H2020-EUJ-02-2016
H2020 Grant Agreement Number 723076
NICT Management Number 18302

Deliverable D7.2

Data Management Plan (v1)

Version V1.0

December 28th, 2016

ABSTRACT:

The CPaaS.io project is articulated around on the one-hand a comprehensive set of real-world scenario where large amount of raw data is produced, and on the other hand, on reasoning and analytics techniques that allows cities to exploit that data – often referred to as the oil of the 21st Century"- through the generation of higher order meaningful knowledge.

This document gives details about all data either collected or generated by the CPaaS.io project and the strategies put in place in order to manage, sustain and share this large amount of information. In particular besides reminding the nature and purpose of the 5 CPaaS.io scenarios data-wise, each CPaaS.io partner provides statement about roles undertaken as far as managing the collected and generated data is concerned. This preliminary version of this living document (updated and maintained continuously) will be complemented with an official final version by the end of the project.

Disclaimer

This document has been produced in the context of the CPaaS.io project which is jointly funded by the European Commission (grant agreement n° 723076) and NICT from Japan (management number 18302). All information provided in this document is provided "as is" and no guarantee or warranty is given that the information is fit for any particular purpose. The user thereof uses the information at its sole risk and liability. For the avoidance of all doubts, the European Commission and NICT have no liability in respect of this document, which is merely representing the view of the project consortium. This document is subject to change without notice.

Document Information

Editors	François Carrez (UNIS)
Authors	EU consortium
Reviewers	Stephan Haller (BFH)
Delivery Type	R
Dissemination Level	Public
Contractual Delivery Date	30th December 2016
Actual Delivery Date	28th December 2016
Keywords	Data management, Data openness, Privacy

Revision History

Rev.	Date	Description	Contributors
v0.1	23/11/2016	Extended ToC	François Carrez
v0.2	08/12/2016	Sections 2 and 3 update, AGT	Martin Strohbach
v0.3	08/12/2016	Sections 2 and 3 update, UoS	François Carrez
v0.4	11/12/2016	Section 3 update, BFH	Stephan Haller
v0.5	13/12/2016	Section 2 and 3 update, TTN	Alexander Overtoom
v0.6	13/12/2016	Section 3 update, OdinS	Antonio Skarmeta
v0.7	14/12/2016	Section 3 update, NEC	Gurkan Solmaz
v0.8	23/12/2016	Final version for review	François Carrez
v0.9		Review comments	Stephan Haller
v1.0		Final version for delivery	François Carrez

Table of Contents

List of Tables

List of Acronyms

Acronym	Definition
DMP	Data Management Plan
HDF	Hierarchical Data Format
XML	eXtensible Markup Language

1 Introduction

As described with the H2020 guidelines, Research funding organisations, as well as organisations undertaking publicly funded research, have an obligation to optimize the use the funds they have been granted. Part of this optimization is that data sets resulting from public funded research must be made available to other researchers to either verify the original results (which is integral part of proper scientific approach), or to build upon them.

In order to achieve this high-level objective, a data management policy has to be implemented and thoroughly followed by the CPaaS.io consortium as a whole, even if –per se- not all CPaaS.io partners will be involved in all aspects of those policies/principles.

The Data Management Plan (DMP) is a living document (with two formal versions of the same deliverable released in M6 and M30 respectively) that describes the data management policy (i.e. the management principles) and collected and generated data sets. It covers all aspects introduced in the "Guidelines on Data Management in Horizon 2020", which are:

1. Precise description of the collected and generated data (nature of data, related domain ontologies, standards and data formats used,…)
2. Detail about various aspects of the data management (how it is stored, by whom, under which responsibility, how it is secured, how it is sustained and backed up)
3. Sharing principles (licensing, access methods,...)
4. Detail about how the privacy is maintained

This first version of the Data Management Plan gives a preliminary description of the data as collected and generated by both the CPaaS.io platform and project partners through their legacy systems. At the time of the document editing, some aspects of data management are still under discussion, mainly because they are strongly depending on some technical decisions pertaining to the CPaaS.io system platform design and how this architecture deals with partners' legacy systems as far as storage, backup and data flows are concerned.

Aspects such as data backup, sustainability, detail about data sharing and archiving will be thoroughly developed through an intermediary and far more complete version of this deliverable.

In this current version we mainly provide detail (as known at M6) about the scenarios and collected data (see Section 2) and roles of partners as far as Data Management in CPaaS.io is concerned (see Section 3).

Due to differences between EU and Japanese formal contracts and differences in data-related rules and constraints, we have focussed in this initial version on scenarios and partners from EU only. However the living document will aim at harmonizing the different views through a single and consultable internal document.

