Deliverable 6.5

Mid-term evaluation report for Data Management Platform including PIA

Grant Agreement number:	688082
Project acronym:	SETA
Project title:	An open, sustainable, ubiquitous data and service ecosystem for efficient, effective, safe, resilient mobility in metropolitan areas
Funding Scheme:	H2020-ICT-2015

Authors

Rafał Janik	rafal.janik@softwaremind.com
Marcin Sieprawski	marcin.sieprawski@softwaremind.com
Marcin Kwasnik	marcin.kwasnik@softwaremind.com
Andrzej Boruch	andrzej.boruch@softwaremind.com

Internal Reviewer

Neil Ireson	n.ireson@sheffield.ac.uk

State:	v.1.0
Distribution:	Confidential

Deliverable History

Date	Author	Changes
03 July 2017	Rafał Janik	Initial version
04 July 2017	Rafał Janik	Evaluation process description
07 July 2017	Marcin Sieprawski	Evaluation process description
10 July 2017	Andrzej Boruch	First set of load tests results
13 July 2017	Marcin Kwasnik	Cluster configuration
14 July 2017	Rafał Janik	Scope and background sections.
17 July 2017	Rafał Janik	Evaluation results section.
18 July 2017	Marcin Sieprawski	Minor changes
19 July 2017	Andrzej Boruch	Evaluation results section.
21 July 2017	Rafał Janik	Recommendations added
21 July 2017	Marcin Sieprawski	Internal Review
22 July 2017	Rafał Janik	Minor Changes
25 July 2017	Neil Ireson	Internal Review
28 July 2017	Rafal Janik	Addressing comments form Internal Review
31 July 2017	Marcin Sieprawski	Final version

Contents

1. Summary

WP6 attaches great importance to the evaluation of cloud-based SETA Data Management Platform, to ensure robust, smooth and scalable collection, indexing, manipulation and sharing of heterogeneous, multimodal, dynamic mobility data. Constant evaluation of functionality and technical efficacy of the SDP was built into agile software development process from the beginning of the project. In addition to that formal evaluation of the Platform is carried out at two points within the project.

This document presents the results of the execution of a suite of functionality, performance and scalability tests of SETA Data Management Platform performed on mid-term formal evaluation (M16-M18). The document presents the evaluation process, description of the environment end results of evaluation tests, followed by recommendations from evaluation.

This deliverable contains also periodic Privacy Impact Assessment (PIA) analysing privacy and data protection risk for mid-term evaluation - submitted as a separate document.

2 Glossary of Terms

API	Application programming interface
CPU	Central Processing Unit
DB	Data Base
HDD	Hard Disk Drive
HDFS	Hadoop Distributed File System
IO	Input / Output operation
ITS	Intelligent Transport Systems
JAX-RS	Java API for RESTful Web Services
JSON	JavaScript Object Notation
JVM	Java Virtual Machine
MB	Mega Bytes
ms	milliseconds
PIA	Privacy Impact Assessment
RAM	Random Access Memory
RDBMS	Relational DataBase Management System
REST	Representational State Transfer
SDP	SETA Data Management Platform
SOA	Service Oriented Architecture
SQL	Structured Query Language
SSD	Solid State Drive
stddev	Standard Deviation
VMware	Hardware and Operating System Virtualization Software
WP	Work Package
WP6	Work Package 6

3 Structure of the document

This document will write up mid-term evaluation of SETA Data Management Platform. Section 4 will describe reasons why the evaluation was performed. Next in section 5, will describe what parts of SETA Data Management Platform (SDP) were evaluated. Description of the steps taken during the evaluation and listing tools and resources used will be a part of section 6. Section 7 will discuss evaluation process execution and results. Recommendations - Section 8 - will sum-up evaluation and suggest a way, how SETA Data Management Platform should be developed during upcoming months.

4 Background

WP6 has to deliver a scalable, efficient, expressive, and continuously available spatial Big Data Management Platform and interfaces to communicate to the data and support analytical queries over large volumes of spatio-temporal data.

This deliverable presents the results of execution of preliminary suite of functionality, performance and scalability tests which address the technical efficacy of the SETA Data Management Platform.

Outcome of this document will be used to address specific improvement actions in second development phase. This deliverable also provides an updated Privacy Impact Assessments (PIA).

5 Scope

The evaluation of the SETA Data Management Platform audits the performance of chosen solutions and covers the initial requirements by platform functionalities. As WP6 manage big data coming from project partners and open data sites, evaluation is based on a large-scale data gathered within the SDP.

The complete description of the data that SDP will store and manage was described in Deliverable 6.2.1 *Requirements and architecture of Data Management Platform including PIA.* The data items and related API calls can be divided into following groups:

- Static sensors and static measurements - spatio-temporal data, static sensors are the sensors with fixed localisations.
- Mobile sensors and mobile measurements, such as mobile applications, sensors located in vehicles - sensors whose localisation surely will be changing - a spatio-temporal data. Measurements produced by such sensors, compared to those of static sensors, include localisation.
- Cities' road networks - this subset provides information about the whole area as a set of nodes and edges - road infrastructure graphs.
- Facilities - this group is a set of diverse objects - spatial data - places, buildings and the like - that has an impact on traffic and/or set the metropolitan topology and network

After the set of consultation with other WPs, WP6 has chosen a set of different kind of queries which will simulate the most popular and expensive (from the server resources point of view) queries. Final scope of evaluation was agreed and it contains end-to-end load tests of storing and retrieving a huge amount of data.

6. Evaluation process

The evaluation focuses on the execution of a suite of functionality, performance and scalability tests which address the the technical efficacy of the SETA Data Management Platform.

6.1 Evaluation as a part of agile software development process

This evaluation process is mainly a built-in part of SDP development and it started from the beginning of the project. All SDP functionalities are covered by unit and end-to-end tests. New features are announced to SETA community and also evaluated by them. WP6 uses Scrum - an Agile framework - which supports the evaluation process; new features are provided in two-weeks sprints so all updates of the Platform functionalities may be tested almost immediately. More details about this process can be found in section 6 of deliverable D6.4 Initial prototype of SETA Data Management Platform.

It is important that WP6 monitors most of the SDP metrics for evaluation purposes. The server farm is continuously analysed against the resources usage.
SETA Data Management Platform monitoring provides statistical gathering collectors for chronological evaluation of CPU, RAM, HDD/SSD storage Input/Output and load average Operating System metrics coming from hundreds of prepared monitoring jobs. The more detailed description of used tools and the whole process was described in D6.4 subsection 6.5. Based on the real-time and historical data provided by monitoring tools, WP6 continuously evaluate SDP by tuning VM settings. SDP settings and resources allocation. The SDP can be easily reconfigured during evaluation thanks to cluster management tools described in D6.4 section 6.7.

6.2 Mid-term evaluation process

As evaluation happens all the time, this deliverable is focused on specific part of evaluation - the mid-term performance of SDP which is crucial from the point of view of robustness and high-availability. From now we will use ‘evaluation’ term for describing those mid-term tests of SDP services saving, indexing, processing and retrieving spatio-temporal data.

SETA Data Management Platform evaluation is mainly focused on benchmarks of data accessibility and availability as well as on performance of data uploading. WP6 has performed a number of different set of scenarios imitating a final usage of the Platform and in each measured response times of queries.
The evaluation process was splitted into two parts, first one including availability of data storing within the SDP, when Platform was loaded with the huge amount of incoming data.
The second phase was mainly focused on querying SDP heavily with the various types of query parameters. During both phases, WP6 performed implementation improvements and environment tuning and acceleration.
What is important to mention, during all evaluation, all communications with SETA Data Management Platform performed by evaluation executors, were sent through SETA API (described in D6.4 Initial prototype of Data Management Platform), so all evaluation scenarios cover the way of usage the platform by other SETA WPs (Fig.1.).

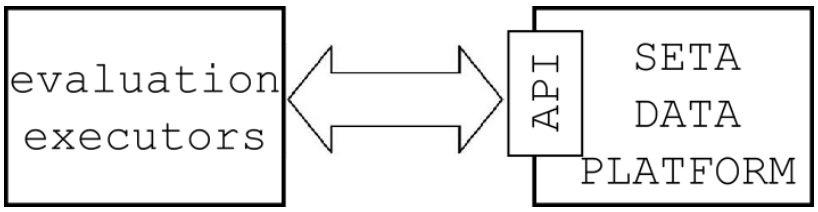

Figure 1. Evaluation via API

Evaluation of SDP has been divided into evaluation of two main subsets of operation:

- Data loading and indexing - tasks performed by specified evaluation executors - stream players (described in 6.3)

- Retrieving results/analysis - executed by jobs called evaluation clients

On both subsets WP6 performed load and stress tests to determine SETA Data Management Platform robustness, availability, and error handling under a heavy load. Load testing procedure – exerts extremely high service pressure on software platform or computing device – and as a result – collecting appropriate software platform response measurements. Load tests are performed so that a baseline software platform behavior gets established - under both normal and anticipated peak loads. Additionally, they serve to identify maximum and minimum - operating capacity of an application. Furthermore, any potential bottlenecks are narrowed down as well as any software platform element causing service degradation.

Usually, stress testing - which are performed during load tests – place heavy load on the system, beyond normal usage patterns - to provide the system's response at unusually high or peak loads.

