PROTEUS

ADAPTIVE MICROFLUIDIC - AND NANO - ENABLED SMART SYSTEMS FOR WATER QUALITY SENSING

Data Management Plan

Deliverable D6.2

Author : **F. Le Gall**

June, 22nd 2015

www.proteus-sensor.eu

This project has received funding from the European Union's H2020 Programme for research, technological development and demonstration under grant agreement No 644852

Editor:	Franck Le Gall, Easy Global Market	
Deliverable nature:	Report	
Dissemination level:	Public	
Contractual/actual delivery date:	M04	M09
Suggested readers:	Example: Executives in entertainment companies and banks, investors	
Version:	1.0	
Keywords:	Open data plan	

Abstract

A project data management plan is proposed. It follows the Open Research Data Pilot of H2020. A review of the project datasets took place and lead to the selection of 8 dataset having high potential for publication as open research data .They cover the fields of: sense city sensors, sensor lab benchmarking, energy availability on the water network, energy harvesting system calibration, on site Energy harvester benchmarking, demonstrator dataset, demonstrator dataset, and Communication technology testing.

Disclaimer

This document contains material, which is the copyright of certain PROTEUS consortium parties, and may not be reproduced or copied without permission.

All PROTEUS consortium parties have agreed to full publication of this document.

Neither the PROTEUS consortium as a whole, nor a certain part of the PROTEUS consortium, warrant that the information contained in this document is capable of use, nor that use of the information is free from risk, accepting no liability for loss or damage suffered by any person using this information.

This project has received funding from the European Union's Horizon 2020 Programme for research, technological development and demonstration under grant agreement n° 644852.

Copyright notice

Revision History

Revision	Date	Description	Author (Organisation)
V0.1	June 2015	Creation	F. Le Gall (EGM)
V0.2	August 2015	Complete drafting, partners contributions insertion	F. Le Gall (EGM)
V0.3	September 2015	Added partners revisions	B. Lebental (Ifsttar) P. Vlacheas (WINGS) F. Cottone (NIPS)
V0.4	October 2015	Added partner's contributions. Built release candidate deliverable	B. Almeida (Unparallel) B. Vergne (PONSEL) F. Le Gall (EGM)
V1.0	Octobre 2015	Final review	F. Le Gall (EGM)

Table of content

Executive summary

This deliverable proposes the data management plan (DMP) to be supplied in the context of the H2020 pilot action on open access to research data. Usual practices in relation to data management plan have been looked at and allowed to identify data sets characteristics to be looked at and being: description of data to be generated or collected (explanation of the different types of data which will be produced, including file formats where possible), standards and metadata (presence and description of the metadata accompanying the dataset), data sharing (provided access level) and archiving and preservation (procedures ensure preservation of the dataset).
The target datasets in the context of the H2020 pilot on research data focuses on providing on-line access to scientific information that is free of charge to the end-user. In the context of R&D, 'scientific information' can refer to **peer-reviewed scientific research articles** (published in academic journals) and also to **scientific research data** (data underlying publications, curated data and/or raw data).
A template to describe the relevant project datasets is defined and includes: reference and name, data set description, metadata, quality assurance, data sharing and archiving and preservation, in-line with the H2020 recommendations.

From the analysis of available datasets, 7 have been selected as candidate for the Open Research data Pilot and cover the fields of: sense city sensors, sensor lab benchmarking, energy availability on the water network, energy harvesting system calibration, on site Energy harvester benchmarking, demonstrator dataset, demonstrator dataset and Communication technology testing.

Evolution of these dataset availability will be monitored at project reviews.

1 Introduction

In Horizon 2020 a limited pilot action on open access to research data is being implemented [1] . The requirements for the Pilot on Open Research Data are based around the G8 Science Ministers' Statement on open scientific research data from June 2013, in particular the point that "open scientific research data should be easily discoverable, accessible, assessable, intelligible, useable, and wherever possible interoperable to specific quality standards" [3] . Projects such as PROTEUS which volunteer to participate in this pilot are requested to submit a Data Management Plan (DMP) as introduced in the Horizon 2020 Work Programme for 2014-15:

"A further new element in Horizon 2020 is the use of Data Management Plans (DMPs) detailing what data the project will generate, whether and how it will be exploited or made accessible for verification and re-use, and how it will be curated and preserved. The use of a Data Management Plan is required for projects participating in the Open Research Data Pilot. Other projects are invited to submit a Data Management Plan if relevant for their planned research."

