where we store significant quantities of data such as company finances and billing, M&A, or litigation
- Sensitive Customer Data: where we store sensitive data owned by customers, such as portfolio details, or data about customers, such as usage data
- Personally Identifiable Information: where we store information that can be used to identify a single person uniquely

DATABASE SCHEMA / DATA MODEL The table below outlines the database schema / data model most relevant to the GDPR research (i.e. Personal / PII data) <Removed from public version>

Part IV

Discussion

10 A unified view of the location based services use cases

In order to pave the way to policy reuse, policy templates, and consent request reduction or reuse, it is useful to elicit a unified view of the location based services (LBS) occurring in the pilots led by Proximus and TLABS.

10.1 A simple location based service

In a generic setup, a location based service can be broken down into the following core processes:

1. Collection of events or other information points
2. Collection of location information concerning those events or information points
3. Collection of live location information from mobile subscribers
4. Collection of other information from those subscribers
5. Merging the collected information into an interest-profile for those subscribers
6. Definition of an area or sector of application
7. Matching of mobile user location information and event location information
8. Sending information about one or more events to the user's mobile device

The actors in *Figure* 4 are named according to their contractual role. Using the technical terms of the data protection legislation, the mobile subscriber is the «*data subject*» and the LBS intermediary is the «*data controller*». The event or service provider is a third party who provides business to business services to the Telco.

Even though the figure looks rather simple, the variety of location based services based on this architecture is near to infinite. In a first iteration of the LBS case as shown by the figure, the LBS intermediary helps achieving a privacy friendly result as it stands in the middle between the event provider and the mobile subscriber. In this example presented the LBS intermediary is a telecommunications enterprise (*Telco*).

The information about the event may include all kinds of further promotional or practical information. We can think e.g. about a link to reserve a ticket for the event that can be directly paid by the user via their mobile phone subscription.

The LBS scenario is a very powerful service concept that allows the filtering of information, such that only the most relevant points are conveyed to the user, thus protecting against information overload.

But it requires the use of sensitive personal data from the mobile subscriber. As such the process enumeration outlined above is incomplete as the user needs to be made aware of and be given the ability to control the use of their personal data, to a certain extent.

SPECIAL bridges this gap by on the one hand making mobile users aware of what data is stored/used and for what specific purpose(s), and by on the other hand providing Telcos with a means to obtain the consent of the data subject. Thus allowing the LBS to operate in a legally compliant way.

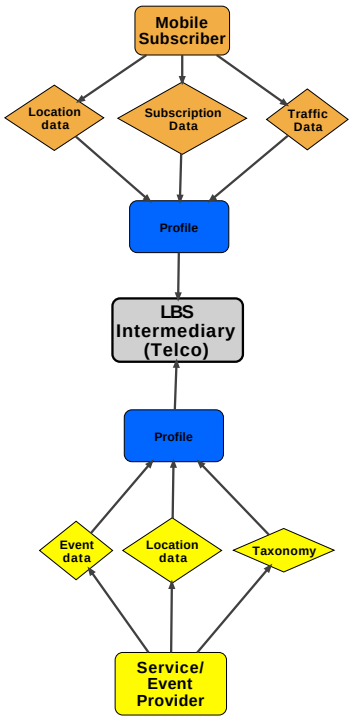


Figure 4: A simple location based service

10.2 Consent for the LBS case

Assuming that the data controller is a Telco and that the data has already been collected from the mobile subscriber in the course of normal operations and is stored in some operational database. In such a scenario, it is highly likely that the initial collection of that data was to fulfil other purposes than to provide location based services, as such it is necessary to seek specific consent from the mobile subscriber for LBSs. A consent and control architecture is thus needed in order to enable the Telco to re-use the existing data in order to provide new services to the client. The SPECIAL LBS case is thus a typical big data case where existing data is re-used for new innovative services. In the Proximus- case, data from different sources is funneled into the mobile subscriber's profile. As Proximusoffers several telecommunications services, it re-uses data from several sources. This matches the case from Deutsche Telekom. They lookup data in their operational databases to find patterns of movement or other patterns that allow Deutsche Telekomto derive concrete needs the mobile subscriber may have. Deutsche Telekommay then look into the list of third party service providers to find a service matching the need.