In SETA project it's important to load and analyse data quickly as huge amount of data comes almost every second. Here, in this evaluation, based on the most frequent queries WP6 has run a set of designed and dedicated benchmarks to measure SETA Data Management Platform performance by analyzing response times. Work Package 6 mission is to accommodate a reliable, big-data, capable of querying, low-latency cloud-based SETA Data Platform – and the evaluation was prepared to check if low latency is a domain of SDP.

6.3 Stream player - real time data

The first set of load tests scenarios depends on testing SETA Data Management Platform data consuming performance. To simulate a real-time load of the Platform feeded with the huge amount of measurement coming from various sensors, WP6 has implemented additional components - stream players (Fig.2.). Idea of stream players was developed during ROBUST[1] project and consists on pushing the historical data, one by one, into a platform as a real stream of measured sensors' values. Usually measurements feed the platform in one minutes intervals. What is important here one instance of stream player in this evaluation calls SDP API about 15,000 per minute so it simulates many clients. One

[1] http://www.robust-project.eu/

stream player instance may be considered as 15,000 of concurrent SDP clients. In this evaluation, WP6 checks SDP load low-latency by increasing the number stream players.

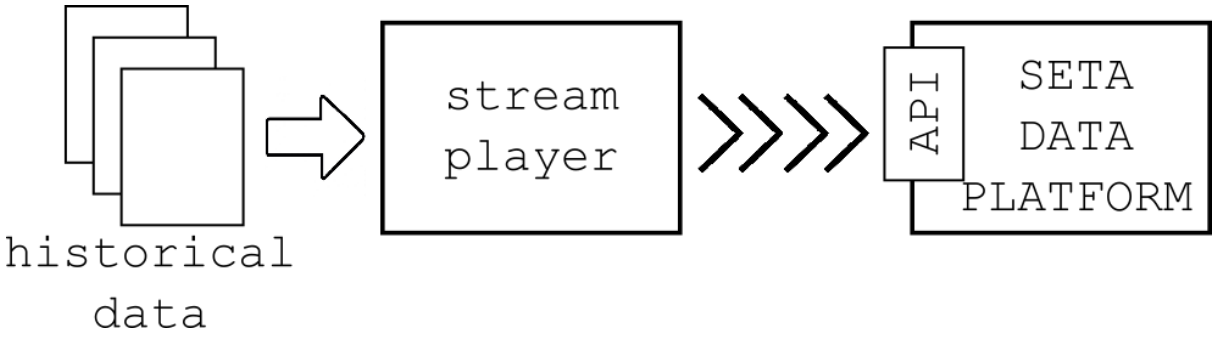

Figure 2. Idea of stream players.

6.4 Randomised queries

Second set of scenarios depends on running different randomized queries on data managed and processed within the SETA Data Management Platform. Randomizing the queries is a very important aspect of performed tests, that ensures the results are not cached, so the actual query execution time is being assessed.

7. Evaluation

Evaluation was performed against SETA cloud based environment. Detailed description of Platform architecture may be found in D6.4 section 7. For a reliably evaluation it's crucial to set-up platform environment properly. Unappropriate configuration may interfere results.

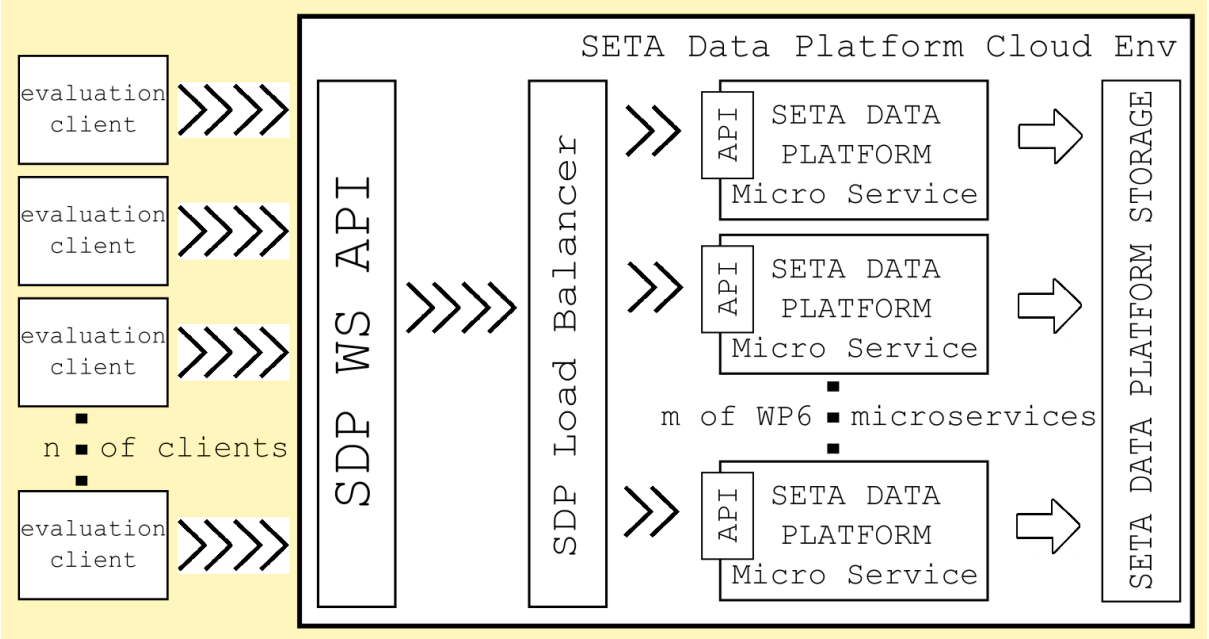

Figure 3. Evaluation architecture.

Evaluation executors, utilized by SETA WP6 is a set of distributed stream players[2] and queries clients - implemented to interpret and send into the Platform a huge number of various kind of data, as well as executors of SDP load tests. Those executors make a huge set of simple RESTful calls on the SDP application - for the purpose of transmitting the split load of hundreds of millions of measurements (Fig.3.).

7.1 Evaluation environment configuration

SDP WS API is a Platform web-service entrance - RESTful interface - exposed for the purpose of handling incoming calls from other WPs, in case of evaluation also the evaluation executors.

To improve SETA Data Management Platform performance, WP6 introduced load balancing solutions.

Load balancing refers to efficiently distributing incoming requests across a set of backend servers, also known as a server farm. Each server in the server farm hosts one instance of a SDP microservice, so that multiple request may be handled at the same time.

[2] see section 6.3 above

As a load balancer WP6 employed software solution High Availability Proxy (HAProxy[3]). HAProxy was chosen as it has a proven track record and provides historical performance feats.

SETA application utilizes microservice architecture within the JAVA programming language realm. All data processing and computations take place here so Virtual Machines hosting SDP application are built on a fast SSD disk. SSD disks improve filesystem I/O operations performance.
As described earlier, SSD disks offer much higher throughput for data storage than regular HDD partitions. Ample amount of RAM made available for microservice architecture - provides for a smooth operational environment within the SDP application cluster.

SDP Storage - shown in the image below (Fig.4.), involves Hadoop HDFS cluster with HBase and OpenTSDB nodes. RESTful microservice interface exposes methods that process, validate and enriches incoming measurements - and - stores them in the HDFS through the HBase and OpenTSDB node.
10 Gigabit Ethernet - built into the SETA WP6 infrastructure - provides for: an near real-time system response and data transfer.