2 CPaaS.io Research Data

This section introduces the different EU-side use-cases as described in the CPaaS.io Description of Work document and the applications built upon them. It also describes the collected data (meaning the semantically annotated raw-data with no extra added value) and the generated data (meaning the semantic value-added information built from the annotated raw-data using various technics like analytics or reasoning). Part of the information described in this section can be found in a more complete form in CPaaS.io deliverable D2.1 [1].

The two scenarios considered in CPaaS.io for the EU-only side are:

- Managing Fun and Sport events
- Waterproof Amsterdam

And the two derived applications are:

- Enhanced User Experience
- Waterproof Amsterdam

2.1 Data from Enhanced User Experience application

Short description

The core idea of this application is to use IoT sensors and analytics to enhance people's experience while visiting or participating at a fun or sports event. Wearables and mobile phones are used as sensors in order to learn about the activities of event participants. Event participants may include members of the audience, but also performing artists or athletes. For instance AGT has previously equipped referees and cheerleaders in basketball matches with wearable sensors and created content based on the analysed data for consumption on site and for distribution via TV broadcasting, social media and other digital distribution channels[1]. Furthermore the application uses sensor deployed at the venue to measure and analyse fan behaviour and engagement.

Data collected for the Enhanced User Experience application (Color Run)

Table 1 summarizes the data from the Enhanced User Experience application as described in D2.1. Please note that although the hosting field specifies that the most of the data is hosted external to the CPaaS.io platform we are considering to use the storage capabilities in the next iterations. Further to the data sets described in D2.1 we have added an additional mobile camera data set.

[1] http://www.euroleague.net/final-four/berlin-2016/news/i/6vokoibj5fsgqg4q/heed-the-event-platform-based-joint-venture-between-wme-img-and-agt-international-makes-first-official-foray-into-sports

Table 1: Data collected for Managing Fun and Sport events scenario

Biometric data	
Detailed Description	We will collect a range of biometric measurements from wearables such as wristbands, chest straps and smart sportswear that provides biometric measurements including heart rate, breathing rate and galvanic skin response, burned calories measurements and skin temperature.
OGD or private data	Private
Personal Data	Yes
Hosting	External
Data Provider	AGT
Format	JSON
Update Frequency	up to every 200ms
Update Size	~1 KB
Data Source	Sensor
Sensor	Wristband, chest strap, smart shirts
Number of Sensors per person	~6

GPS Traces	
Detailed Description	GPS traces include positional data including altitude information as delivered by GPS devices.
OGD or private data	Private
Personal Data	Yes
Hosting	External
Data Provider	AGT
Format	common GPS formats (GPX, KML, CSV, NMEA)
Update Frequency	Up to 1s
Update Size	< 1KB
Data Source	Sensor
Sensor	GPS sensor in wristbands and mobile phones
Number of Sensors per person	1-2

Motion Data	
Detailed Description	Motion data that measures hand and body movements based on accelerometer and gyroscope sensors
OGD or private data	Private
Personal Data	Yes
Hosting	External
Data Provider	AGT
Format	JSON
Update Frequency	Up to every 16 ms
Update Size	~ 200 byte per sensor reading
Data Source	Sensors

Sensor	Accelerometer and gyroscope sensors of mobile phones, wristband and other wearables
Number of Sensors per person	2-3

Step Counts	
Detailed Description	This data set contains step counts.
OGD or private data	Private
Personal Data	Yes
Hosting	External
Data Provider	AGT
Format	JSON
Update Frequency	Up to 1Hz
Update Size	~ 200 byte per sensor reading
Data Source	Sensors
Sensor	Step count measurement of wristband
Number of Sensors per person	1-2

Environmental Data	
Detailed Description	This data set environmental data such light intensity and barometric pressure. The data is primarily collected from wearable sensors.
OGD or private data	Private
Personal Data	Yes (tbc)
Hosting	External
Data Provider	AGT
Format	JSON
Update Frequency	Up to 1Hz
Update Size	~ 200 byte per sensor reading
Data Source	Sensors
Sensor	Sensors in wristband
Number of Sensors per person	1-2

Mobile Camera videos	
Detailed Description	This data set contains videos recorded by mobile cameras worn by Color Run participants.
OGD or private data	Private
Personal Data	Yes
Hosting	External
Data Provider	AGT
Format	MP4
Update Frequency	30fps
Update Size	(~45kbps)
Data Source	Mobile Camera

Sensor	GoPro Hero4 Camera
Number of Sensors per person	1

Data generated by the Enhanced User Experience application (Color Run)
The Enhanced User Experience application generates three types of data

1) User Activity
2) Dominant Colors
3) Clothing Analysis

User activity is based mainly on motion data and therefore private information. A user activity is always linked to a user and therefore personal information. The re-use of the data is possible within the boundaries defined in the consent forms used to collect the data.

