**Grant Agreement No.: 611001**

# UCN


Instrument: Collaborative Project

Call Identifier: FP7-ICT-2013-10

Objective:

# D1.1: PIH Requirements Document

Due date of deliverable: 30.06.2014

Actual submission date: 30.06.2014

Start date of project: October 1st 2013 Duration: 36 months

Project Manager: Henrik Lundgren, Technicolor SA

Author(s): University of Cambridge, The University of Nottingham, Technicolor

Editor:

Revision: v.1.0

## Abstract

D1.1 is the first of the three deliverables to be produced within WP1 of the User Centric Networking (UCN) project. The overall aim of the UCN project is to take advantage of the information collected about users, to develop content recommendation and content delivery methods.

Central to UCN is a repository for storing all the data collected about users. We envision a repository for each user and since it will contain personal information about the user and will be owned by him or her, we call it the Personal Information Hub (PIH). The main aim of WP1 is the design and implementation of the underlying software infrastructure needed to support the implementation of the PIH. Conceptually, the PIH is a personal permanent storage space, not necessarily centralised, used by individuals for storing their personal data collected by their devices (e.g., mobile phone and smart meters) and sharing, possibly after processing it, with other parties, all under constraints imposed by the individuals. As such, the PIH is expected to offer facilities (APIs) for storing, processing and retrieving personal data.

The PIH needs to satisfy all the constraints imposed by individuals. Thus, it needs to provide mechanisms for enforcing them. One of the most challenging constraints and of particular interest in the UCN projects is privacy. In addition, to make the PIH functional in practical applications, its design needs to meet several additional requirements. The aim of D1.1 is to document our preliminary ideas about these requirements and expose them to public scrutiny with the intention of receiving feedback. Note that in WP1 we are following the agile development methodology, which is based on interactive design that includes actual testing. Thus, the requirements discussed in D1.1 will be tested in preliminary implementations and are very likely to be refined in subsequent documents such as D1.2 and D1.3.

## Table of Contents

# 1 INTRODUCTION

At the heart of the UCN project is the concept of User Centric Networking—a new paradigm leveraging user information at large to deliver novel content recommendation systems and content delivery framework.

The aim of the UCN project is to develop recommendations and content delivery systems that can leverage in-depth knowledge about users and be used, for example, for finding relevant content, identify nearby network resources and plan how to deliver the actual content at the best time and under consideration of the particularities of the users' devices. Within the UCN project, we have identified three closely related but independent research issues: i) understanding of the user context, ii) profiling and predicting user interests and iii) personalization of content delivery.

Understanding the user context involves the collection of data about users' activities, needs and interests. Thus the user context is the basic building block in UCN as it provides the input data needed by the recommendation and delivery methods. In other words, the understanding of the user context involves the collection of data about users and processing it so that it can be used as input to other tasks such as building user profiles.

In the UCN project we define *personal data* as a set of records collected about a given individual. In the project we aim at solutions that are general enough and orthogonal to the specific nature of the personal data. Yet it is worth mentioning that illustrative examples of personal data are the physical location of the user, the device that he or she is currently using, the social network activities of the user (her profile and social graph), the browsing activities of the user (web pages frequently visited) and her consumption records. As explained in D2.1 (Specification of User Context Metrics), in the UCN project we will focus our attention on personal data collected from four vantage points: end-user devices, sensors, home gateways and service provider networks. A list of specific examples of data of interest in provided in D2.1 and includes the location (GPS coordinates) and device configuration (brand, OS, CPU, memory, applications, etc.) of the user, his or her domestic power consumption, home temperature, health records (heart rate, weight, blood pressure, etc.), TV activities (channels and programmes) and social network activities (friends, likes, wall posts, e-mail addresses, etc.).

A particularity of personal data is that it contains very sensitive personal information about individuals. For this reason, most individuals feel uncomfortable about sharing it with the general public. We take this issue into consideration in the UCN project. We will provide users with means for being in control of personal data collected and stored about them.

By "being in control" we mean that the user owns her (or his) personal data and is in position to impose conditions about sharing it with other parties, including other individuals and enterprises. For example, Alice decides with whom to share a specific record (for example, age, marital status, salary, etc.) of her personal data, when and how. By 'conditions' we mean

policies imposed by the owner of the personal data and to be observed by the parties with which the owner shares her personal data. An example of condition is *"Do not reveal that I am on holiday and away from home to non-members of my inner circle of family and friends"*.

Typical examples of enterprises that might be interested in personal data about individuals are banks, entertainment media, supermarkets, advertisers and other institutions that can use personal data for building accurate consumers' profiles and offer personalised services to their customers. A specific example would be advertisement targeted at individuals as opposite to groups of them. As another example, take "*Cortana"*, the personal digital assistant of Microsoft which relies on personal data collected about individuals.