[3] http://www.haproxy.org/

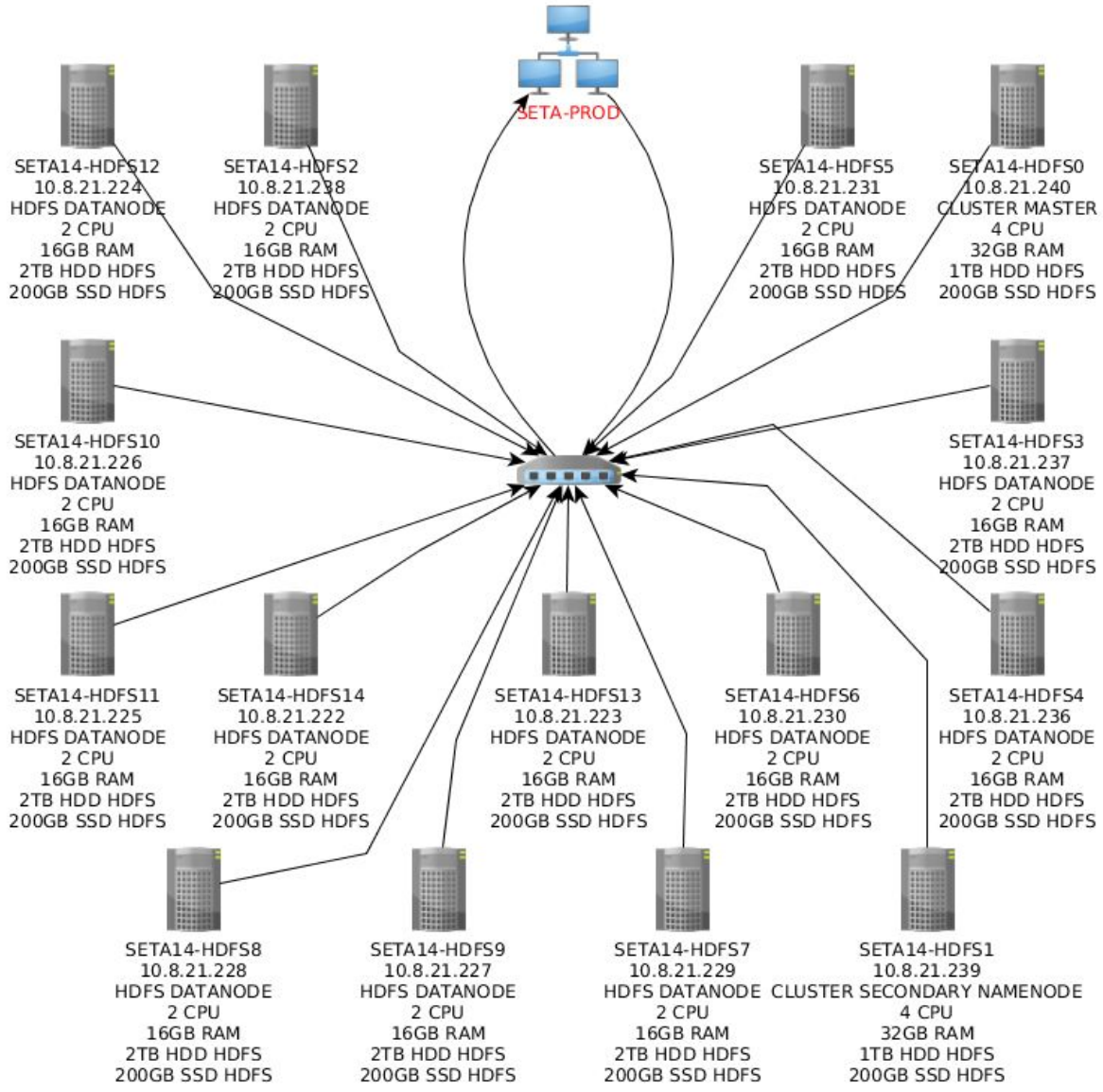

Figure 4. Diagram of evaluated architecture backend.

Seta storage cluster presented above was prepared for mid-term evaluation. To achieve better performance results it will be extended with new servers as well as with new resources and tweaked during a second phase of the project.

7.2 Static sensors' data evaluation

Evaluation scenarios of loading static sensor data into a platform simulated a daily stream of 130 GB, corresponding a typical constant load from 450,000 sensors/sources. In this test SDP was feeded with the continuous stream of requests with measurements to be loaded and indexed. WP6 used stream players[4] to simulate a real time stream of constant data coming into platform from external sources. Each stream player simulated 15,000 clients (sensors/sources): as typical sensors (and other sources like mobile apps) provide max. one new data item/measurement per minute[5], a stream player sent 15,000 measurements each minute, each measurement was sent in a separate API call (one measurement every 4 ms). Stream players took real measurements from historical dataset, the data was related to static sensors and included information such as traffic load, occupancy, average speed, etc.
WP6 has tested SDP here against increasing number of concurrent evaluation executors (stream players). In this set of tests number of stream players was increased from 1 to 30 during evaluation (15,000-450,000 clients) to check how time of full stack of storing data fluctuates (Fig.5.).
Data Management Platform contained 6 microservices attached.

Number of parallel stream players	Median [ms]	Standard deviation [ms]
1	35.98	1.43
3	36.81	1.91
6	37.17	1.05
11	42.68	2.00
16	50.58	3.06
24	55.19	8.73
30	75.81	8.71

[4] See 6.3
[5] Most sources provide 4-12 measurements per minute

Table 1. Saving static data results.

What can be remarked here is that the results were stable until 25 stream players (375,000 clients), after that response time slow down to about 70ms for 30 stream players (450,000 clients) .

Further SDP extensions in second phase of the project will allow to increase this value.

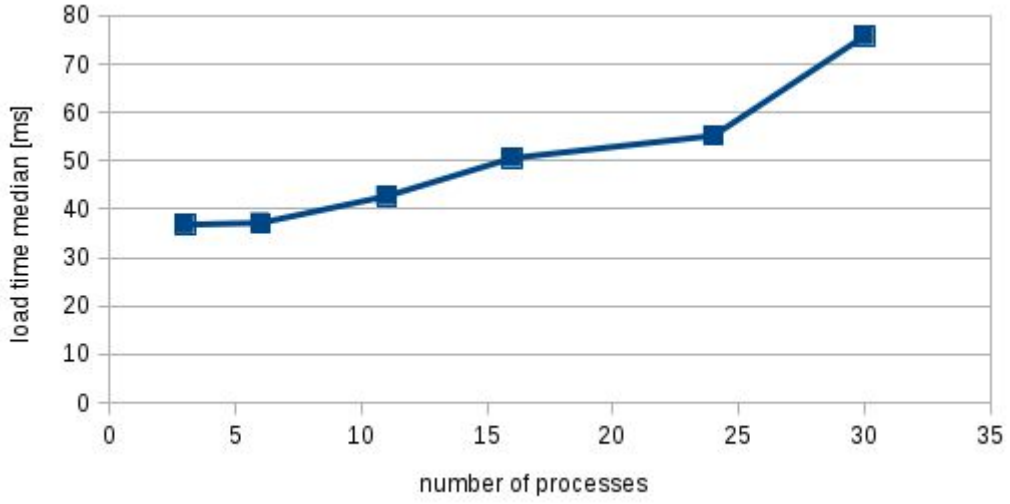

Figure 5. Load time median as a function of parallel processes.

To illustrate a distribution of results WP6 has also run tests to show histogram of data loading times (Fig.6.) for set of 5 millions of API calls from each of 2 stream players at the same time one by one. The size of the each request was about 200B. Results are presented in Figure 6.

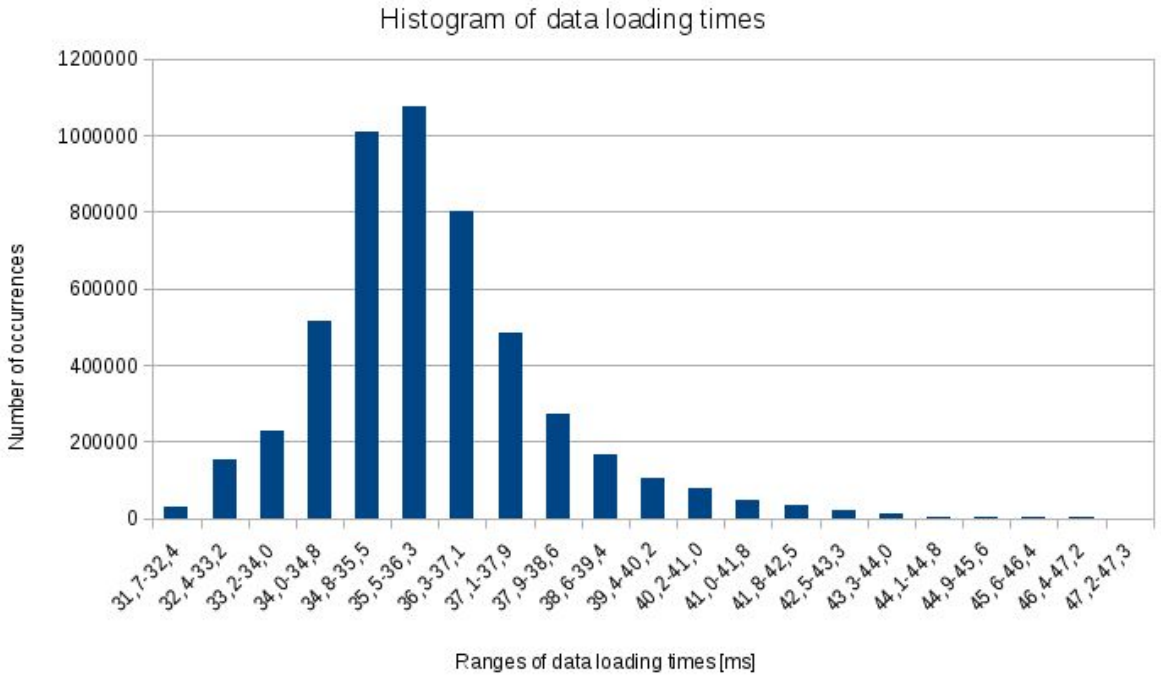

Figure 6. Histogram of data loading times.

The histogram (Figure 6.) shows that loading time is stable and predictable as well as most of results are located in small ranges near a minimal response time: loading time at 95 percentile is 39.86 ms (10% more than median), and only 0.001% of results exceeded the value 47.3 ms.

7.3 Mobile sensors' data evaluation

Next to the data related to static sensors, SETA Data Management Platform was feed with the stream of mobile measurements. The total number of measurements (and - what is equal - the number API calls) was 400 000 000. This is a set of historical data of bike positions, floating cars and buses. First iteration of evaluation showed that during the process of registering mobile sensor, checking sensor information (this operation was performed to check if sensors has been already registered in the platform) and saving the data, the saving of data was taking a inordinate amount of time (Table 2.).

Operation	Response time median [ms]
Registering sensor	31.47
Saving measurement	372.51
Checking sensor status	11.03

Table 2. Saving mobile data results - 1st iteration.

The long response time is caused by matching the geospatial data to the network graph. As was discovered this operation took about 95% of the whole time of request processing.
WP6 decided to reimplement this part of functionality and move more computation into RAM memory to speed-up a time of saving new mobile measurements. After that WP6 has repeated this set of tests with the same input data.