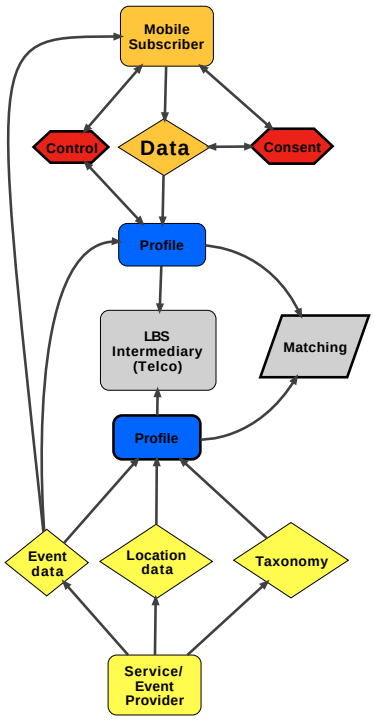


Figure 5: Location based scenario with consent/control

10.3 Service delivery

The service can be delivered in two ways. Either the Telco uses the third party services as a mere data processor. Or it uses the service as a subcontractor with its own rights.

10.3.1 Data processor

By definition, a data processor is transparent to the user, because a data processor does not have their «*own rights*» on the data and processes the data «*on behalf*» of the data controller. In the Proximuscase, the data controller is the Telco that sits in the middle between mobile subscriber and event provider. The Telco may use some SME to host the creation of the profiles according to a detailed data processing agreement. But this SME would not have own rights and only returns the results of the processing to the Telco. In case, the SME returns results to the mobile subscriber, there is no direct need to mention the data processor. But if a mobile subscriber is told that his data will only go to his Telco and he will get results from another source, it may have implications on the trust towards that Telco. Consequently, despite not being legally necessary, the consent should contain information about the data processor in case his data processing is visible to the user.

10.3.2 Service concertation

Today, a single news site is often the result of the dynamic contribution of up to 300 sources. Often, those sources are not owned or controlled by the maker of the news site. The news site essentially embedded additional services into their own service. The same can go for location based services. The Telco has the information about the user. Based on that information, third parties can provide adapted additional useful services to the mobile subscriber. In our LBS case, the Telco has already all the data. The data is thus not newly collected from the third party. The Telco discloses the data to the third party in order to allow the adaptation of the service provided. But in the European legislative framework, arbitrary disclosure of personal data to the third parties is not allowed. The Telco needs to ask the user for consent. In a scenario with a single service, this may be commensurable. But if we think about the initial news site scenario with a 300 providers, we can imagine how difficult it will be to get consent for all 300 from a data subject using their mobile phone.

The deliverable on legal requirements will explore in how far the SPECIAL interface can provide a contextual form to allow to reduce complexity. It will also have to explore how data categories can be used to convey useful information in order to get get agreement from the data subject for additional processing.

11 Policy features occurring in the pilots

Here we summarise the features of the data usage policies occurring in the above pilots, and point out some of the challenges in turning these features into a machine-understandable policy language. Indeed, such features constitute a minimal set of requirements for SPECIAL's policy language. More requirements may result from SPECIAL's strive for generality, aiming at addressing a wide spectrum of applications.

11.1 Personal information categories

Before providing an overview of the formal policy statements and related challenges, we first summarise the categories of personal data and the types of profiles that are required in order to support the pilots.

Categories of personal data: The personal information collected and processed by the pilots spans over the following categories[3]:

1. the typical contents of personal records, such as name and other PII (e.g. tax IDs), birthdate, address, and the like; (all pilots)
2. behavioral information, such as internet navigation history and TV program choices; (PROX's pilot)
3. billing information (PROX and TLAB's pilots);
4. location data (PROX and TLAB's pilots);

[3]Note that location data and telephone call data will be subject to the ePrivacy regulation.

5. telephone call data (PROX's pilot);
6. information about income and financial activities (TR's pilot);
7. news collected from public sources (TR's pilot).

The SPECIAL project needs to investigate if any available ontologies describe (thoroughly or partially) the above data categories. This analysis will consider work by the W3C regarding Geospatial Ontologies[4], news ontologies such as those provided by the BBC[6], to name but a few. More generally, a standard vocabulary for categorising personal information is highly desired and relevant to the project.