The guidelines further detail information to be provided within the DMP as explained within the University of Bristol guidelines [4] :

- **Description of data to be generated or collected**: explanation of the different types of data which will be produced, including file formats where possible. An accurate description of the dataset to be produced from the start is difficult. It is therefore not expected to list everything which is subsequently produced, but to highlight which datasets the consortium agrees are the most significant and likely to have long-term value. It is also expected to state the origins of any collected data, for whom the data might be useful, and whether or not it will underpin a scientific publication. Highlighting whether similar data already exists and the possibilities for integration and reuse will also be helpful. For these reasons, regular updates of the DMP are expected.
- **Standards and metadata**: a specific attention has to be paid to the metadata accompanying the dataset. Such metadata includes the filenames and folder structure which need to be self-explanatory for 3rd party readers.
- **Data sharing**: The EC require clarity on the level of access that will be provided; will the data be widely and publicly open or will there be restrictions on who can access the data? If the latter is the case access procedures to be put in place need to be outlined. Also, if for any reason there is a need for an embargo period on the data you, this should be clearly stated and rationale given. Information on any software and other tools necessary for enabling re-use of datasets should also be provided.
- **Archiving and preservation**: Procedures which will ensure the preservation of your data, including backup and storage need to be specified. It is important to remember that some cloud based storage options may not be governed by EU legislation (such as the Data Protection Directive), and so would not be suitable for storing sensitive data.

These are the basic requirements for DMP and are further identified and developed, in this document in the context of PROTEUS.

2 Scope

2.1 Description

Before detailing aspects of the DMP, it is required to clearly identify the scope of a DMP in the context of the H2020 pilot on Open Research Data. It is indeed not any data being generated or use by the project which has to be considered. As mentioned in the fact sheet to Open Access in Horizon 2020, open access can be defined as the practice of providing on-line access to scientific information that is **free of charge** to the end-user. In the context of R&D, 'scientific information' can refer to peer-reviewed **scientific research articles** (published in academic journals) and also to **scientific research data** (data underlying publications, curated data and/or raw data).

The notion of openness is central for the selection of research data to be included within the PROTEUS participation to the pilot on open research data. Openness is defined by Open Knowledge [6] as having the following characteristics:

- **Availability and access**: the data must be available as a whole as and at no more than a reasonable reproduction cost, preferably by downloading over the internet. The data must also be available in a convenient and modifiable form.
- **Reuse and redistribution**: the data must be provided under terms that permit reuse and redistribution including the intermixing with other datasets. The data must be machine-readable.
- **Universal participation**: everyone must be able to use, reuse and redistribute — there should be no discrimination against fields of endeavour or against persons or groups. For example, 'non-commercial' restrictions that would prevent 'commercial' use, or restrictions of use for certain purposes (e.g. only in education), are not allowed.

The fact sheet makes clear that Open access is not a requirement to publish, as researchers are free to publish or not, nor does it interfere with the decision to exploit research results commercially e.g. through patenting. Indeed, the decision on whether to publish open access documents must come after the more general decision on whether to go for a publication directly or to seek first protection using Intellectual Property Rights (IPR). This is depicted on Figure 1.

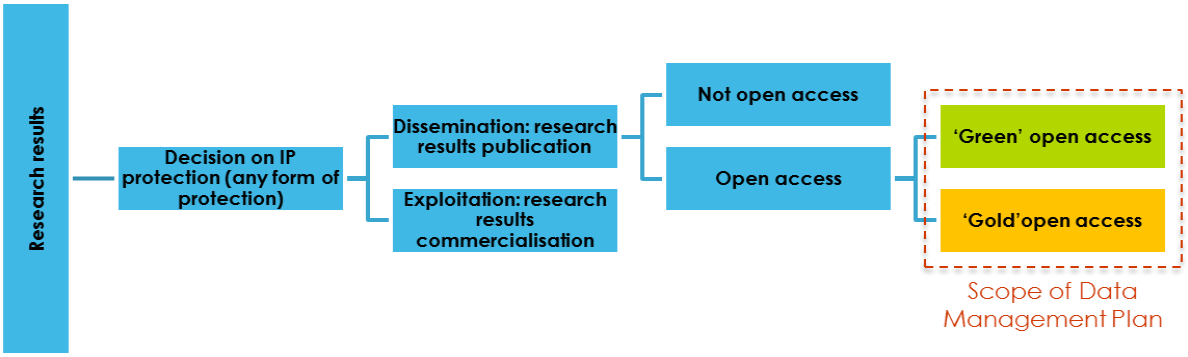

Figure 1: open research data decision process.