Number of parallel stream players	Median [ms]	Standard deviation [ms]
2	15.93	1.88
3	17.18	1.97
5	18.67	2.01
12	20.32	2.08

Table 3. Saving mobile data results - 2nd iteration.

After moving a part of computations into memory, SDP response time improve and is stable (Table 3.). Also an increasing number of concurrent stream players has no major influential impact on response times. SETA Data Management Platform stayed stable and reliable during this set of evaluation executions.

7.4 Querying static data

SDP was also tested against a continuous stream of requests querying for a measurements from static sensors. In this scenario evaluation querying clients asked SDP for randomly

chosen sensors and time intervals in each call to make sure that queries results are not cached.

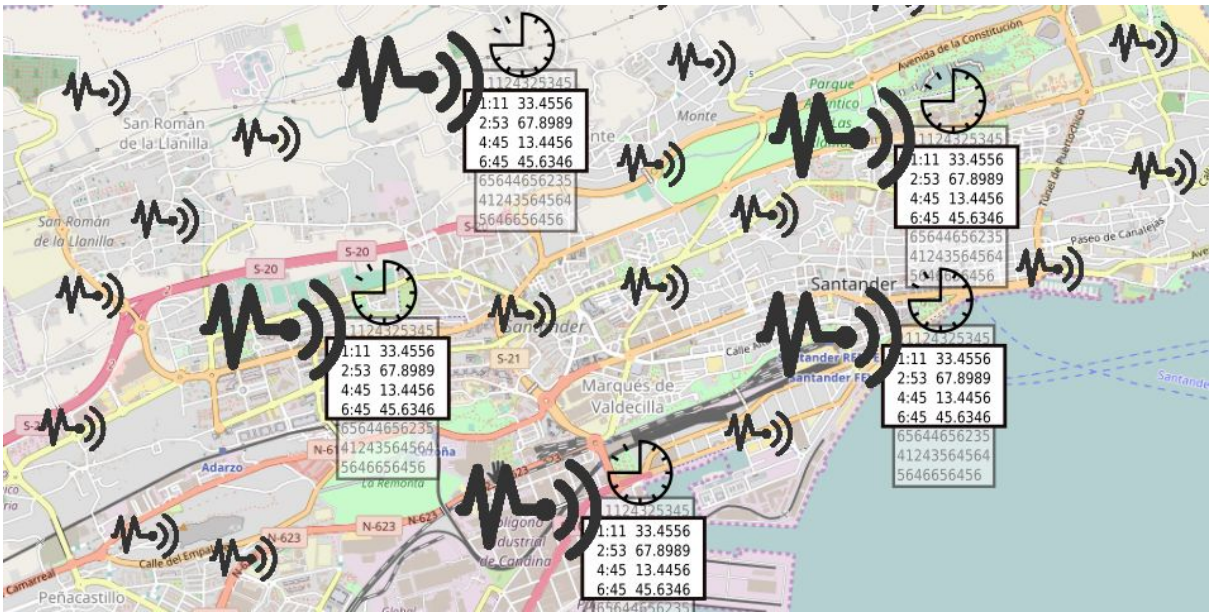

Figure 7. Querying for measurements from random sensors.

Total number of queries was 32 000,000 splitted into n evaluation clients running at the same time, where n was 10, 20 and 100. Average size of returned response was between 520 - 650 of chars.

Number of parallel clients	Median [ms]	Standard deviation [ms]
10	44.87	3.5
20	72.50	9.16
100	152.13	47.49

Table 4. Querying static data results.

The results (Table 4.) show that response time even for 100 concurrent clients at the same are under 0.2s (where response times under 1s are acceptable for human-computer

interactions[6] - SDP query mechanism serves as a backend for WP5 applications). For ten concurrent evaluation clients standard deviation is really small which means that response times are very stable.

7.5 Querying mobile data

WP6 has performed a number of tests on querying the SDP for spatio-temporal data from mobile sensors. This data - mentioned in 7.3 - comes from mobile applications, floating cars, cities' bicycles, environmental sensors etc. In this scenario SDP was tested against a 32 000 000 requests querying for a measurements from mobile sensors. Evaluation executors set requests to the SDP for randomly chosen area and time slots to make sure that queries results are not cached (see Fig.8.).

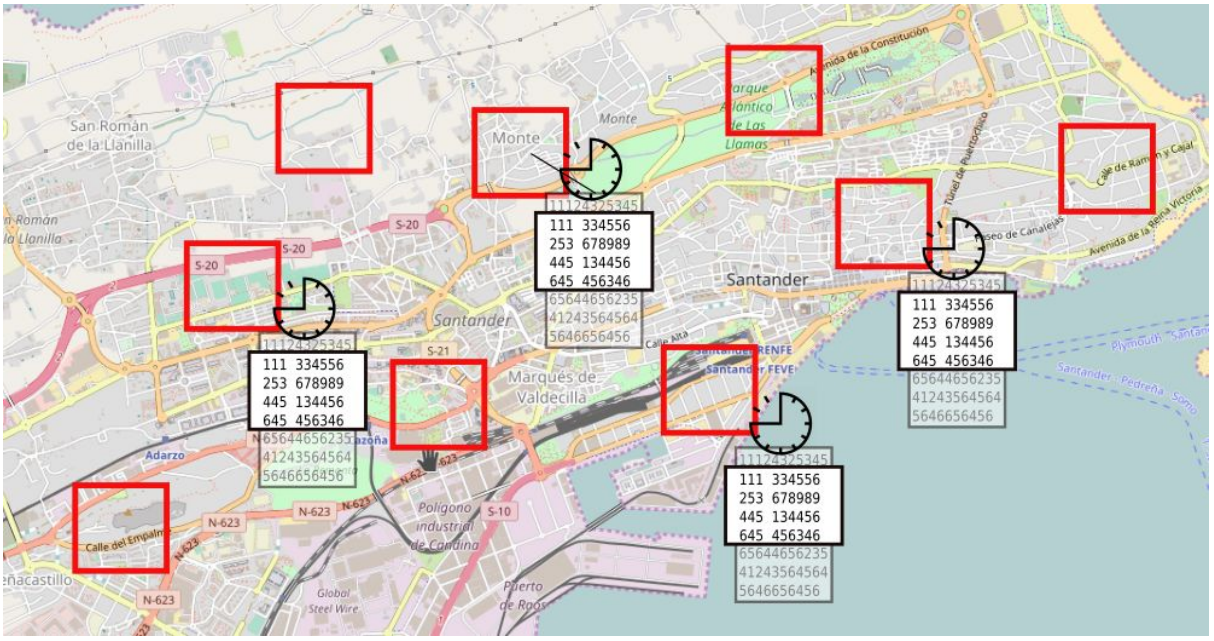

Figure 8. Querying for mobile measurements in requested area 320mx320m and specific for each area time period.

An area used in search process is a square which size increase in next iterations (see Fig.8. and Fig.9.):

- 0.004 radians - about 320m 320m

[6] https://www.nngroup.com/articles/response-times-3-important-limits/

- 0,008 radians, - about 640m x 640m
- 0,016 radian - about 1280m x 1280m.

The size of area is represented in a table by "area side length[radians]" - the bounding box of specific size is chosen randomly, the test is repeated for various sizes and various number of executors. The location of the square is based on existing centroids. Time interval was fixed length set to 1 hour but the start time was chosen randomly.

Number of parallel clients	Area side length[radians]	Median [ms]	Standard deviation [ms]
10	0.004	51.6	42.1
10	0.008	69.7	49.2
10	0.016	77.5	53.4
50	0.004	502.7	278.8
50	0.008	608	301.9
50	0.016	757.9	412.9
220	0.004	1176.5	304.5
220	0.008	1380.7	383.3
220	0.016	1902.6	491.25

Table 5. Querying mobile data results.

Results presented in Table 5 show that the response times for larger set of clients is too high. Despite this, number of clients during overall SETA evaluation won’t be so big so SDP will handle the process.

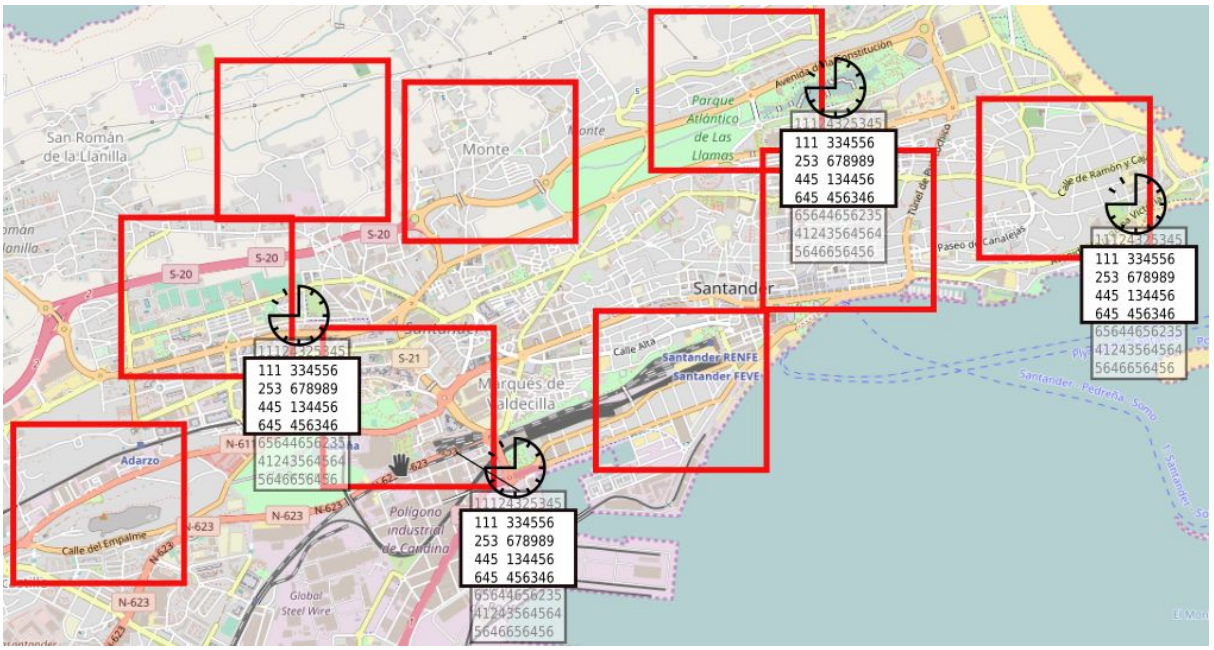

Figure 8. Querying in bigger, random chosen areas .