Dominant Colour provides information about the prevailing colour in a video feed and is used for detecting colour stations in the Color Run. The output is a colour value, duration and location. The generated can be provided in anonymised form, but requires further examination to what degree it can be opened.

Clothing Analysis uses deep learning techniques to determine metrics based on clothing styles derived from images. By nature this metrics are linked to user and therefore reflect private data that can only be reused in the boundaries of the consent forms used to collect the data.

Table 2: Data generated for the Enhanced User Experience application

Types of generated data	Based on...	Anonymised Y/N	Open Y/N
User Activity	Motion Data	N	reusable, but not open
Dominant Colour	Mobile Camera Videos	Y	Reusable, but not fully open
Clothing Analysis	Mobile Camera Videos, Public Images	N	Reusable, but not open

2.2 Data from Waterproof Amsterdam

Short description

Extreme rainfall and periods of continued drought are occurring more and more often in urban areas. Because of the rainfall, peak pressure on a municipality's sewerage infrastructure needs to be load balanced to prevent flooding of streets and basements. With drought, smart water management is required to allow for optimal availability of water, both underground as well as above ground.

The Things Network develops the Amsterdam Waterproof application, which is a software tool creating a network of smart, connected rain buffers, be it rain barrels, retention rooftops or buffer otherwise, that can be both monitored and controlled centrally by the water management authority. Third party hardware providers will connect their buffers to this tool for uplink and downlink data transmission.

External data such as weather data and sewerage capacity are added, in order to calculate the optimal filling degree of each buffer and so operate a pump or valve in the device. Waternet is the local water management company who will be the main user of the application.

Data collected for the Waterproof Amsterdam application

In the section below are the data sets used for the Waternet application. It consists of device data (rain buffer information), public weather data and government data about physical infrastructure. Device data will be stored in the application and could be stored in CPaaS, especially as it contains private data like name and address of device owner. As this stage however we cannot determine whether this privacy data will be shared by the vendors of the devices, who are also the ones maintaining them. They are the only actor who has direct contact with the end user and/or owner of the device. (Historical) weather data is publicly available on the web, so there is no need to store this data. It will be provided by a subscription data feed from the web. The third data set is already owned and stored by Waternet, so there is also no need for storage capabilities.

Table 3: Data collected for the Waterproof Amsterdam scenario

Weather data	
Detailed Description	Upcoming weather displaying periods of heavy rain or drought
OGD or private data	OGD
Personal Data	No
Hosting	Platform
Data Provider	KNMI – Dutch weather forecast agency
Format	HDF5/JSON
Update Frequency	Hourly
Update Size	20kb
Data Source	Sensors
Sensor	Water sensor
Number of Sensors	unknown

Rain buffer information	
Detailed Description	Specific information about each rainbuffer (rooftop, barrel, underground storage) • Buffer size and type • Filling degree • Temperature • Location • Battery status • Pump/valve capacity • Active pump/valve hours • Owner name, address, contact information
OGD or private data	Private
Personal Data	Yes – anonymised and not open
Hosting	Platform
Data Provider	Rain buffer hardware provider
Format	JSON
Update Frequency	Hourly
Update Size	10b

Data Source	Sensors
Sensor	Water sensor or infrared sensor
Number of Sensors	1 per buffer

Sewerage processing capacity	
Detailed Description	Geographical data on water infrastructure depicting remaining capacity of sewerage
OGD or private data	Private
Personal Data	No
Hosting	External
Data Provider	Waternet
Format	XML
Update Frequency	Hourly
Update Size	1kb
Data Source	Sensors, maps
Sensor	Water sensor
Number of Sensors	unknown

Data generated by the Waterproof Amsterdam application

The Waterproof Amsterdam generates different types of data.

1. Open/close command per buffer. This is the most important data generated, as it determines when an actuator inside a buffer should be operated (valve open or pump on). Based on all data sources available, an algorithm will determine which conditions are required to perform a certain command. The commands can be open and close, or a value in between as different water discharge mechanisms have different capacities (i.e. a percentage of full capacity)
2. Aggregated remaining buffer capacity per area. Waternet as the primary user of the application needs to monitor the total remaining capacity to buffer rain water, to understand whether there will be plenty capacity to catch up rain water in moments of heavy rainfall.
3. Aggregated litres of rain water processed per area. This is a metric to be used to show the impact the micro buffer network has generated over time. These insights may be used for PR and marketing purposes to stimulate individuals and companies to also buy and install such rain buffers.