The fact sheet also defines two main and non-mutually exclusive routes towards open access to publications:

- **Self-archiving (also called 'Green' open access)** means that the published article or the final peer reviewed manuscript is archived by the researcher – or a representative - in an online repository before, after or alongside its publication. Access to the article is often – but not necessarily - delayed ('embargo period') as some scientific publishers may wish to recoup their investment by selling subscriptions and charging pay-per-download/view fees during an exclusivity period.
- **Open access publishing (also called 'Gold' open access)** means that an article is immediately provided in open access mode by the scientific publisher. The associated costs are shifted away from readers, and instead to (for example) the university or research institute to which the researcher is affiliated, or to the funding agency supporting the research.

2.2 Application to PROTEUS case

The PROTEUS DMP is restricted to the 2 routes described above and elaborates on the characteristics of identified data described hereunder. These are inspired from the on-line DMPTool [7] :

2.2.1 Reference and name

An identifier for the data set has to be produced

2.2.2 Data set description

Description of the data that will be generated or collected, its origin (distinguish between newly collected data and data being re-used from other projects), nature, scale and resolution and to whom it could be useful, and whether it underpins a scientific publication. Information on the existence (or not) of similar data and the possibilities for integration and reuse.

Data types could include text, spreadsheets, images, 3D models, software, audio files, video files, reports, surveys, etc.

2.2.3 Metadata

Reference to existing suitable standards for metadata of the discipline. If these do not exist, an outline on how and what metadata will be created.

As examples, spatial data sets must be documented using either the FGDC version 2.0 or the ISO 19115 metadata standard. The Biological Data Profile standard (associated with FGDC) is very useful for creating documentation of field- and lab-based work. We recommend use of a metadata documentation tool, e.g., Metavist (http://nrs.fs.fed.us/pubs/2737).

2.2.4 Quality Assurance (QA)

This section describes the steps that will be used to process and ensure quality the data. Procedure need to include: data proofing and validation, including data collection, entry, transmission, and storage. Criteria related to quality assurance (e.g. documentation, calibration, validation, monitoring, versioning, etc.)should appear here.

If any, descriptive or analytical statistics to be run on the data for quality assurance should be described.

2.2.5 Data sharing

Description of how data will be shared, including access procedures, embargo periods (if any), outlines of technical mechanisms for dissemination and necessary software and other tools for enabling re-use, and definition of whether access will be widely open or restricted to specific groups. Identification of the repository where data will be stored, if already existing and identified, indicating in particular the type of repository (institutional, standard repository for the discipline, etc.).

This section will include any necessary limitations to protect sensitive data as well as how to ensure security of data considering in particular the EU's Data Protection Directive.

2.2.6 Archiving and preservation

Description of the procedures that will be put in place for short and long-term preservation (data storage and backup) of the data. Indication of how long the data should be preserved, what is its approximated end volume, what the associated costs are and how these are planned to be covered.

3 PROTEUS data sets

An identification of the possible dataset of interest within the PROTEUS, project has been made. All partners have been requested to identify the datasets to be produced in the context of the project. From that list, several datasets have been characterized being in-line with the scope of the PROTEUS DMP as outlined in section 2. These datasets are detailed here after.

It should be noticed that no mention of scientific publication is made here. However, it is clear the as a research project, PROTEUS will issue publications. These will monitored as part of the project dissemination plan.