8. Recommendations

This evaluation shows that SDP performance is good but not linear for increasing number of evaluation executors running. SETA Data Management Platform scalability should be closer to linear function. Another conclusion coming from evaluation is the way of development and maintenance is very good, allows to extend and scale-up SDP and should be continued in the same format. Gathered recommendations just complement existing process.

As a result of evaluation process, WP6 gathered a following list of recommendations:

- In-memory computation - based on 7.3
- Improvement of connection between Persistence Layer and microservices - based on 7.2, 7.3 7.4 and 7.5
- Tuning servers settings - based on 7.2, 7.3 7.4 and 7.5

- Querying and indexing acceleration (GPU) - based on 7.5

which are briefly described in following subsections.

8.1 In-memory computation for crucial elements of request processing

After initial tests in the evaluation process WP6 reimplemented a part of SETA Data Management Platform computation and moved a part of data processing into RAM which resulted in a big performance improvement. The recommendations here is to review existing algorithms in order to perform computation in-memory as often as possible. Some important structures have volumes enabling storing them in RAM of typical processing node (eg. road infrastructure graph for metropolitan areas) and can be easily scaled up (in case of larger areas).

8.2 Improvement of connection between Persistence Layer and microservices

This evaluation shows that response times for increasing number of evaluation executors is getting longer. Adding new SETA microservices improves the performance but there is also another place which might be a potential bottleneck. Investigation performed during evaluation figured out that persistence layer had responded longer, as opposed to microservices layer, where processing time was stable and constant. The recommendations regarding this issue is to scale out SETA Data Platform Storage (see Fig.3.) and increase a number of machines of storage cluster (see Fig. 4), and improve the scalability of the connection between microservices and Persistance Layer. This will increase platform performance and reliability and ensure better data availability for microservices layer.

8.3 Tuning servers settings

Evaluation has shown that hardware tuning is important. WP6, thanks to chosen solutions (described in section 6,1), was able to smoothly improve SDP performance after each iteration of the tests by changing servers' resources as CPU, RAM as well as replacing HDD with SSD where disk IO operations were intensive.

Thanks to the evaluation we also know that the proper environment configuration depend on the implementation so the servers configuration should be monitored and verified after each

release of SDP. Figure 9 shows how memory utilization grown during a set of evaluation tests. WP6 was able to react and add more RAM. In the same way other metrics were monitored. Monitoring will allow to tune the SDP properly.

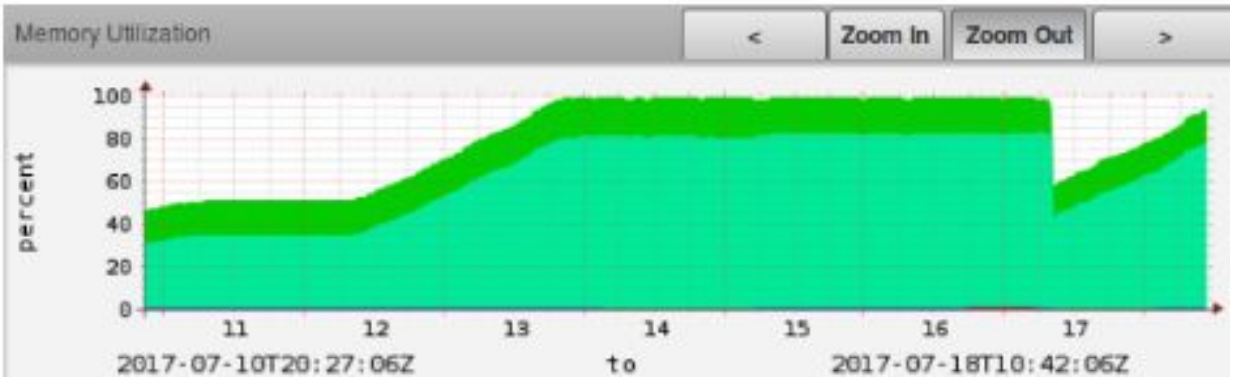

Figure 9. Memory utilization monitoring during a part of evaluation .

8.4 Acceleration of spatio-temporal querying and indexing

The evaluation tests for spatio-temporal queries against mobile sensors data showed that response times for querying the SDP are quite good for small and medium number of parallel clients (below 0.1s), but they quickly grow with increasing number of clients and queries per second. Those operations are CPU-intensive - which was confirmed by WP6 cluster resource management and monitoring tools during the tests. WP6 will employ GPU for accelerating spatio-temporal queries and indexing - for example new algorithms for data indexing based on R-trees will be implemented to make use of GPU capabilities. R-trees are popular spatial indexing techniques that have been widely used in many geospatial applications. GPU's highly improve performance of R-trees algorithms[7].

[7] http://www-cs.ccny.cuny.edu/~jzhang/papers/rtree_tr.pdf

Deliverable PIA (part of D6.5)

Privacy Impact Assessment (Version 2)

Grant Agreement number:	688082
Project acronym:	SETA
Project title:	An open, sustainable, ubiquitous data and service ecosystem for efficient, effective, safe, resilient mobility in metropolitan areas
Funding Scheme:	H2020-ICT-2015

Authors
Vitaveska Lanfranchi v.lanfranchi@sheffield.ac.uk

Internal Reviewer
Marcin Sieprawski marcin.sieprawski@softwaremind.pl

External Reviewer
Anna Donovan, anna.donovan@trilateralresearch.com
Trilateral Research Ltd

State: FINAL
Distribution: Confidential

Deliverable History

Date	Author	Changes
15 July 2017	V. Lanfranchi	Initial version
21 July 2017	V. Lanfranchi	Adding contributions from partners
28 July – 31 July 2017	Trilateral Research Ltd.	Review
31 July 2017	V. Lanfranchi	Addressing reviewers' comments and final version

Contents

1. Summary

The following document provides details of the revised Privacy Impact Assessment (PIA) carried out for the Seta project at the end of Month 18 (Phase 1). This is the second iteration of the PIA.

As the PIA has been defined by David Wright as "a process which should begin at the earliest possible stages, when there are still opportunities to influence the outcome of a project. It is a process that should continue until and even after the project has been deployed."[1], SETA has decided to undertake the first PIA during the requirements phase then revise it at Month 18 after the first evaluation feedback.
The main changes, analysed in Section 3 and Section 4, compared to the initial PIA are related to:

- Introduction of a new data provider partner, 5T. more details are provided in Section 3.3.
- Change to the implementation of the mobile app, that now asks (optionally) the user for age, weight and gender, to calculate calories consumed
- A change in the technical requirement that meant one of the partners, The Floow, was asked to provided a dataset with higher risks of identifying individuals. This was discussed with The Floow and it was decided to not proceed, due to the risk being too high (see Section 4.4).

In SETA, the PIA is used to proactively identify and analyse the most relevant privacy and data protection risks as well as to suggest and monitor associated solutions or mitigation strategies to avoid, minimise, transfer or share these risks.

This deliverable presents the updated privacy risks on SETA.

2. Glossary of Terms

Annex I	*Otherwise known as the DoW*
CA	*Consortium Agreement*
DoW	*Description of Work*
DPA	*Data Protection Act*
GA	*Grant Agreement*
GDPR	*General Data Protection Regulation*
ICO	*Information Commissioner's Office (UK)*
PIA	*Privacy Impact Assessment*
WP	*Work Package*

[1] Wright, D. (2012) 'The state of the art in privacy impact assessment' in Computer Law and Security Review, 28, pp.54-61

3. The process

SETA has committed from the beginning of the project to use a Privacy by Design methodology (defined in WP6 'Management of large scale date for large scale pervasive smart mobility' and presented in D6.2) and to implement the PIA process to identify and analyse the most relevant privacy and data protection risks as well as associated solutions or mitigation strategies.

In SETA we have adopted a twofold approach to privacy and data protection based on:

- Constant monitoring of relevant legislation
- PIA: The purpose of the PIA is not legal compliance. Rather, the PIA is designed to ensure that privacy and data issues are sufficiently covered to respect an individual's right to privacy.