The open data in the table below can be reused to perform analytics on historical data, and could be open data through a public (graphical or application) interface for third parties to interact with.

Table 4: Data generated for the Waterproof Amsterdam application

Types of generated data	Based on…	Anonymised Y/N	Open Y/N
Open/close command per buffer	All data sets	Y	N
Aggregated remaining buffer capacity (street, area, city level)	Individual rain buffers filling degree and location, map	Y	Y
Aggregated litres processed by the buffers	Individual rain buffer pump hours run and pump capacity, map	Y	Y

3 CPaaS.io Research Data management plan

CPaaS.io project follows the principle that research data will be handled and managed by those organisations/institutions that either collects or generates the research data. The CPaaS.io project comprise a number of partners that are involved directly in either:

- Producing the actual data during the trials, or
- Developing tools and enablers (e.g. analytics, reasoners, etc.) that are needed as core elements in the CPaaS.io system architecture, or
- Elaborating upon the produced data (using the aforementioned enablers) in order to produce new value-added knowledge.

The individual roles and duties of such partners and the research data management plans that are in place in the organisations taking part in CPaaS.io are described in the following sub-sections.

3.1 AGT International (AGT)

Data collection (from sensors)
The data collected by AGT has been described in Section 2.1 and is used for generated the data as described in Table 2 and for developing the Enhanced User Experience application. As described in D2.2 the collected data is enriched with additional metadata.

Data generation
The data generated by AGT has been described in Table 2 and is used in the Enhanced User Experience application.

Data Management

We have implemented appropriate technical and organizational measures to ensure generated data is protected from unauthorized or unlawful processing, accidental loss, destruction or damage. We review our information collection, storage and processing practices regularly, including physical security measures, to guard against unauthorized access to our systems. We restrict access to generated data to only those employees, contractors and agents who strictly need access to this information, and who are subject to strict contractual confidentiality obligations.

3.2 University of Surrey (UoS)

ICS at the University of Surrey is not involved neither in the production of raw data nor in the exploitation or generation of higher-level information out of it. However, UoS is focussing on architecture work where particular attention is paid to ensuring that 1/ all privacy-related requirements are thoroughly taken into account 2/ important part of the data is publicly available following the project Open Data policy.

To this respect UoS is aiming at providing a bridge between CPaaS.io and another FIRE project called FIESTA-IoT, two projects where UoS is actively involved. UoS will in particular aim at involving CPaaS.io in either the 2nd Open Call of FIESTA-IoT or as a fellow contributor to that project via a cooperation agreement to be discussed between the two projects after both POs have been consulted on that matter. In both cases, CPaaS.io could play two non-exclusive distinct roles:

- Data-provider: playing this role the CPaaS.io project would inject its data or part of its data (either raw data or inferred data) to the FIESTA-IoT so that so-called experimenters can make use of it using the FIESTA-IoT enablers; or
- Experimenter: playing this role, CPaaS.io could reuse additional data sets produced by the FIESTA-IoT collaborators for testing our new own algorithms (e.g. Analytics) and techniques.

Data collection (from sensors)
UoS does not participate in any data collection

Data generation
UoS does not generate any new data from the project data sets

Data Management
UoS does not manage any gathered or generated data

3.3 Bern University of Applied Sciences (BFH)

The BFH is not directly involved in the implementation of the envisaged use cases. Its main research focus is in the data management concepts – in particular the usage of Linked Data and Open Government Data as well as data quality annotations, the application of MyData approaches, and in the validation of the use cases. Hence it is not collecting, generating or storing any data.

However, as part of its exploitation, validation and knowledge transfer activities, BFH is planning to connect some sensors via the LoRa testbed network that another institute (Institute for Energy and Mobility Research in Biel) is currently setting up. What data will be collected and for what purposes exactly will be defined at a later stage; a related data management plan will be drawn up before any data collection starts.

Data collection
BFH is not collecting any data for the main use cases of CPaaS.io. It may collect and make available some sensor data through the LoRa network at BFH for testing and validation purposes; details will be determined at a later stage.

Data generation
BFH is not generating any data for the main use cases of CPaaS.io. It may link public data sources (e.g., from the Swiss Open Government Data portal at www.opendata.swiss) with the sensor data collected through the LoRa network at BFH for testing and validation purposes; potential use cases will be determined at a later stage.

Data Management
BFH is not managing any data for the main use cases of CPaaS.io. Data collected and generated for testing and validation purposes through the LoRa network at BFH will likely be made available publicly, in the spirit of open data research, unless the data could allow to infer any information about individuals. Details are to be determined at a later stage.