Personal profiles: As a result of the processing of the above data, different types of personal profiles are created, stored, disclosed and/or used for further elaboration:

1. personal interest profiles (PROX and TLAB's pilots);
2. risk-related personal profiles (TR's pilot).

In some pilots (e.g. the PROX pilot), the detailed personal information collected could even be deleted right after using it for creating or updating the subject's interest profile.

11.2 Formal policy statements and related challenges

The policy language needs to be able to describe the purposes of the processing, recipients or categories of recipients of the personal data, the period for which the personal data will be stored, how policies are associated with data and obligations for the data controller and/or processor.

Purposes of the processing: Usage policies shall specify the *purpose* for information collection and processing. The list of purposes occurring in the pilots fall within the following categories:

1. recommendations;
2. targeted advertisement;
3. (traffic) alerts;
4. public infrastructure planning;
5. risk assessment (in the KYC sense).

Some of the existing policy languages provide a simple categorisation of purposes that may used as a starting point to develop a purpose ontology suited to the goals of SPECIAL. The finer the level of detail required, the more challenging the formalisation would be, given the ample range of purposes conceivable in principle. In general, it may be necessary to combine a machine-understandable description with more articulated,

[4]W3C Geospatial Ontologies, https://www.w3.org/2005/Incubator/geo/XGR-geo-ont-20071023/, existing telecommunication vocabularies [5]

[6]BBC ontologies, http://www.bbc.co.uk/ontologies

human-readable text. Still, we expect most applications to be focused on a restricted set of common purposes.

Policies should also specify *how data are processed*. Describing data processing in a machine understandable way is a challenge, given the variety of possible aggregation, anonymisation and analysis procedures. Even if the focus is placed on the *results* of the processing (e.g. interest profiles, risk assessment etc., as opposed to the algorithm or mechanism applied) the space of possibility is wide enough to make the develoment of a unique standardised vocabulary a difficult task. For the same reason, describing the derived data shared with third parties is similarly difficult. In the TR pilot, data processing may have to be related to mixed human/automated workflows.

The *location* where data are stored may also be specified if needed. Usually, the granularity needed for this feature concerns company borders and the nation(s) in which involved companies are located. The latter indicates which privacy regulations are applicable, and whether data may be transferred to those countries. Accordingly, it is relevant to know whether the country is located in the EU, US or other unions/federations of states with shared privacy laws. Formalising such information is not expected to pose any major problems.

Recipients or categories of recipients of the personal data: *Disclosure to third parties* comes into play in two of the three pilots (TR & TLAB's). The policy should describe the recipients, that in our pilots typically include: public offices, service providers, financial companies and banks. An ontology is needed for describing such entities. Moreover, the policy should indicate which information is disclosed to the third party. This may be challenging, as explained in the following point.

Period for which the personal data will be stored: Currently the pilots' usage policies need *temporal constraints* on information persistency (e.g. deleting information not sooner or not later than a specified amount of time due to applicable laws). The GDPR and ePrivacy may introduce further needs, e.g. personal data should be kept only for as long as necessary for carrying out the provided service.

Associating policies with data: There must be a way of *linking policies and data*. It should be possible to identify which data are subject to a given usage policy, as well as which policy applies to a given piece of data. In SPECIAL, data and policies shall be identifiable via suitable Uniform Resource Identifiers (URIs) so as to leverage semantic web languages and standards.

Obligations derived from the use cases: At this stage, the pilots seem to require no further description of *obligations*, beyond the persistency and deletion guarantees implied by the aforementioned temporal constraints.

11.3 Enabling consent, transparency and compliance

In order to guarantee that data processing is aligned with the consent obtained by the data subject, SPECIAL plans to adopt a unique internal policy format that is used both for generating consent requests (that are presented to the data subject) and for compliance checking (by the data controller and/or processor).

Consent requests: Accordingly, the usage policy description that will be presented to the data subject in the *consent request* will be automatically derived from the internal format. Policy presentation may be dynamic, especially when the policy is complex. The idea is that users can interactively explore data collection and usage modalitites, focusing on the aspects that they are interested in and selecting the level of detail appropriate to their needs. Automated policy verbalisation and explanation systems have been studied and realised in the past: for example, the tools for visualising P3P policies in a human-readable manner (cf. the tools listed on https://www.w3.org/P3P/implementations.html, such as P3P Display, Privacy Bird and the policy management components of Internet Explorer; see also [21]), the second generation verbalisation and explanation system for Protune [4, 3], Android's tools [9] and more [20]. Still, making such dynamic policy presentation systems effective is an open research topic, cf. [20, 11, 10, 16].