3.1 Sense city sensors

Reference name	Sense-City sensor data
Contributing partner	Ifsttar
Description	The dataset provides data measured by various sets of Sense-City sensors: 1) Sense-City infrastructure of reference sensors measuring Sense-City conditions as a benchmark (weather stations, air quality, water quality, thermal camera...). 2) Sense-City-deployed PROTEUS sensors 3) Other sets of experimental sensors deployed within Sense-City during other projects Scales, units and resolution of data provided by the sensors vary from one sensor to the other. This information is made available within sensor datasheets. Sense-City reference sensor infrastructure is regularly evolving, as well as the number and type of experimental sensors, so the volume of data is not known. Rough projections indicate that the volume of collected data may reach up to 100Mo/day.
Metadata	Data are stored in ASCII format in MongoDB. They can be extracted manually and automatically from the database into .csv or .txt files with filename and extent of the dataset (time range, selected observables) determined by the end-user. Each sensor is associated to its GPS data in decimal degrees (DD) format. The geoposition of each sensor can be easily extracted independently or included into the data extracted from the database. However, at this time, no standard has been imposed to Sense-City geodata. This may be a later development.
Quality Assurance	Data is directly fed from the sensors to the database. Each data is associated to time & GPS stamp and sensor ID The datasheet, available calibration data (with date and condition of calibration) and deployment notes for each sensor can be downloaded by

	the end user. However, the Sense-City consortium does not guaranty the proper calibration of the sensors at the time of the measurement. As a consequence, raw data (before conversion into the quantity of interest) are kept available.
Data sharing	The principles of data sharing to third parties are still being set by Sense-City consortium, but it is expected that: Data from reference sensor infrastructure and from Proteus sensor is freely available online to freely registered users for non-profit purposes. Access rules will be defined for business usage of the data. The availability of data from other sets of experimental sensors will be defined project by project. Sense-City consortium will promote online availability to freely registered users for non-profit purposes The type of licence is still being discussed within Sense-City consortium. Decision is expected before end of 2015. Data are available via a web interface with access restricted to registered users. Registration is free for non-profit users. Data and subsets of data can be extracted manually and automatically from the database into .csv or .txt files with filename and extent of the dataset (time range, selected observables) determined by the end-user. Tools for online data representation/charts are also available.
Archiving and preservation	Modalities of archival are still being drafted out. Duration of data archival will be at least three years. The system is setup for daily to weekly backup of the data. Sense-City data are already available via the modalities previously described, but the reference sensor infrastructure is coming online progressively. New data are available is quasi real-time.

3.2 Sensor lab benchmarking

Reference name	Data from sensor lab benchmarking
Contributing partner	Ifsttar
Description	This dataset contains the measurements realised to benchmark the sensors, evaluate their cross-sensitivity and assess their reliability. The measurements will follow a predefined experimental plan, identical from sensor to sensor. Data is collected automatically during essays by Labview program and stored in several xls/txt/csv file.
Metadata	Data is made available in folders named by sensor ID. Data is timestamped. Data is accompanied with configuration files providing the relevant parameters of the corresponding essay and with error log providing if needed a list of anomalies having occurred during the essay
Quality Assurance	Automated creation, naming and stamping of the files, automated storage in automatically-named folder by Labview program, including configuration and error files. Creation of detailed description of the essay accompanied with a user guide for the data. Availability of automatically created configuration and error files.
Data sharing	Data is freely available online to registered users for non-profit purposes Access rules remain to be defined for business usage of the data. A licence may be used, the type is to be defined later on Data will be made available electronically upon mail request to the consortium. Access via a web interface is not planned at the time. Automated extraction of a subset of data is not planned.
Archiving and preservation	The data will be preserved at least until the end of Proteus. The plan for data archiving and preserving will be defined during the course of PROTEUS depending on actual use of the data reported within and outside of the consortium partners during the project.

3.3 Energy availability on the water network

Reference name	Monitoring of available energy on the water network
Contributing partner	NiPS
Description	This dataset contains measures realised in the field from Almada water network and provides measures of available energy. Measured sources are: water flowrate, solar energy (lux), vibrations (accelerometer), temperature difference, traffic (piezoelectric pavement). Data is stored in ASCII text files and its size should not be more than 1Gb. The accuracy of the measured parameters depends on the instrumentation sensor, in any case it is below 5%. Most of the parameters, excluding water flux provided by SMAS infrastructure, are acquired onsite. Once the PROTEUS wireless sensor network is implemented, they could be accessed remotely.
Metadata	Measures are timestamped with GPS information and sensor ID At this time, no standard has been imposed to geodata. This may be a later development with other partners.
Quality Assurance	data collected automatically during tests by DAQ Card/LabView and stored in files /xls/txt/csv Most of the commercial sensors such as the ones for acceleration, luminosity, temperature are provided with certificate of calibration and datasheet that we make available on the PROTEUS website together with corresponding data. The other sensors will be calibrated in laboratory according to standard methodology.
Data sharing	Data is freely available online to registered users for non-profit purposes Access rules remain to be defined for business usage of the data. No license is required for the raw sensor data. The post-processed data license will be discussed within PROTEUS consortium. Data are available via a web interface with access restricted to registered users. Registration is free for non-profit users. Data and subsets can be extracted manually and automatically from the database into .csv or .txt files with filename determined by the end-user. At the moment, online tools for analysis and representation are not foreseen. This point will be discussed within PROTEUS consortium.
Archiving and preservation	Archiving methodology and preservation are still being drafted out. The system is setup for daily to weekly backup of the data. Data will be made available online after verification and validation. Duration of data archival will be at least three years.