3.1. Legislation Monitoring

SETA is constantly monitoring national and European legislation about privacy and data protection according to the following process, led by the Coordinating Partner (USFD):

- Each partner has been asked at the beginning of the project and every 3 months (before each project meeting) to contact their National Privacy and Data Protection Authority and/or their Legal office to verify the project compliance with national rules
- USFD is constantly reviewing EU legislation (in cooperation with the University Legal Office) to ensure that all data processing and storage will be carried out in accordance with the General Data Protection Regulation (GDPR) as this is the most up to date and relevant piece of legislation in the area of data protection in the EU. The project is also monitoring other legislation where appropriate[2]
- Any relevant changes in GDPR or national legislation are discussed in the Project Steering Committee Meetings and is necessary, the requirements for the project's data anonymisation, data protection and privacy protection are updated.

3.2. PIA

The project has chosen a systematic approach to privacy analysis, by conducting the PIA alongside key phases of the Seta methodologies and technologies design and development.

[2] *Including the Convention for the Protection of Human Rights and Fundamental Freedoms (in particular Article 8); the Charter of Fundamental Rights of the EU; the EU Directives 95/46/EC and 2006/24/EC; Article 29 Working Group's Opinion 8/2010; as well as applicable national data protection legislation and other applicable laws .*

The first PIA (presented in D6.2) was undertaken first stage of the Seta project (M1-M6) that was dedicated to state of the art analysis and requirements gathering. As all new methodologies and technologies carry privacy and data protection concerns, all project partners were required to complete a PIA questionnaire (developed with the support of Trilateral Research) in Month 6. This fed into the design and development of all technologies and the Data Management Platform (this is the component that deals with data storage, data access and data protection), as shown in Table 1. At Month 6, not all the methodologies and technologies had been fully designed and it was acknowledged that further privacy and data protection implications may arise.

Risk title	Mitigation strategies translated into technologies
Identifying individuals	Limit the amount of personal data requested when using the mobile app Make registration optional Use of encryption when sending information from the mobile app to the server.
Combining data	Creation of separate databases on separate servers to store personal information separately from other data
Obtaining consent	Creation of a comprehensive set of terms and conditions of use for both the mobile app and the web apps, that were validated and approved by the Project Coordinator Legal Office (the University of Sheffield)
Tracking individuals	Share information only in aggregate form
Personalised data	Share information only in aggregate form
Data transfer	Use of OAuth protocol. Definition of data sharing protocols. Use of encryption when sending information from the mobile app to the server

Table 1- How privacy risks from D6.2 were translated into technology implementation will be presented in Section 3.2).

The second key date chosen for the revised PIA was Month 18 (this document) as it marked the end of the first technical evaluation of the technologies: as SETA is an iterative user-centred design process this is also the moment when new requirements for Phase 2 arise and need to be taken into account in the PIA.

At month 18 all partners were given a new PIA questionnaire (presented in Appendix A), following feedback and advice from Trilateral Research (our independent external ethics advisors) to fill and the responses were analysed and integrated to create this intermediate PIA (presented in Section 4).

All partners completed the questionnaire. Follow up 1 to 1 discussions with key partners were held when necessary or appropriate (separate consultations with The

Floow were held on 4th of February, 11th and 14th of July, as detailed in Section 3.1 and with K-Now and USFD on the 22nd of March).

Moreover, SETA is ensuring proactive privacy risks management by reviewing the key risks identified during consortium project meetings by the project steering group (e.g. the changes requested to The Floow for the data to be shared, detailed in Section 3.1).

The key risks are shared with all partners in a register (online spreadsheet) that is reviewed and updated at every Project Steering Committee Meeting.

The register is composed by the following fields:

- Work package where the risks is originated
- Unique ID
- Risk description
- Likelihood
- Impact
- Mitigation/action

The PIA process is led by the co-ordinator of the project, The University of Sheffield, with significant input from other consortium members and with expert advice being provided by Trilateral Research as our independent external ethics advisors.

This process will be repeated at the end of the project to ensure that a comprehensive assessment of privacy risks presented by the Seta technologies has been undertaken and any steps have been implemented to avoid and minimise any risks that were identified during the PIA process.

3.3. Partners

A new partner has been added to the consortium, 5T (Turin) since phase 1 of the PIA process. 5T is a company that manages traffic control for Turin and represents public Authorities and traffic control centre stakeholders in the project.

At the point of joining 5T was asked to sign the data sharing agreement (presented in D6.2) and to fill in a PIA questionnaire. The role of 5T in the consortium could potentially add privacy risks as 5T is a data provider for Turin's mobility.

The results of the questionnaire did not highlight any specific privacy or data risks, as all the data has been collected and publicly shared in anonymous form as Open Data, accordingly to National and EU regulations.

All other partners remained the same.

4. Risks identified and mitigation

The following risks have been identified during project steering meetings or through the PIA questionnaire process.

Compared to the initial PIA, the main change has been the raising of tracking and identifying individuals. This is due to changes in the technologies developed, in particular:

- The mobile app now asks the user for age, weight and gender, to calculate calories consumed
- The Floow was asked by technical partners to provide a more fine-grained dataset to improve modelling. This required further discussion and a detailed analysis of privacy implications (this has resulted in increasing the risk of Tracking Individuals to HIGH); the specific results are presented in Section 4.4.

An asterisk is placed near to the risk levels that have changed after the initial PIA due to changes in the technologies.

Whilst the collected data is then analysed in aggregated form we believe this is of high likelihood, impact and risk.

The project used an agile implementation of PRINCE2 methodology for risk assessment, based on qualitative assessment of the risks outlines in the questionnaire replies. This included the following steps:

- Risk identification: The starting point here was a risk prompt list based on risks identified in other similar project that was followed during a brainstorming session.
- Risk Assessment: Risks identified during the brainstorming were compared to those outlined in the PIA questionnaire and replies were aggregated into categories and then assessed for likelihood, impact and level of risk. The risks are then added to the risk register.
- Risk Plan: replies to the PIA questionnaires were used to identify mitigation strategies to help reduce or avoid the threats. When questionnaire replies were not sufficient of the situation changed rapidly (see USFD/K-NOW and The Floow in Section 4.4) separate consultations with partners were held.
- Risk Plan Implementation: the mitigation strategies identified in step 4 were integrated in the technology development.
- Risk Communication: the risks and mitigation strategies are constantly discussed with partners and project stakeholders.

The risks are outlined in the table below which allocates a risk level to each risk based on the likelihood (taking into consideration the proposed mitigation) and potential impact of each one. More detail is then given about each risk and proposed solutions. Details about each risk are provided in Section 4.

Risk title	Likelihood	Impact	Level of risk
Identifying individuals	Medium*	High*	High*
Combining data	Medium	Medium	Medium*
Obtaining consent	Low	Medium	Low
Tracking individuals	High*	High*	Very High*

Personalised data	Medium	Medium	Medium
(Current) Technologies will have unforeseen privacy implications	Low	High	Medium
(New) Technology development	Medium	High	High
Data transfer	Medium	Medium	Medium

Table 2 - Risks identification and assesment

Level of risks is based on a 3x3 risk matrix (see Table 3)with a five class ranking as detailed below: the level attributed to each risks was calculated by identifying how many times the risk was mentioned in relation to different technologies/datasets/partners.

Likelihood				
	High	Medium	High	Very high
	Med	Low	Medium	High
	Low	Insignificant	Low	Medium
		Low	Med	High
	Impact			

Table 3 - Risks matrix

4.1. Identifying individuals – High risk

4.1.1. The risk

That data collected within the project by the SETA mobile app could be used to identify individuals and risk invading an individual's privacy or put them at risk of being profiled.

The (optional) personal data collected by the project technologies is:

- Age
- Weight
- Gender
- Email address
- Big Birmingham Bike membership (Birmingham Only)
- Leisure Flex ID (Birmingham Only)
- City

The Device ID is also collected.

CCTV and camera video streams are also collected by Turin, Birmingham and Santander city council and shared with Scyfer for automatic information on traffic/bus occupancy but no personal data is extracted.

Data collected to track individuals' movements has also potential privacy implications that are strongly correlated with the possibility of identifying individuals, as described in Section 4.4.

As the projects aims to model travel behaviour, this may allow us to identify individuals and/or regular destinations which could include home, work, leisure destinations and patterns of behaviour.

4.1.2. The solution

To minimise the privacy risks of collecting personal data, the following steps have been taken: although personal data is collected, the data is:

- passed to the data processors in aggregated, anonymised form (anonymous identifier and time and location information at regular intervals) using secure communication and data transfer procedures.
- Used only to personalised the app features therefore are only stored on the user's device
- held separately to information recorded from applications and not shared with third parties (even within the consortium) by storing them on secure servers in USFD and K-Now, to which no other partner has access.

For what regards video and CCTV cameras streams, no personal data is extracted as the only analysis that is made is aggregated data on the number of vehicles/passengers per time interval. Individual passengers/cars are not considered and just used to generate a count by the algorithm.

4.2. Combining data – medium risk

4.2.1. The risk

When combining data from different sources it is theoretically possible to end up knowing more than was anticipated about an individual. This means that it can be possible to identify an individual even when this would not be possible from individual data sources.

4.2.2. The solution

The risk has been raised to Medium as Birmingham City Council has asked to collect optional details such as:

- Big Birmingham Bike membership (Birmingham Only)
- Leisure Flex ID (Birmingham Only)

Therefore Birmingham City Council will have the opportunity to BCC to cross reference their Leisure Flex ID and Big Birmingham Bike membership if the users provide information and consent.