Checking compliance: Concerning *compliance checking*, this may be instantiated as a sort of decision support system, to assess the compliance of personal data processing workflows with the GDPR and ePrivacy, or it may be conceived as an enforcement system in charge of:

1. access control decisions;
2. action scheduling (such as obligatory data deletions).

Policy changes: The pilots allow for user-driven after the initial consent.[7] We foresee mainly the addition of data processing purposes and third parties to whom information is disclosed, as a result of new consent requests. In some applications (e.g. PROX's pilot) the data subjects can indirectly tune the policy by changing their personal interest profiles; this affects the flow of recommendations that the user receives, therefore – to some extent – the purpose of data usage. Moreover, in several pilots (PROX, TR), the data subjects may give their consent to using only a subset of the personal data that might potentially be used; of course, this may affect the results of the processing (e.g. less focussed recommendations/advertisement and incomplete risk assessement reports).

12 Legal challenges

The regulatory environment of the use cases is changing. With the advent of the new European Data Protection Regulation[23][8], existing regulation around Data Protection will be updated or replaced. The replacement will be partly automatic because the new framework is a European regulation that is directly applicable in all Member states of the Union. According to its article 99, the GDPR will come into force on 25 May 2018. It is therefore not advised to view the use cases under the current legal framework. SPECIAL decided to only consider the GDPR and other applicable legal provisions. A look at the new ePrivacy Regulation replacing Directive 2002/58EC[8] will complement the legal analysis of the use cases.

[7] Here we do not refer to control-related actions such as information deletion requests and opt out requests, that we consider as different concepts.

[8] GDPR in the following

One of the basic principles in data protection in the EU is inherited from the German model. It is a usual legislative technique in German public law to reverse the paradigm of democratic freedom. While the democratic default is, that everything not prohibited shall be allowed, we have the opposite situation here. Article 6 of the GDPR says that *Processing shall be lawful only if and to the extent that at least one of the following applies* and subsequently lists the exceptions. This means implicitly that all processing of personal data is prohibited, if it is not allowed by the GDPR or any other provision as determined by article 6. The general prohibition strikes where personal data as defined by the GDPR is concerned and where there is no permission given. Consequently, data processing can escape the prohibition by only dealing with data that is not personal data. Or acquire a legal or other permission to process.

The GDPR itself, but also Directive 2002/58EC contain themselves many permissions. A good example is the article 4 of Directive 2002/58EC that allows the collection and processing of traffic data for the security of the electronic communication service. In case there is no legal provision allowing for the collection and processing of personal data, article 6 (1)1. requires the data subject's consent for such processing.

But not all collection allows for all processing. Nearly all our use cases also contain cases where data already collected is re-used for a different purpose. This is one of the core use cases in big data. Large amounts of data were collected under a certain permission and it is now extremely tempting and potentially very beneficial to the enterprise or even to society to reuse the data already present. Especially in the scenarios around the location based services, the telecommunication enterprise has normally already collected most of the necessary data under their normal networking operations. Those networking operations are legal under provisions from the GDPR, but also because of provisions from the very densly regulated telecommunications area.

In case personal data is collected, the permission has to exist before the actual collection happens. In case data was legally collected, the re-use may be privileged, e.g. for statistical, historical or archiving purposes. Statistical information is only privileged if the aggregation is such that from the statistical result one can not trace back to the individual data record or data subject. In case the privilege applies, there is no need for an additional permission by the user. This may affect several branches of the SPECIAL use cases.

The legal discussions during the exploration of the SPECIAL use cases were of a remarkable complexity. But even more striking were the general hesitations and doubts about the scope of application of the GDPR and the extend to which permissions would cover the intended data processing. The challenge for the legal team is to help the use case partners to navigate through the difficult terrain. Instead of just giving legal scrutiny to a suggested scenario, the legal team has to be already involved in the design. The further challenge is that several legal requirements, e.g. requests for permission, are inherently disruptive of the user experience. In this tension, legal team and technical team have to find compromises and smart escape lines to allow not only for a legal, but also for a usable implementation of the suggested use cases.