3.4 Energy harvesting system calibration

Reference name	Energy harvesting system characterization in laboratory
Contributing partner	NiPS
Description	This dataset contains characterization of energy harvesting system (EHS) made in laboratory. Data is stored in various formats (e.g. ascii, txt, Jpeg, AVI, docx, xlsx, pptx) its size varies from a few Mb per measurement to 1 Gb per device characterization, depending on duration of lab tests. The accuracy of the measured parameters depends on the instrumentation sensor, in any case it is below 5%.Q? How to access sensors information?
Metadata	Measures are timestamped with sensor ID (No need of GPS data for EHS characterization in lab)
Quality Assurance	Data is collected automatically during tests by DAQ Card/LabView and stored in files Tools for EHS characterization will be provided with certificate of calibration and datasheet when available together with corresponding data. The home-made instruments will be calibrated in laboratory according to standard methodology. Before publication, data will be validated and verified by comparing them with different measurement techniques.
Data sharing	Subsets of EHS characterization data (e.g. EH power performance) will beData freely available online to registered users for non-profit purposes Access rules remain to be defined for business usage of the data. No license will be needed for chosen data subsets the EHS characterization. Other characteristics related to EH design and working principle will be made available at specific request after check on intellectual property rights. Data are available via a web interface with access restricted to registered users. Registration is free for non-profit users.
Archiving and preservation	Archiving methodology and preservation are still being drafted out. The system is setup for daily to weekly backup of the data. Data will be made available online after verification and validation. Duration of data archival will be at least three years.

3.5 On site Energy harvester benchmarking

Reference name	On site Energy harvester benchmarking campaign
Contributing partner	NiPS
Description	This dataset contains characterization and comparison of energy harvester deployed in the field. Data is stored in various formats: txt, .xlsx, for collecting and docx, pptx, tex to be converted into .pdf for open publications (e.g. ascii, txt, Jpeg, AVI, docx, xlsx, pptx). Its size is up to 100Mb/s per energy harvester device. The accuracy of the measured parameters depends on the instrumentation sensor, in any case it is below 5%. Once the PROTEUS wireless sensor network will be implemented, the characterization data of EHS will be accessible remotely (the accuracy, range and features of the sensors will be provided in the datasheet and calibration certificate).
Metadata	Measures are timestamped with GPS information and sensor ID At this time, no standard has been imposed to geodata. This may be a later development with other partners.
Quality Assurance	Data collected by measurements onsite on the energy harvesting systems and elaborated by post-processing analysis. Manipulation and/or conversion of raw data into easy-to-access data. For instance from acceleration, pressure, water and solar flux into available and generated electrical energy. Graphical representation. Before publication, data will be validated and verified by comparing them with different measurement techniques.
Data sharing	Reports published through main stream scientific/technical journals, technical notes, presentations, freely available online through the PROTEUS website and free newsletters of NiPS (ICT Energy Letters) Raw data sets used will be made available to 3rd parties so they can reproduced published results. Data are available via a web interface with access restricted to registered users. Registration is free for non-profit users. At the moment, online tools for analysis and representation are not foreseen. If necessary, dataset will be structured to be suitable for filtering/visualization tools developed within the PROTEUS consortium.
Archiving and preservation	Archiving methodology and preservation are still being drafted out. The system is setup for daily to weekly backup of the data. Data will be made available online after verification and validation. Duration of data archival will be at least three years.