We will continue to monitor this risk as part of project procedures by reviewing it as every Steering Committee meeting, recording any changes in the risk register.

4.3. Obtaining consent – low risk

4.3.1. The risk

Data controllers are obligated where possible to obtain consent from individuals when collecting their data. New SETA technologies and services must pay attention to this obligation.

4.3.2. The solution

For mobile applications and other activities where it is possible to do so, informed opt-in consent has been and will continue to be obtained. The informed consent has been checked and approved by the relevant Data Protection Officers. For example the consent form for the mobile app and all the relevant Term&Conditions were approved by USFD Data Protection Officer.

Where this is not possible, notably regarding traffic monitoring cameras/ CCTV, the understanding is that the municipal authorities have already explored privacy implications regarding relevant laws.[3]

4.4. Tracking individuals – High risk

4.4.1. The risk

As the project is about mobility it is possible to track an individual's movements and make assumptions about items such as home address or work address.

The tracking of individuals carried out by SETA involves capturing the following data:

- Location data
- Point Sensor data
- Origin Destination data regions

4.4.2. The solution

To minimise the privacy risks of tracking individuals, the following steps have been taken: although location data is collected relative to an individual, the data is:

- stored separately from personal data
- shared in anonymised form
- amalgamated and placed at a less specific geographic level/ time frequency level to avoid recognising individuals.

A particular concern for the project has been the collection of vehicular mobility carried out by The Floow ltd.

The Floow as a company focus upon gathering and processing high resolution mobility sensor data in order to understand and minimise driving risk. Gathered data is first obtained from various devices ranging between smartphones, on-board diagnostic port devices, white-box, Black box devices, blue-tooth phone pairing

[3] The operation of CCTV systems must be undertaken with due regard to the following legislation: • The Data Protection Act 1998 • The Human Rights Act 1998 • The Regulation of Investigatory Powers Act 2000 • The Freedom of Information Act 2000

devices and in vehicle embedded electronics. This gathered data regardless of the technical means of its capture is obtained with consent from opt-in drivers alongside contracted agreements in motor insurance, auto company services and breakdown cover. For each end user, The Floow collates second by second positional and sensor data so they can provide mobility analysis and related services for clients and end users primarily aiming to understand and importantly improve driving risk.

In order to provide Telematics analytic services, The Floow compares drivers against other drivers in similar circumstances and locations. To this extent The Floow has collected consent to reuse gathered data in aggregate and anonymous formats divorced from personal data for scoring analysis and other purposes which allow data reuse. Historical mass data helps The Floow build background comparison data, risk scoring data, mapping segment and route data, mobility analysis and other services to end drivers, insurers or other agencies. These commercial activities make use of mass mobility data covering a proportion of the UK drivers in fine grained detail over extended periods of time.

This very large data source has sensitive mobility information of many individuals' geo-positional movements thus has strict mandated controlled processing in order to ensure data protection and ethical processing of data. Ultimately in order to use data externally all data must be prepared into anonymous and aggregated forms to prevent misuse and to protect data originators.

Individual data points collected in raw format by the Floow are precisely timestamped so with various readings they could potentially be used to recreate the route of individual vehicles. To maintain anonymity for data originators it is therefore essential to ensure to collate samples so data does not reveal information about individuals route and locations; thus to protect privacy it is essential to aggregate data in two dimensions, namely space and time.

In February 2017 The Floow was asked by technical partner to release a finer-grained dataset for Birmingham region. Although this could have substantially improved the modelling outcomes it also raised more concerns about privacy.

The table below (Table 4) shows the possible levels of aggregation and the corresponded statistical relevance for modelling.

Aggregation plan number	Aggregation type	Geographical Resolution	Temporal Resolution
1	Background Segment Analysis 1	Segment level resolution	24/7 all days and hours combined
2	Background Segment Analysis 2	Segment level resolution	Morning peak/Mid Peak/Evening Peal/Off Peak
3	Background Segment Analysis 3	Segment level resolution	Morning peak/Mid Peak/Evening Peal/Off Peak – by Day of the Week
4	Background Segment Analysis 4	Segment level resolution	Data buckets to hour (or multi hour periods) – Data buckets split by volume of data in each – by Day of

			the Week
5	Background Segment Analysis 5	Segment level resolution	Data buckets to hour (or multi hour periods) – Data buckets split by volume of data in each – by Day of the Week
6	Background Segment Analysis 5	Segment level resolution (but only where ten data samples are recording in the timeframe)	15min aggregation on a day of the week basis

Table 4 - The Floow data aggregation proposals considered in the discussion

After discussions and consultation with technical partners it has been agreed that whilst aggregation in row1-5 does not pose privacy risks, the aggregation in row 6 would have too high privacy risks, as on small roads it may highlight individual journeys around journey endpoints, therefore The Floow could not proceed with Aggregation 6.

For what regards video analysis on roads this has been engineered to avoid licence plate recognition - instead relying on the shape, model and colour of a vehicle to track its movements around the immediate monitored metropolitan area only.

4.5. Personalised data – medium risk

4.5.1. The risk

The project aims to provide personal decision-making capacity for individual citizens and this will involve being able to understand individual attributes, requirements and habits. This could potentially lead to a breach in an individual's privacy, should the data be used for behavioural profiling (this is not the case in SETA).

4.5.2. The solution

All technologies that provide personalised interaction do not share personal information and any personal data is stored separately from mobility data. In addition, the technologies that collect personal data (i.e. the Mobile app) obtains informed opt-in consent from individuals through a form that links to the full terms and conditions of usage.

4.6. (Current) Technologies will have unseen privacy consequences – medium risk

4.6.1. The risk

That the current technologies have privacy flaws that we are not aware of which could lead to privacy breaches. For example in section 4.4 we have detailed the case of The Floow, where all data is shared only in anonymised and aggregated form but a new request from project partners led to the discovery that aggregating data in 15 minutes intervals could cause privacy breaches presenting an privacy concern if provided.

4.6.2. The solution

The consortium partners are experienced in this area and most of the currently scoped technologies are based on or are extensions of existing systems or research. For that reason there has already been examination and resolution of the privacy implications of these technologies, as highlighted for example in section 4.4 for vehicular data.

4.7. (New) Technology and service development – high risk

4.7.1. The risk

Because the technology arena is evolving quickly and the project uses an iterative and agile design approach the technologies design and implementation may change during the project. There may be developments or new technologies which will have different privacy implications.

4.7.2. The solution

Privacy issues may arise at any stage of the project and it is for this reason that the PIA is an iterative process. Consortium partners will be asked to focus on newly developed technologies and services in reviews of the PIA and to notify the project coordinator when any change to the technology is made and what are the relevant privacy implications, risks and mitigation strategies. The risks are recorder in the risk registry (presented in Section 3) and constantly revised. In addition, the consortium co-ordinator and project manager will monitor new developments to ensure the high visibility of privacy issues.

4.8. Data transfer and sharing – medium risk

4.8.1. The risk

Because of the varying partners and roles in the project there is a risk that the data may be intercepted or misplaced during data transfer. The data transfer carried oud in the project involves:

- Transfer of anonymised and aggregated location data for non-motorised mobility
- Transfer of anonymised and aggregated location data for vehicular mobility as detailed in Section 4.4)

All other personal data is not shared.

4.8.2. The solution

To avoid any issues with data transfer we will invoke strict data transfer protocols and agreements between the parties involved. The transferred data has already been anonymised and amalgamated but will still be treated especially cautiously throughout the data sharing protocol design and implementation.

To ensure a unified and manageable data security, privacy and authorisation capability across all the components in the Platform SETA uses OAuth protocol. Compartmentalization process have defined together with data access roles. The Data sharing protocols across the project are constantly monitored by the Project Manager to ensure that the agreements are in place to the satisfaction of all parties involved and that actual data transfer is carried out according to the agreed

protocols. Connections are protected through encryption and data is stored on certified data centers with suitable levels of protection built in.

This caution, and the pursuant agreements and protocols, also extends to the secure storage and disposal of any project data.

4.9. Further risks

SETA is aware of potential for changes to national or European law regarding privacy which may prevent us using our anticipated technologies and methods for collecting data. For example, the GDPR has a broader definition of personal data than provided for under the Data Protection Act (DPA). This means that information such as an online identifier can now be classified as personal data. Changes to personal data definitions such as this could lead to privacy issues over the course of the project lifetime – hence part of the requirement for ongoing monitoring.

4.9.1. The solution

- Each partner has been asked at the beginning of the project and every 3 months (before each project meeting) to contact their National Privacy and Data Protection Authority and/or their Legal office to verify the project compliance with national rules
- USFD is constantly reviewing EU legislation (in cooperation with the University Legal Office) to ensure that all data processing and storage will be carried out in accordance with the General Data Protection Regulation (GDPR) as this is the most up to date and relevant piece of legislation in the area of data protection in the EU. The project is also monitoring other legislation where appropriate[4]
- Any relevant changes in GDPR or national legislation are discussed in the Project Steering Committee Meetings and is necessary, the requirements for the project's data anonymisation, data protection and privacy protection are updated.

5. Recommendations

The PIA recommends the following points to the attention of the project steering group.

5.1. Technical recommendations

- To continue to focus on identifying any risks associated with aggregating data as the technologies and services are further developed, and to avoid, mitigate or minimise these appropriately.