12.1 Challenges regarding anonymisation

Some data is not personal data by nature, e.g. weather data or geo-sensor data about the state of some soil and similar sources of data. Other data has a strong personal

link. This is especially true for telecommunication data as the communication between people is social by nature. Telecommunications enterprises collect a lot of data. This data has a high potential to generate benefits for the entire society. But asking the data subjects for each and every data reuse would create just annoyance without serving the real data protection goals. It is therefore important to have anonymisation as an escape line. The challenge here is to avoid re-identification.

Anonymity criteria can be categorised in two families: (i) k-anonymity and its evolution, such as l-diversity and t-closeness, just to name a few; (ii) ϵ-differential privacy and its refinements. For each anonymity citerion there exist algorithms and mechanisms that achieve that kind of anonymity. The methods applying to the first family typically operate by removing information (e.g. by deleting day and month from birth dates); the methods applying to the second family typically introduce noise in query answers or directly in the data, in a controlled way (special methods are available for non-numeric data). Still there exist relationships between the two families, see for example [18, 13]. A nice overview of anonymisation criteria and techniques, including a discussion of their pros and cons, can be found in [2].

Clearly, when data can be anonymised, legal re-purposing and sharing become extremely easy, as it is no longer deemed personal data and as such no further contact with the user is required. Unfortunately, it is currently not clear when data can be considered anonimysed, from legal as well as technical viewpoints.

A first difficulty is that the anonymity criteria are parametric; the parameters – roughly speaking – are related to the difficulty of re-identifying the data subjects. Fixing the parameters is not an easy task [17, 5]. Unfortunately the law does not fix any official threshold for any of these parameters, so it is not clear when the data can be considered sufficiently anonymised. Furthermore, the probability of re-identification can hardly be reduced to zero, and the variety of additional information sources and algorithms that can be used to enrich the (supposedly) anonymised data (thereby reducing the anonymisation level, i.e. changing the anonymity parameters) is so wide that in practice one can never be sure that the chosen parameters actually prevent re-identification – nor that the parameters of the enriched information remain below any privacy threshold established by future laws. The vulnerability of all privacy enhancing mechanisms to attacks based on available background knowledge have been discussed in several papers, including [24, 14, 5, 19]. The approaches derived from k-anonymity suffer from an additional drawback, namely, they turned out to be vulnerable to novel attack models. While solutions have been proposed for each new attack, currently no results guarantee that these methods are not vulnerable to any further, unforeseen attack models (cf. [7, 24, 5]).

A second, well known difficulty is that anonymisation decreases data quality and – accordingly – data *utility* [2, 22, 14, 17, 5]. The required anonymisation level may turn out to be incompatible with the necessary data quality in many applications. In particular, this consideration shows that a cautious approach, based on choosing extremely "conservative" privacy parameters, should not be expected to be viable in practice.

A third difficulty relates directly to the second one. The science about de-anonymisation of anonimysed data sets is advancing swiftly. The more there is computing power and the more we understand data, the more science is able to single out individuals from a large anonymised data set[12]. But the stronger the anonymisation needed, the less

meaningful information will be left in the resulting database.

Relating the above discussion to SPECIAL's objectives, if the pilots' realisation were uniquely centred around anonymisation, then they would most likely be difficult to assess legally, due to the lack of precise definitions of what can be considered anonymous. Moreover, any target anonymisation level should be matched against the quality and utility of the anonymised data, for each use case and each scenario.

The above discussion leads us to acknowledge that anonymisation techniques are effectively applicable in a restricted number of cases, and pushes us to leverage informed consent, instead.

12.2 Consent and its inherent challenges

The legal model of the current data protection framework is coherent and the GDPR only makes a few changes to that model. To overcome the general prohibition mentioned above, the data subject has to give their consent to the intended data processing. It is very important to recall the relevant rule in the GDPR:

> **Article 4** defines «*consent*» this way:
> (11) «*consent*» of the data subject means any freely given, specific, informed and unambiguous indication of the data subject's wishes by which he or she, by a statement or by a clear affirmative action, signifies agreement to the processing of personal data relating to him or her;

In theory, everything is fine in the legal world. Data self determination is achieved because the data subject is informed of all data processing that concerns them and they agree to that data processing. The data controller creates some evidence of the agreement. Like for a contract, the information contains the data items collected and the purpose pursued as well as some additional information, e.g. data retention times.