3.6 Demonstrator dataset

Reference name	Data from warter monitoring demonstrator
Contributing partner	Ponsel
Description	This dataset contains data from sensors installed in real conditions for long term validation (step before pilote installation) Its size is about few Ko/sensor/day. The data will store using our dataloggers to PONSEL databases. The format is not defined yet but expected excel file or .txt format Q? What are the accuracy and ranges of measured parameters? What is the schematic of sensors deployment ? Accuracy and ranged of published sensor data is not yet known as depending of the application and the performances of the sensor. Sensor information will be accessed through the PONSEL SCADA (Arlequin).
Metadata	Measure are stored with a timestamp, sensor ID and datalogger ID No geodata is available at that time but it could be possible if needed
Quality Assurance	Data is collected automatically by Ponsel dataloger and supervision There is postprocessing for customers in order to give a accurate measurement. PONSEL will access to non postprocessing treatment Sanitization check is performed in order to eliminate bad values
Data sharing	Data is freely available through the demo system. A web interface is provided to access and filter the dataset.
Archiving and preservation	Work is undergoing with an external supplier (OVH) to archive all PONSEL information

3.7 Demonstrator dataset

Reference name	Water operator's monitoring priorities
Contributing partner	Wings
Description	This dataset will describe the water operator's monitoring priorities, requirements, control levels, monitoring profiles This information is provided as C/C++ Header file for in-node definition or Config file (e.g. JSON, INI, TXT) for in-SCADA definition. The volume will be of ~1KB for 1 node initialization Accuracy and range of provided parameters will be also included in the dataset.
Metadata	In files comments Each dataset may have the geo information of the area that the sensor node will be located in the name of the file.
Quality Assurance	Data will be tested during trial phase of PROTEUS
Data sharing	User friendly representation will be provided so as to be read easily through a text editor. Data will be made available electronically upon mail request to the consortium. Access via PROTEUS web site will be investigated.
Archiving and preservation	The data will be preserved at least until the end of Proteus. The plan for data archiving and preserving will be defined during the course of PROTEUS depending on actual use of the data reported within and outside of the consortium partners during the project and after agreement with SMAS which is the actual data provider.

3.8 Communication technology testing

Reference name	Data from testing with different communication technologies
Contributing partner	UI
Description	This dataset contains the results obtained from the testing of several communication technologies. The testing focuses in the characterization of energy consumption of different communication technologies for different values signal amplification, communication range, message payload, bitrates, and presence of obstacles. The energy consumption is measured using the current required by a device that is exchanging communication messages. The current is measured in milliampere (mA) and is analysed over time, in seconds. Measurements have a range of ±3.2 A with resolution of 0.8 mA. This dataset will be published in the context of a master thesis, being data presented in tables and graphs. This dataset may be of interest to anyone that needs to compare different communications technologies to in order to select the must suitable for his or her specific use case.
Metadata	Identification of the communication modems and types of antennas used in each testing.
Quality Assurance	Power consumption data are automatically collected by a device equipped with a current sensor. According to the datasheet of the current sensor, measurements are performed with 1% precision.
Data sharing	As this dataset is produced in the context of a master thesis, it will be under the copyright of the author, the college *Faculdade de Ciências e Tecnologias*, and university *Universidade Nova de Lisboa.* College and University have rights to archive and publish the dissertation in any support (paper, digital, etc.) and allow its distribution and use for educational or research, non-commercial purposes. A digital version of the dissertation, and therefore the dataset, will be available the on the University's online repository (run.unl.pt). Creative Commons License protects all documents in this repository. A paper version will be available on the college's library.
Archiving and preservation	Document preservation is managed by the university, where it is not defined any expiration period for documents inserted into the repository.

4 Conclusion

From the guidelines, DMP should be reviewed at regular intervals. This will be synchronised with project reviews planning. Updates to the DMP will appear within the dissemination report of the project.