[4] *Including the Convention for the Protection of Human Rights and Fundamental Freedoms (in particular Article 8); the Charter of Fundamental Rights of the EU; the EU Directives 95/46/EC and 2006/24/EC; Article 29 Working Group's Opinion 8/2010; as well as applicable national data protection legislation and other applicable laws .*

5.2. Procedural recommendations

- The project co-ordinator should monitor the development and content of any additional data sharing protocols between partners.

- To continue monitoring the Project Risk Register and particularly the section on privacy risks. Every partner is requested to notify the project co-ordinator immediately of any change in their technologies/implementation or national country legislation. The risks identified in the PIA process will be added to the Project Risk Register.

- To continue reviewing external privacy and other relevant regulations in the national and European contexts but in particular monitoring of any developments with interpretations of the GDPR.

- To ensure the ongoing assessment and analysis of risks, especially for tracking individuals, related to the project by carrying out further PIAs. As risks related to privacy may emerge at any stage of the project the steering group should also regularly review privacy implications outside of the formal review process and in particular when new technologies and/ or services are being discussed and designed.

Appendix A -

Privacy Impact Assessment Questionnaire

Version 2 – July 2017

Note: This questionnaire is sent to each SETA partner for you to detail, to the best of your knowledge, a) the technologies you will make available or will be designing, developing and implementing as part of the SETA project and b) the data which will be used. The purpose of a Privacy Impact Assessment Questionnaire is to ensure privacy risks are minimised while allowing the aims of the project to be met whenever possible. The results of this questionnaire will be used to inform the Privacy Impact Assessment section of D6.5 and to guide the technology design and development.

The questionnaire has a specific focus on personal data, i.e. any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person [Article 4(1) of the GDPR].

Answer the questions below so that there is a clear understanding about how the information will be used, who will use it etc. Remember to think about all the data that will be collected. If you do not have answers to all the questions at this time, simply record what you do know.

Organisation name:	
Name of lead:	
Date of completion:	

1 TECHNOLOGY

Please describe in simple terms the system(s)/ technology(ies)/ tool(s) that you are planning to develop for the SETA Project?

[Repeat this section for each system/ technology/ tool.]

1.1 Name or brief description of system/ technology/ tool.

1.2 What is the purpose of the technology?

1.3 What is its role in SETA?

1.4 Does the technology involve any surveillance of individuals (i.e. monitoring of behaviour and activity of an individual)? If yes, provide brief details.

1.5 Does the technology involve any collection of biometric information (this may include things such as body scanning, iris recognition, gait recognition, etc.)? If yes, provide brief details.

1.6 Does the technology use CCTV or other video surveillance technologies? If yes, provide brief details.

1.5 Does the technology use any form of user tracking (e.g. geo-location, etc.)? If yes, provide brief details.

1.6 Does the technology use profiling technologies (i.e. the use of algorithms or other mathematical techniques that allow the discovery of patterns or correlations in large quantities of data, that in return are used to identify or represent people)? If yes, provide brief details.

1.7 Could the technology affect any vulnerable groups (directly or indirectly, i.e. technology that could provide abusers with the tools and information necessary to carry out abusive or controlling behaviour)? If yes, provide brief details.

1.8 Could the technology benefit some groups and exclude others? If yes, provide brief details.

1.9 Can you think of any other privacy issues that have not been considered above?

2 DATA - OVERVIEW

2.1 In the table below please detail what type of personal data is going to be collected and/or processed? Mark for special attention any that is regarded as 'sensitive'. Add new rows as required.
Note – Sensitive personal data is personal data consisting of information "revealing race or ethnic origin, political opinions, religion or beliefs, trade-union membership, and the processing of genetic data or data concerning health or sex life or criminal convictions or related security measures shall be prohibited." [Article 9 of General Data Protection Regulation].

No.	Type of data (e.g. name, age, email address)?	Is the data 'sensitive'? Yes/No	is any data collected utilised for any other purpose than that which originally stated?	Source? (provided, observed, derived, inferred)	Specific purpose?	What processing will be performed?	From how many individuals will data be collected?
1							
2							
3							
4							

2.2 For each type of data outlined above please complete the following table concerning individual consent. Add new rows as required.

No.	Will specific consent be obtained? Yes/ No	Give details of how/ why not	What type of notice(s) are provided to the individual on the scope of information collected, the opportunity to consent to uses of said information, the opportunity to decline to provide information, who is collecting such information	Do individuals have an opportunity and/or right to decline to provide all or part of their information*? If so, what is the procedure by which an individual would achieve this? Note: We should aim for systems that are Opt-In not Opt-Out to follow GDPR guidelines.
1				
2				
3				
4				

3 DATA COLLECTION AND PROCESSING

3.1 How many individuals do you intend collect data from (if any)?

3.2 What procedures will be in place for checking that the data collection procedures are adequate, relevant and not excessive in relation to the purpose for which the data will be processed?
3.3 Do you have informed consent procedures in place? Please outline these.

3.4 Will the personal data be checked for accuracy? Please outline this process.

3.5 Has the personal data been evaluated to determine whether its processing could cause damage or distress to data subjects (by whom and what methodology was employed)?

3.6 Will there be any handling of types of personal data that might be of particular concern to individuals? This could include information considered as 'sensitive' (see first page).

3.7 Will personal details about individuals in an existing database be used for a different purpose? If so, does the original informed consent procedure provide this information to individuals?

3.8 Will there be consolidation, inter-linking, cross-referencing or matching of personal data from multiple sources. If so, briefly explain.
3.9 Please outline whether and how users have been informed/ provided consent for this activity.

3.10 Will there be collection policies or practices that may be unclear or intrusive?
Note: Anything that is designed that may be in this category should be discussed by the consortium as a whole to obtain partner consent. Such practices may present legal jeopardy for all organisations involved.

3.11 What data quality assurance or processes are in place? Please outline these.
Note: Anything that is designed that may be in this category should be discussed by the consortium as a whole to obtain partner consent.

3.12 'What data security policies and procedures are in place? Please outline these.
Note: Anything that is designed that may be in this category should be discussed by the consortium as a whole to obtain partner consent.

3.13 What data access arrangements are in place? Please outline these.
Note: Anything that is designed that may be in this category should be discussed by the consortium as a whole to obtain partner consent. All data access should default to a 'need to know' basis.

3.14 What, if any, data retention policies and procedures are in place? Please outline these.
Note: Extensive may be defined as anything more than a year following the end of the project. Anything that is designed that may be in this category should be discussed by the consortium as a whole to obtain partner consent.

3.15 Will any additional data be made available in the public domain?

Note: Anything that is designed that may be in this category should be discussed by the consortium as a whole to obtain partner consent.

4 Transfer to third parties

4.1 Will the data be transmitted/ released to third parties? Please note that data can only be transferred if authorised by the conditions of use (e.g. via user consent) and to a partner providing security equivalent to the one provided by your organisation.

If yes, please answer the following questions. Describe any sharing arrangements and what the level of access is. It may help to produce a diagram to show the data flows.

Will data be shared with any third parties?
If yes, please provide details.

2) What data will be shared?

3) How will data be transmitted?

4) Are there policies and procedures in place for information sharing? Please outline these.

5) Have individuals been informed of the potential for their data to be shared with third parties?

6) Have individuals provided informed consent for the sharing of their data with third parties?

5 DATA STORAGE

5.1 How long is data retained for?
Note: This refers to all parties who have access to the information.
Will there be set retention periods in place in relation to the storage of the personal data? If there are different retention periods for the different types of data please detail this in a table. If any retention will take place for more than a year beyond the end of the project please detail this and provide reasons for it.

5.2 How is data kept accurate and up to date?

5.3 Please outline the process of deletion once data is no longer required.

5.4 Are individuals provided the right to issue a subject access request?
If yes, is the extraction of this data possible without infringing on third party data?
Note: Due to the research nature of the project we can avoid restrictions on secondary processing and on processing sensitive categories of data (Article 6(4); Recital 50). As long as appropriate safeguards are in place we may also have an argument to override a data subject's right to object to processing and to seek the erasure of personal data (Article 89).

5.5 Is it possible to rectify, erase or block access to individual records if required? Please note that in some cases users may require their data to be removed within 72 hours.

5.6 What technical and organisational security measures will be in place to prevent any unauthorised or unlawful access to or processing of the personal data?
5.7 Who in your organisation/ in the project is responsible for judging the sufficiency of these measures?

5.8 Describe the procedure for ensuring the destruction of personal data once its retention is no longer necessary?

5.9 Will you be transferring personal data to any country outside of the European Economic Area? If so where, and what arrangements will be in place to ensure that there are adequate safeguards over the data?
Note: Data must remain within the EEA region without a data agreement agreed by the end user and of sufficient protection as to be equivalent to GDPR law (e.g. US safe harbour etc.).
All data in the SETA project is currently anticipated to remain in the EEA region.

6 Stakeholder analysis

Note: Please detail any stakeholder groups who have been specifically consulted about privacy for the SETA project (and data it handles). Also list any further stakeholder groups who will be consulted about privacy on the project (or data it handles) in the future.

7 Environmental scan

Note: Please detail previous projects either within or outside your organisation that have helped to inform design features or processes that will be used as 'lessons learned' in this project.

8 Compliance with Privacy Laws

Note: Please detail how data collection, storage, retention and disposal fits with your local (national) and European privacy legislation and regulations.