Central to the UCN project is the Personal Information Hub (PIH)--a platform for collecting, storing and sharing personal data. Each individual will own a PIH and be able to decide and impose his or her conditions for sharing the personal data stored in his or her PIH.

One of the first technical problems faces in UCN, and the responsibility of WP1, is the design and implementation of the underlying software infrastructure (for example, databases and communication mechanisms) for building the PIH. In other words, the responsibility of the WP1 is to build the underlying hardware and software components of the PIH and pipelines for interconnecting them, to abstract away implementation details and present neat APIs to the above software layers of the PIH.

We can anticipate that different users will impose different conditions. Yet we argue that security related concerns, such as privacy, are fundamental and likely to be shared among most users. On this basis we take security as one of the driven requirements in the design of the bottom layer of the PIH. In the following sections, we present a conceptual architecture of the PIH and next we discuss the design requirements that should be satisfied by the PIH in order to meet security-related concerns. We also include additional requirements that should be satisfied by the PIH to make it practically functional.

Note that this document (D1.1) is the first one (out of three) to be produced by WP1 where we follow the agile development methodology. The ideas outlined in D1.1 are thus expected to be tested and refined in subsequent documents such as D1.2 and D1.3. The aim of D1.1 is to document and expose to public scrutiny the design requirement that we have identified so far as fundamental of the PIH.

# 2 CONCEPTUAL ARCHITECTURE OF THE PIH

A conceptual view of the architecture of the PIH is shown in Fig. 1. We present an overview of its main components and functionality in this section and elaborate on implementation details and design requirements in subsequent sections.

In the figure, Alice represents an arbitrary user who is interested in taking control of her Personal Data (PD).

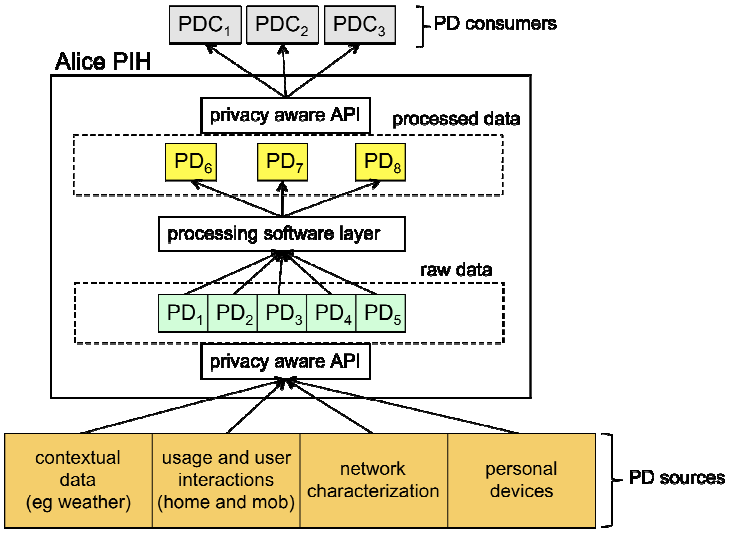


Fig. 1 Conceptual architecture of the PIH.

Conceptually, the PIH is a permanent storage repository where Alice's personal data is stored. The PIH is directly related to PD sources and PD consumers. Both of them access the PIH through privacy-aware APIs.

The PD sources are the producers of Alice's PD. Four examples of potential sources of PD are shown in the figure. Yet the actual nature of the source is irrelevant provided that it can interact with the API that the PIH will provide. We assume that the sources produce records of raw data which are represented by $PD_1$, $PD_2$, $PD_3$, $PD_4$ and $PD_5$. They are raw data in the sense that they are not privacy aware and might contain information (e.g., current geographical location) that Alice does not wish to share with other parties. The PD consumers ($PDC_1$, $PDC_2$, and $PDC_3$) represent the parties with whom Alice has decided to share pieces of her PD from her PIH.

We assume that both the sources and consumers have communication means from depositing and retrieving, respectively, PD from Alice's PIH. Also we assume that both have means (for example cryptography technology) for protecting Alice's PD in transit to guarantee that Alice's security concerns are observed. An example of a personal data collector and user of the API provided by the PIH, are the data collectors to be developed within WP2-Data Collection.

Note that to guarantee the observance of Alice's privacy requirements, the raw data is processed by the *processing software layer* before making it available to the consumers.

# 3 REQUIREMENTS

In our view, personal data represent the electronic life of individuals and is gradually collected along his or her lifetime. Thus we conceive personal data as a set of incremental records that are kept during the individual's lifetime (normally decades) and possibly beyond in the form of historical archives. A particularity of personal data is that it includes highly sensitive information about the individual's private life that he or she is willing to share only with parties of his or her choice or with nobody.

An implementation of the architecture shown in Fig. 1 needs to satisfy a set of requirements to ensure that the observations discussed above are taken into account. The following list includes the requirements that we have identified so far as being fundamental.