Reviews will go further the information presented in section 3 by answering the following questions related to the produced scientific research data:

- **Discoverable**: are the data and associated software produced and/or used in the project discoverable (and readily located), identifiable by means of a standard identification mechanism (e.g. Digital Object Identifier)?
- **Accessible**: are the data and associated software produced and/or used in the project accessible and in what modalities, scope, licenses (e.g. licencing framework for research and education, embargo periods, commercial exploitation, etc.)?
- **Assessable and intelligible**: are the data and associated software produced and/or used in the project assessable for and intelligible to third parties in contexts such as scientific scrutiny and peer review (e.g. are the minimal datasets handled together with scientific papers for the purpose of peer review, are data is provided in a way that judgments can be made about their reliability and the competence of those who created them)?
- **Useable beyond the original purpose for which it was collected**: are the data and associated software produced and/or used in the project useable by third parties even long time after the collection of the data (e.g. is the data safely stored in certified repositories for long term preservation and curation; is it stored together with the minimum software, metadata and documentation to make it useful; is the data useful for the wider public needs and usable for the likely purposes of non-specialists)?
- **Interoperable to specific quality standards**: are the data and associated software produced and/or used in the project interoperable allowing data exchange between researchers, institutions, organisations, countries, etc. (e.g. adhering to standards for data annotation, data exchange, compliant with available software applications, and allowing re-combinations with different datasets from different origins)?

5 References

[1] “Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020”, http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf

[2] “Guidelines on Data Management in Horizon 2020”, http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

[3] “G8 Science Ministers’ Statement”, 12 June 2013 https://www.gov.uk/government/news/g8-science-ministers-statement

[4] “Data Management Planning - EC Horizon 2020 Pilot on Open Research Data applicants”, University of Bristol, 2014, http://data.bris.ac.uk/files/2014/04/data.bris-ECHorizon2020-guide.pdf

[5] “Fact sheet: Open Access in Horizon 2020”, December 2013, https://ec.europa.eu/programmes/horizon2020/sites/horizon2020/files/FactSheet_Open_Access.pdf

[6] “What is Open?”, Open Knowledge, https://okfn.org/opendata/

[7] DMPTool, University of California Curation Center of the California Digital Library, https://dmptool.org/

6 Annex: identified datasets

Data to be collected		Data creation & collection				Support		
Data description	Format & volume	Standards & methods	Structure	Versioning	QA	Exploitation	Tools	Metadata
Data from Sense-City sensors (PROTEUS sensors+Sense-City sensor infrastructure)	up to 100Mo/day; ASCII	data collected automatically by supervisor and stored in MongoDB database	each data associated to time & GPS stamp and sensor ID		quality of data depends on sensor quality ==> data associated with sensor datasheet	data freely available online to registered users for non-profit purposes to be definied for business usage of the dazta		
Data from sensors lab characterization	from a few Mo/sensor to 10Go/sensor depending on extent of characterization; ascii or tiff	data collected automatically during essays by Labview program and stored in several xls/txt/csv file	automatic labelling of file names by Labview program with time stamp automatic storage into folder named by sensor ID automatic creation of configuration files for each essay directly stored into the folder	not relevant (all the essays are stored for later analysis)	definition of process for data management by the software and the user	data freely available for reaseach purposes		
Benchmarking data for each sensor	at most a few Mo/sensor; format to be defined (probably .xls file plus .doc report to be converted into .pdf)	Two options: 1) during development of methodology, data created from analysis of extensive lab characterization 2) after stabilization of methodology, automatic creation by Labview program from sensor lab-scale characterization	standard format to be defined for the .xls file and the .doc report	ime stamp and number of version in file name	to be defined with sensor distributor and end user	data communicated to the distributor who will decide modality of dissemination and exploitatin		
Data from sensors for the monitoring of available energy on the water network: water flowrate, solar energy (lux), vibrations (accelerometer), temperature difference, traffic (piezoelectric pavement)	ASCII, txt; max 1 Gb	data collected automatically during tests by DAQ Card/LabView and stored in files /xls/txt/csv	each data associated to time & GPS stamp and instruments ID	data collected by experimental tests in situ	quality of data depends on instrumentation precision and sensor characteristics	data freely available online to registered users for non-profit purposes to be definied for business usage of the data	Data preserved in storage systems: hard disks, pen drives, cloud, web pages	