In practice it is not that simple because the legal system doesn't scale to the needs of our networked and digitised society. If we talk about big data, the velocity and variety of data collection and reuse exceeds every possible human interaction. Declarations are then generalised to data categories and to generic purposes. Beautified text lures the data subject into agreements that they would never have agreed to, had they understood the full extend of the data processing. As a direct result of this practice the difficulty in getting consent is increasing.

But even with generalisation the challenge persists because of the number of consent requests. This can be especially frustrating when consent requests are very similar to each other, to the point that data subjects might not see the rationale for such repeated requests. Already the Primelife project[15] addressed the challenge of finding a good interface to collect consent.

The issue was known during the deliberations on the GDPR in the Parliament, the Commission and the Council. It is important to explore the options given. We first have to look into the explanations given in the relevant considerations:

> **(32)** Consent should be given by a clear affirmative act establishing a freely given, specific, informed and unambiguous indication of the data subject's agreement to the processing of personal data relating to him or her, such as by a written statement, including by electronic means, or an oral statement. This could include ticking a box when visiting

an internet website, choosing technical settings for information society services or another statement or conduct which clearly indicates in this context the data subject's acceptance of the proposed processing of his or her personal data. Silence, pre-ticked boxes or inactivity should not therefore constitute consent. Consent should cover all processing activities carried out for the same purpose or purposes. When the processing has multiple purposes, consent should be given for all of them. If the data subject's consent is to be given following a request by electronic means, the request must be clear, concise and not unnecessarily disruptive to the use of the service for which it is provided.

(42) Where processing is based on the data subject's consent, the controller should be able to demonstrate that the data subject has given consent to the processing operation. In particular in the context of a written declaration on another matter, safeguards should ensure that the data subject is aware of the fact that and the extent to which consent is given. In accordance with Council Directive 93/13/EEC (1) a declaration of consent pre- formulated by the controller should be provided in an intelligible and easily accessible form, using clear and plain language and it should not contain unfair terms. For consent to be informed, the data subject should be aware at least of the identity of the controller and the purposes of the processing for which the personal data are intended. Consent should not be regarded as freely given if the data subject has no genuine or free choice or is unable to refuse or withdraw consent without detriment.

(43) In order to ensure that consent is freely given, consent should not provide a valid legal ground for the processing of personal data in a specific case where there is a clear imbalance between the data subject and the controller, in particular where the controller is a public authority and it is therefore unlikely that consent was freely given in all the circumstances of that specific situation. Consent is presumed not to be freely given if it does not allow separate consent to be given to different personal data processing operations despite it being appropriate in the individual case, or if the performance of a contract, including the provision of a service, is dependent on the consent despite such consent not being necessary for such performance.

The goal of SPECIAL is to provide a context sensitive way of collecting consent that fulfils the requirements above. This means an affirmative action is needed. This affirmative action has to be embedded into the context of the actual service environment that is the starting point of our use cases. But we have additional constraints as two of the use cases have a mobile component. Thus a SPECIAL solution has to take the particular constraints of the mobile device into account. In this context it becomes clear that it is not really reasonable to confront the data subject with a large PDF that they would have to read on their mobile device.

The unified account of LBS outlined previously provides some concrete examples of this issue. Would a data subject who already opted in for the travel insurance ads appreciate a new consent request where the only difference is that ads are now about fitness? This is questionable, especially if the data subject has fine-grained control on the ads topics of her interest and can tune *ex post* incoming ads messages, e.g. by adding or blocking topics. In this respect, the Proximus pilot – where users can modify their interest profile as they please – is quite interesting and raises a research question in the legal domain:

1. *Can fine-grained control balance generality in consent requests, and to what extent?*
2. *Can we contextualise the consent request and add to the sum of consented things over time?*

Another major concern of SPECIAL's industrial partners is that new business opportunities themselves are frequently discovered by mining personal data and analysing

customer interests. So we run into a chicken-and-egg problem: *Companies need user consent to analyse their personal data, but at that moment of the request they can't articulate precisely the purpose because it is not yet known* (as the purpose is the result of the analysis). Again the challenge is *crafting a valid consent request that the law requires to be very specific, while the current knowledge of the requester does not allow that level of detail.* Also in this case, it would be interesting to investigate whether fine-grained control may compensate the lack of details of the initial request.