3.4 OdinS

OdinS as a partner involved on the security and privacy aspects, will check and support the project to check that data access and sharing activities will be implemented in compliance with the privacy and data collection rules and regulations, as they are applied nationally and in the EU, as well as with the H2020 rules. Concerning the results of the project, these will become publicly available based on the IPRs as described in the Consortium Agreement.

Due to the nature of the data involved, some of the results that will be generated by each project phase will be restricted to authorized users, while other results will be publicly available. Data access and sharing activities will be rigorously implemented in compliance with the privacy and data collection rules and regulations, as they are applied nationally and in the EU, as well as with the H2020 rules.

Data collection (from sensors)
OdinS will not be involved in the data generation of data from sensors, working exclusively in the architecture aspects of the data collections and its consequence over the security and privacy components.

Data generation
OdinS is not involved in the production of raw data, but as part of the Task 4.1 User Empowerment Component Definition and the definition of access control policies and use consent solution, OdinS will generate information associated to data for controlling access and sharing data between entities and components that will use the platform.

Data Management
As the raw data included in the data sources, will be gathered from sensor nodes and information management systems, those could be seen as highly sensitive. Therefore, access to raw data can only take place between the specific end users based on the policies associated and the partners involved in the analysis of the data. For the models to function correctly, the data will have to be included into the CPaaS.io repository. The results of the data analytics are set to be anonymised and made available to the subsequent layers of the framework, which will then allow the possibility for external industry stakeholders to use the results of the project for their own purposes.

3.5 NEC

NEC is not directly involved in the production or raw data. NEC’s focuses are in the architecture (system integration including transferability and semantic interoperability) area and cloud-edge processing of the data. FIWARE’s resources such as the Generic Enablers and NEC’s IoT Platform can support storage and exploitation of data from use cases for generating higher-level analytical results. NEC pays particular attention to privacy related requirements as well as the Open Data policy of CPaaS.io.

Data collection
NEC is not planning to collect any raw data for the use cases of CPaaS.io.

Data generation
NEC is not generating data for the main use cases, NEC may exploit shared data from use cases and generate higher level data as a result. Potential use cases will be determined at a later stage.

Data management

While NEC is not directly involved with the use cases, it will take part in data transferability and management via the provided IoT Platform. NEC has implemented necessary organizational and technical measures for the usage of the data and its protection from unauthorized persons.

3.6 The Things Network

Data collection (from sensors)

The data collected by The Things Network has been described in Section 2.2 and is used for generated the data as described in table 2 and for developing the Waterproof Amsterdam application. As described in D2.4 the collected data is enriched with additional metadata.

Data generation

The data generated by The Things Network has been described in and is used in the Waterproof Amsterdam application. Private data from owners of a rain buffer is anonymised. Based on an algorithm, data from various sources is processed by the application to determine the optimal filling degree for each individual rain buffer. The results may be used for automated control of buffers, or push notifications to trigger manual control.

Data Management

Open data such as weather data will be streamed into the application and not stored locally.

Private data from external sources such as device location will be stored in the application and only released in an anonymised and aggregated manner. Personal details about a device, such as name, address and contact details will also be stored in the application in a secure account server. These data may be transferred to CPaaS.io at some time, easing security and privacy demands on the application end and transferring those to CPaaS.io

Parts of the personal data, such as buffer location, size and processed litres, will be released in an aggregated, anonymised manner (e.g. on a heat map) per area of a city or the city as a whole.

Readily available data from Waternet about sewerage capacity will abide by the policies of Waternet. These policies are not yet clear at the moment.

We restrict access to generated data to only those employees, contractors and agents who strictly need access to this information, and who are subject to strict contractual confidentiality obligations.

4 Conclusions & Next Steps

In this deliverable we presented the CPaaS.io approach towards data management as handled by the EU CPaaS.io consortium. However at this early stage (M6), we do not have yet very precise information about the data collected or generated by the project. Some functional aspects are also still under discussion which prevents giving much detail about type and location of data storage, backup procedures, techniques used for generating data, and architecture-related detail in general.

However, being a living document, future iterations of this deliverable (even if not official deliverables) will provide increasing level of detail about all data sets collected and generated by the project (including the Japanese part, in order to provide a complete view). We will hopefully also be able to describe very soon pre-requisite for reusing the public data sets and possibly concrete example of such reuse by third-parties (some contacts have been already taken with the FIESTA-IoT FIRE H2020 project for instance).

5 References

[1] CPaaS.io Deliverable D2.1: “Requirements Specification”.