1. The user shall be able to store personal data in his PIH incrementally during his lifetime.
2. The user shall be able to retrieve any record stored in the past that has not been explicitly deleted by him from the PIH.
3. The PIH shall provide mechanisms for seamless migration to new technology.
4. The PIH shall guarantee that the user's data is not lost accidentally.
5. The PIH shall offer protection against software and hardware crashes and accidental misuse (e.g. accidental deletion of records or lost of devices or encryption keys) of the user.
6. The user shall be in control of his PIH and be able to impose privacy policies about sharing specific records stored in his PIH.
7. The user shall be able to alter his privacy policies as needed.
8. The user shall be able to delete records from his PIH permanently if he wishes to.
9. The user shall be able to access his PIH for uploading and retrieving personal data, from different devices regardless of their physical and logical locations.
10. The PIH should provide mechanisms for encrypting personal data in transit and on storage.
11. The user shall be able to upgrade the devices he uses to realise and interact with his PIH.
12. The PIH shall offer interfaces to both producers and consumers of personal data.
13. The PIH shall provide means for integrating new devices seamlessly and potentially integrate scores of them with his PIH.
14. The user should be able to store some of his records in storage provided by cloud providers (e.g. Amazon S3) without revealing private data to the cloud provider.
15. The PIH shall provide mechanisms for permanently deleting personal data stored in discarded devices.
16. The PIH shall provide mechanisms that abstract away details and impairments of current devices (e.g., failures, exhausted batteries, etc.) communication technology (e.g. middle-boxes), from the user's view.
17. The user, who is not necessarily an expert in computer technology, shall be able to manage his PIH with little or no technical assistance.
18. The PIH shall provide mechanisms for protecting the PIH (the personal devices for instance) against hackers.

# 4 IMPLEMENTATION ARCHITECTURE

We regard and will implement the PIH as a distributed system that interconnects several components including a large number of personal devices and storage space provided by cloud providers like Amazon S3. We assume that the personal devices are provided with complete computation facilities such as CPU, memory, permanent storage, wireless communication and possibly cryptographic facilities.

A preliminary implementation architecture of the PIH is shown in Fig. 2. The figure shows the pieces of technology we are planning to use to realise the conceptual architecture of the PIH shown in Fig. 1

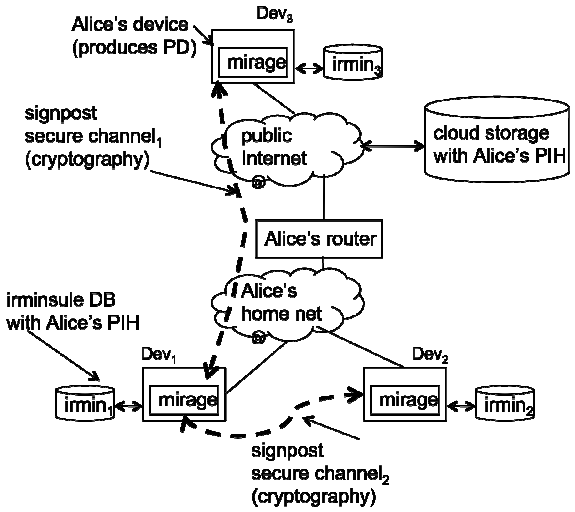


Fig. 2 Implementation architecture of the PIH.

In the figure, Alice is an arbitrary user in possession of several devices and a home network accessed through a home router. Though the figure shows only three devices ($Dev_1$, $Dev_2$ and $Dev_3$), Alice might have scores of them integrated to her home network. The devices are provided with permanent storages that are used for realising the storage repository of the PIH as a distributed database ($irmin_1$, $irmin_2$, $irmin_3$) implemented in irmin. The devices run the *Mirage* operating system and communicate with each though *signposts* secure channels regardless of their logical and physical locations. The devices shown in the figure collect personal data, store it in their local storages ($irmin_1$, $irmin_2$, $irmin_3$) and synchronise their records with each other.

We will explain more details about these technologies and the reasons for using them in the following sections.

# 5 STORAGE OF PERSONAL DATA

To address requirements 1-5 we are planning to use irmin for implementing the storage repository of the PIH. irmin[1] is currently being developed at University of Cambridge and is a library database implemented in OCaml and based on the GitHub paradigm. OCaml is a well documented functional programming language that has proven to be mature enough for implementing commercial and research applications[2]. As shown in[3], it is still under active development at University of Cambridge.

We will use irmin because it provides features that are essential in the implementation of the PIH, namely: 1) distribution, 2) version control and 3) synchronisation. For instance, to reduce the risk of losing Alice's personal data we will distribute and replicate Alice's personal data on several devices; the synchronisation facilities offered by irmin will ease the problem of replica synchronisation; whereas the version control system offered by irmin will help build mechanisms for recovering pieces of personal data accidentally lost.