Solving this issue is crucial for supporting innovation in the European industry. The strictest interpretations of the GDPR and ePrivacy rule out the data analyses that are needed to identify new services and new business opportunities, thereby hindering the competitiveness of the companies operating on the European territory. The effect of this may even be counterproductive from a privacy perspective, since the data subjects who do not find certain services in Europe may be induced to resort to extra-European providers and have their data processed with much weaker privacy guarantees.

But SPECIAL will have a contextual consent request as it appears while data subjects are using their mobile device. The affirmative action is recorded via SMS or email. In the case of the Proximus employees, the consent can be recorded «*out of band*» on a paper signed at the workplace. But the goal is really to have a fully electronic consent mechanism. As the system will allow the user to control their own profile at any time, the affirmative action tries to be the least disruptive possible. This can be as simple as a reply to an SMS explain

Making consent usable for the data subject will lead to more permissions. Having a control interface for later alterations will allow to be less strict on the first encounter of the data collection or purpose change. Because SPECIAL adds a control layer into the technology stack, companies can allow for control as they are able to fulfil the promise in their backend.

The challenge remains to enable technology innovation in such a way that it is in line with the GDPR and to identify potential gaps in the ePrivacy Regulation currently under discussion.

Bibliography

[1] *Twelfth Symposium on Usable Privacy and Security, SOUPS 2016, Denver, CO, USA, June 22-24, 2016*. USENIX Association, 2016.

[2] G. D' Acquisto, J. Domingo-Ferrer, P. Kikiras, V. Torra, Y. de Montjoye, and A. Bourka. Privacy by design in big data – an overview of privacy enhancing technologies in the era of big data analytics. V 1.0, ENISA Report, 2015. https://www.enisa.europa.eu/publications/big-data-protection.

[3] Piero A. Bonatti, Juri Luca De Coi, Daniel Olmedilla, and Luigi Sauro. A rule-based trust negotiation system. *IEEE Trans. Knowl. Data Eng.*, 22(11):1507–1520, 2010.

[4] Piero A. Bonatti, Daniel Olmedilla, and Joachim Peer. Advanced policy explanations on the web. In Gerhard Brewka, Silvia Coradeschi, Anna Perini, and Paolo Traverso, editors, *ECAI 2006, 17th European Conference on Artificial Intelligence, August 29 - September 1, 2006, Riva del Garda, Italy, Including Prestigious Applications of Intelligent Systems (PAIS 2006), Proceedings*, volume 141 of *Frontiers in Artificial Intelligence and Applications*, pages 200–204. IOS Press, 2006.

[5] C. Clifton and T. Tassa. On syntactic anonymity and differential privacy. In *2013 IEEE 29th International Conference on Data Engineering Workshops (ICDEW)*, pages 88–93, April 2013.

[6] Lorrie Faith Cranor, editor. *Symposium On Usable Privacy and Security, SOUPS '12, Washington, DC, USA - July 11 - 13, 2012*. ACM, 2012.

[7] Josep Domingo-Ferrer and Vicenç Torra. A critique of k-anonymity and some of its enhancements. In *Availability, Reliability and Security, 2008. ARES 08. Third International Conference on*, pages 990–993. IEEE, 2008.

[8] European Parliament and Council Directive. Directive 2002/58/ec of the european parliament and of the council: concerning the processing of personal data and the protection of privacy in the electronic communications sector (directive on privacy and electronic communications). Official Journal of the European Communities, 2002.

[9] Adrienne Porter Felt, Elizabeth Ha, Serge Egelman, Ariel Haney, Erika Chin, and David Wagner. Android permissions: user attention, comprehension, and behavior. In Cranor [6], page 3.

[10] Adrienne Porter Felt, Robert W. Reeder, Alex Ainslie, Helen Harris, Max Walker, Christopher Thompson, Mustafa Embre Acer, Elisabeth Morant, and Sunny Consolvo. Rethinking connection security indicators. In *Twelfth Symposium on Usable Privacy and Security, SOUPS 2016, Denver, CO, USA, June 22-24, 2016* [1], pages 1–14.