Data to be collected		Data creation & collection				Support		
Data description	Format & volume	Standards & methods	Structure	Versioning	QA	Exploitation	Tools	Metadata
Data from energy harvesting system characterization in laboratory	ascii, txt, Jpeg, AVI, docx, xlsx, pptx from a few Mb per measurement to 1 Gb per device characterization depending on duration of lab tests;	data collected automatically during tests by DAQ card, Oscilloscope, Multimeters and stored in several xls/txt/csv file	automatic labelling of file names by acquisition software with time stamp; automatic storage into folder named by DEVICE type; summary of test results in document format docx, pptx, jpeg, png	not relevant (all the essays are stored for later analysis)	data management performed by the user	data freely available after specific request for reasearch purposes	Data preserved in storage systems: hard disks, pen drives, cloud, web pages	
Onsite benchmarking and reporting data for each energy harvesting device	up to 100 Mb per Energy harvesting device formats .txt, .xlsx, for reporting: docx, pptx, tex to be converted into .pdf for open publications	data collected by measurements onsite on the energy harvesting systems and elaborated by post-processing analysis	filed for benchmarking and reporting in format .xlsx, .docx, pptx to be converted in .pdf for open access	time stamp and number of version in file name	Quality of data depends on measurement tools and post-processing by the user	reports published through main stream scientific/technical journals, technical notes, presentations, freely available online through the PROTEUS website and free newsletters of NiPS (ICT Energy Letters)	Data preserved in storage systems: hard disks, pen drives, cloud, web pages	
4	1 Mo to 10 Mo/day. Format must be defined	data collected automatically form our R&D computers	not defined yet (excel ?)	rot relevevant	data associated with tests and conditions of the test	Only for Ponsel R&D team and Proteus Team		
Calibration data from sensors during the manufacturing of the sensor	few Ko to 1 Mo/sensor	data collected during the calibration	.csv or integration into a database	not relevant	data associated with sensor SN	Only for Ponsel R&D team		
Data from sensors installed in real conditions for long term validation (step before pilote installation)	few Ko/sensor/day	data colected automatically by Ponsel dataloger and supervision	database, associeted to time, sensor ID and datalogger ID	not relevant	Depend of the results and interest of the end user	datafreely available (demo system)		

Data to be collected		Data creation & collection				Support		
Data description	Format & volume	Standards & methods	Structure	Versioning	QA	Exploitation	Tools	Metadata
CMOS chip (SoC) database: achip circuit and system schematics is collected in a special database, OpenAccess DB,	OpenAccess DB. This is the database supported by the chip design enviroment.	OpenAccess DB (http://www.si2.org/)		SVN		Data will be stored in the UNINOVA private servers. Restrict access due to IP and signed NDA with the silicon foundry.	Chip design EDA tool: Cadence IC6	
CMOS chip (SoC) mask layout: chip mask information is collected in a special file format. This file is used by the semiconductor foundry to built the chip	gdsII.This is the mask layout data file supported by the Foundry.	gdsII		SVN		Data will be stored in the UNINOVA private servers. Restrict access due to IP and signed NDA with the silicon foundry.	Chip design EDA tool: Cadence IC6	
R&D scientific information created duri	Latex, Word, PDF	Follow closely the IEEE publication templates		SVN (for Latex)	Content quality validation by the IEEE peer reviewing process	Accepted publications available form IEEE explore	PDF reader, MS word	
Water operator's monitoring priorities, requirements, control levels, monitoring profiles	Volume: 1 file (~1KB) used for node initialization. Format: C/C++ Header file for in-node definition or Config file (e.g. JSON, INI, TXT) for in-SCADA definition	C/C++ Header file or Config file (e.g. JSON, INI, TXT)	C/C++ Header file or Config file (e.g. JSON, INI, TXT)	Node identifier at the file name	Data quality depends on water operator policies	Open (high level description exists in D1.1)	User friendly representation will be provided so as to be read easily through a text editor	C/C++ Header file or Config file (e.g. JSON, INI, TXT)
Extrapolation models and functions	Volume: It depends on where extrapolation will run; in the node or in SCADA; and on the number of critical parameters to extrapolate (at least 1 - the more critical parameter per use case). Format: C/C++ Header and Source files	C/C++ Header and Source files	C/C++ Header and Source files	In-node/In-Scada execution identifier and corresponding parameter and use case in the file name	Comparison of predicted value with real value from SMAS (or even PONSEL)	Some extrapolation results (predicted/processed data) together with the associated raw data (real data from measurements) may be considered to be available in some open mode	Guidelines on input/output of models/functions will be specified	C/C++ Header and Source files
State machine specification	Volume: 1 file. Format: State diagram	Word, Powerpoint	State diagram	N/A	Behaviour tested in lab scale, model deployment and real deployment	Open (Description will exist in D1.2)	Description will exist in D1.2	Description (definition of states, states input and output, state transitions) will exist in D1.2