For convenience of access to consumers of PD, we are planning to deploy replicas of some pieces of Alice's personal data with cloud service providers such as Amazon S3.

We will use state of the art crypto technology to provide privacy. For instance, we will encrypt personal data stored in the PIH and on transit (explained below). We will also use crypto technology to encrypt personal data stored within cloud providers. We will take advantage of crypto-facilities offered by some storage providers such as client or server side encryption offered by Amazon to S3 customers.

The distributed approach provided by irmin will help us address issues related to accidental lost such as issue number 4. We reduce the risk of losing personal data by means of using several devices to keep replicas of the personal data.

# 6 COLLECTION OF PERSONAL DATA

Requirement number 10 covers the integrity, authenticity and privacy of the personal data in transit. We will use state of the art crypto technology to address these issues. For instance, personal data produced by the sources will be encrypted in transit.

Requirement number 9 demands that all devices can seamlessly communicate with each other to provide the personal data they collect and to synchronise their local repositories with each other, regardless of their physical and logical locations.

Seamless communication is not a trivial requirement to satisfy with current Internet due to the existence of middle boxes (such as firewalls and NATs, and load balancers) that are frequently deployed between communicating entities and are known to interfere with their communications. For instance, some of them alter IP addresses, port numbers and even payloads. A recent discussion of the annoyances caused by middleboxes is presented in [4].

We will address requirement number 16 with the help of *signposts* ---a network service being developed at University of Cambridge[5] to provide ubiquity, reachability and security. As demonstrated by an early prototype[6], signposts helps an individual integrate his personal devices into a personal network where devices communicate with each other over secure channels regardless of their physical and logical locations.
In this order, we will use signposts to help Alice create a personal network of her personal devices that are responsible for collecting and contributing personal data to Alice's PIH.

It is worth emphasizing that signposts takes advantage of the functionalities provided by DNSSEC to automate the process of creating and maintaining the secure communication channels[7]—in this way, signposts contributes to meeting requirement number 17.

# 7 INSTRUMENTATION OF PERSONAL DEVICES

Requirement 18 is related to the potential security vulnerabilities of the software deployed in the devices responsible for collecting and storing personal data. It is well known that hackers exploit different vulnerabilities to compromise systems. Yet it has been widely documented that vulnerabilities related to memory-overwriting stacks are the most routinely exploited by hackers[8].

We will address this issue at operating system level by deploying the Mirage operating systems in the devices. Mirage is a unikernel operating system being developed in the OCaml functional language at University of Cambridge[9]. As explained in [10], Mirage offers several advantages. For instance, the strong static type checking provided by OCaml eliminates the risk of memory-overwriting. In addition its small size makes it very suitable for deployment on small devices such as home appliances such as sensors.

# 8 CONCLUSIONS

This D1.1 delivery is part of WP1 (responsible for the design and implementation of the PIH) and outlines the requirements that the implementation of the PIH is expected to satisfy. It is meant to help the designer and implementers of the architecture of the PIH. It is the first out of three documents that will be produced by WP1. Thus the requirements discussed in D1.1 are expected to be tested and refined in subsequent documents.

## REFERENCES

[1] Irminsule, nymote.org/software/irminsule

[2] OCaml Labs, http://www.cl.cam.ac.uk/projects/ocamllabs/

[3] A. Madhavapeddy, Y. Minsky and J. Hickey, *Real World OCaml: functional programming for the masses,* O'Reilly Associates, 2013.

[4] C. Paasch and Olivier Bonaventure, *"Multipath TCP"*, Queue, V.12, Issue 2, Feb 2014.

[5] Signposts nymote.org/software/signposts

[6] A. Aucinas, A. M. Chaudhry, J. Crowcroft, S. P. Eide, S. Hand, A. Madhavapeddy, A. W. Moore, R. Mortier, C. Rotsos and N. Vallina-Rodriguez, *"Signposts: end-to-end networking in a world of middleboxes"*, In Proc. ACM SIGCOMM'12 Conf. on Applications, Technologies, Architectures, and Protocols for Computer Communication, Helsinki, 2012.

[7] C. Rotsos, Heidi Howard, D. Sheets, R. Mortier, A. Madhavapeddy, A. Chaudhry and J. Crowcroft, *"Lost in the Edge: Finding your Way with Signposts"*, In Proc. IEEE Symposium on Foundations of Computational Intelligence (FOCI 2013).

[8] R. Anderson, *Security Engineering,* John Wiley and Sons, 2001.

[9] A programming framework for building type-safe, modular systems, www.openmirage.org

[10] A. Madhavapeddy and D. J. Scott. "*Unikernels: Rise of the Virtual Library Operating System*" acmqueu, Jan. 2014.