[11] Joshua Gluck, Florian Schaub, Amy Friedman, Hana Habib, Norman M. Sadeh, Lorrie Faith Cranor, and Yuvraj Agarwal. How short is too short? implications of length and framing on the effectiveness of privacy notices. In *Twelfth Symposium on Usable Privacy and Security, SOUPS 2016, Denver, CO, USA, June 22-24, 2016* [1], pages 321–340.

[12] Shouling Ji, Prateek Mittal, and Raheem Beyah. Graph data anonymization, de-anonymization attacks, and de-anonymizability quantification: A survey. *IEEE Communications Surveys & Tutorials*, 2016.

[13] Shiva Prasad Kasiviswanathan and Adam D. Smith. A note on differential privacy: Defining resistance to arbitrary side information. *CoRR*, abs/0803.3946, 2008.

[14] Daniel Kifer and Ashwin Machanavajjhala. No free lunch in data privacy. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, pages 193–204. ACM, 2011.

[15] Christina Köffel and Peter Wolkerstorfer. Primelife virtual usability laboratory – results of trial 1. *PrimeLife poject, internal report*, 2009.

[16] Jonathan Lazar, Julio Abascal, Simone D. J. Barbosa, Jeremy T. Barksdale, Batya Friedman, Jens Grossklags, Jan Gulliksen, Jeff A. Johnson, Tom McEwan, Loïc Martínez Normand, Wibke Michalk, Janice Y. Tsai, Gerrit C. van der Veer, Hans von Axelson, Åke Walldius, Gill Whitney, Marco Winckler, Volker Wulf, Elizabeth F. Churchill, Lorrie Faith Cranor, Janet Davis, Alan Hedge, Harry Hochheiser, Juan Pablo Hourcade, Clayton Lewis, Lisa P. Nathan, Fabio Paternò, Blake Reid, Whitney Quesenbery, Ted Selker, and Brian Wentz. Human-computer interaction and international public policymaking: A framework for understanding and taking future actions. *Foundations and Trends in Human-Computer Interaction*, 9(2):69–149, 2016.

[17] Jaewoo Lee and Chris Clifton. How much is enough? choosing ϵ for differential privacy. In Xuejia Lai, Jianying Zhou, and Hui Li, editors, *Information Security, 14th International Conference, ISC 2011, Xi'an, China, October 26-29, 2011. Proceedings*, volume 7001 of *Lecture Notes in Computer Science*, pages 325–340. Springer, 2011.

[18] Ninghui Li, Wahbeh H. Qardaji, and Dong Su. On sampling, anonymization, and differential privacy or, k-anonymization meets differential privacy. In Heung Youl Youm and Yoojae Won, editors, *7th ACM Symposium on Information, Compuer and Communications Security, ASIACCS '12, Seoul, Korea, May 2-4, 2012*, pages 32–33. ACM, 2012.

[19] Changchang Liu, Prateek Mittal, and Supriyo Chakraborty. Dependence makes you vulnberable: Differential privacy under dependent tuples. In *NDSS*. The Internet Society, 2016.

[20] Alessandra Mazzia, Kristen LeFevre, and Eytan Adar. The pviz comprehension tool for social network privacy settings. In Cranor [6], page 13.

[21] Robert W. Reeder, Patrick Gage Kelley, Aleecia M. McDonald, and Lorrie Faith Cranor. A user study of the expandable grid applied to P3P privacy policy visualization. In Lorrie Faith Cranor, editor, *Proceedings of the 5th Symposium on Usable Privacy and Security, SOUPS 2009, Mountain View, California, USA, July 15-17, 2009*, ACM International Conference Proceeding Series. ACM, 2009.

[22] Rathindra Sarathy and Krishnamurty Muralidhar. Evaluating laplace noise addition to satisfy differential privacy for numeric data. *Trans. Data Privacy*, 4(1):1–17, 2011.

[23] European Union. Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation), May 2016.

[24] Xiaokui Xiao, Yufei Tao, and Nick Koudas. Transparent anonymization: Thwarting adversaries who know the algorithm. *ACM Trans. Database Syst.*, 35(2):8:1–8:48, 2010.