H2020 Research and Innovation action

Topic: MG-8.3-2016 Assessing future requirements for skills and jobs across transport modes and systems

Grant Agreement number: 723989

Start date: 01 October 2016

Duration: 36 months

Skills and competences development of future transportation professionals at all levels

SKILLFUL

Deliverable D5.2

Data Management Plan

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 723989

Project Coordinator	Dr. Thierry Goger, FEHRL, Bld de la Woluwe, 42/b3, 1200 Brussels, Belgium. Tel: +32 2 775 82 34 E-mail: thierry.goger@fehrl.org Website: www.skillfulproject.eu

Main Editor(s)	Dr. Evangelos Bekiaris, CERTH/ HIT, Greece +30 2310 498261, abek@certh.gr Matina Loukea, CERTH/ HIT, Greece +30 211 10 69 556, mloukea@certh.gr
Due Date	30/09/2017 (Month 12)
Delivery Date	
Task number	A5.1
Dissemination level	Public

Contributor(s)

Main Contributor(s)	Athina Dimou, CERTH/ HIT, Greece Panou Maria, CERTH/ HIT, Greece

Review

Reviewer(s)	1. Arto Kyytinen, TTS, Finland 2. Dana Sitanyiova, UNIZA, Slovakia

Control Sheet

Version History			
Version	Date	Editor	Summary of Modifications
v1.0	20/09/2017	Evangelos Bekiaris	1st version of D5.2
V1.1	20/10/2017	Evangelos Bekiaris	2nd version of D5.2
V1.2	24/10/2017	Matina Loukea	3rd version of D5.2, integrating comments and feedback from SKILLFUL 2nd Plenary Meeting
V1.3	26/10/2017	Matina Loukea	4th version of D5.2, integrating comments and feedback from reviewers

Final Version released by		Circulated to	
Name	Date	Recipient	Date
Claudia Ciuca	06/11/2017	Coordinator	30/10/2017
		Consortium	06/11/2017
		European Commission	06/11/2017

Table of Contents

Abbreviations

Abbreviation	Meaning
CD	Compact Disc
D	Deliverable
DMP	Data Management Plan
DVD	Digital Versatile Disk
EIs	Evaluation Indicators
FAIR	Findable, accessible, interoperable and re-usable
GHz	Giga Hertz
h/w	Hardware
ID	Identification
KPIs	Key Performance Indicators
M	Month
MaaS	Mobility as a Service
O	Objective
Q	Quantitative
QL	Qualitative
S	Subjective
s/w	Software
USB	Alternate Storage Memory
WP	Work package

List of Tables

Executive Summary

This report constitutes the Deliverable 5.2 of SKILLFUL project, part of the WP5 (Pilots). SKILLFUL project aims to identify the skills and competences needed by the Transport workforce of the future (2020, 2030 and 2050 respectively) and define the training methods and tools to meet them. Within this context SKILLFUL will develop new training schemes, which will be adapted to the particular needs of the transportation professionals of the present and the future.

These training schemes will be validated through the realization of Pilots in 13 different Pilot sites in Finland, Spain, Belgium, Netherlands, UK, Germany, Italy, Slovakia, and Ireland, as defined so far.

A preliminary Data Management Plan (DMP) was prepared for the service and pilot data to be collected during the project, following the regulations of the Pilot action on Open Access to Research Data of Horizon 2020.

In this version of the Deliverable (that will be updated on Months 24 & 36 of the project) the necessary aspects of the Data Management Plan framework are set:

- **Data types**
- **Data privacy policy**
- **Data access**
- **File naming procedures and ownership**
- **Archiving and preservation**

The Data Management Plan is a deliverable directly connected to forthcoming evaluation and pilot plans for each of the pilot sites (WP5) and the decisions, taking into consideration feedback from the regional Ethics committees and Data Protection Authorities. The final version of the deliverable will include description for the complete datasets that will be created during the pilots and the analyses to follow.

1 Introduction

SKILLFUL (http://skillfulproject.eu/) is dealing with one of the main challenge for the transportation sector, which is the ability to attract new employees, as well as equip the existing ones with the competences required for addressing the needs of the constantly changing and developing transportation sector. Its vision is to identify the skills and competences needed by the Transport workforce of the future (2020, 2030 and 2050 respectively) and define the training methods and tools to meet them. Within this context, the project objectives can be described as following:

- ✓ to critically review the existing, emerging and future knowledge and skills requirements of workers at all levels in the transportation sector, with emphasis on competences required by important game changers and paradigm shifters (such as electrification and greening of transport, automation, MaaS, etc.);
- ✓ to structure the key specifications and components of the curricula and training courses that will be needed to meet these competence requirements optimally, with emphasis on multidisciplinary education and training programmes;
- ✓ to identify and propose new business roles in the education and training chain, in particular those of "knowledge aggregator", "training certifier" and "training promoter", in order to achieve European wide competence development and take-up in a sustainable way.

For the aforementioned objectives to be achieved, the whole project process its structured that way that it can be divided into three major categories/ steps:

- **Step 1**: Identification of Future Trends/ Needs & Best Practices
- **Step 2**: Development of Training Schemes & Definition of Profiles and Competences
- **Step 3**: Verification and Optimization of training schemes

During the third step of the SKILLFUL project and the procedure of the training schemes piloting and verification, data will be gathered during the pilots to be conducted as part of WP5 activities. The Data Management Plan is a deliverable directly connected to forthcoming evaluation and pilot plans for each of the 13 pilot sites (WP5).

Post-processed datasets free from any private/personal and identifiable information will reside in the SKILLFUL data repositories, which will be described in the next revised versions of this deliverable (M24 & M36). The final version of the deliverable will include description for the complete and shareable datasets that will be created during the pilots and the analyses to follow.

Data Management Plans (DMPs) are a key element of good data management. A DMP describes the data management life cycle for the data to be collected, processed and/or generated by a Horizon 2020 project. As part of making research data findable, accessible, interoperable and re-usable (FAIR), a DMP should include information on:

- the handling of research data during and after the end of the project
- what data will be collected, processed and/or generated
- which methodology and standards will be applied
- whether data will be shared/made open access and
- how data will be curated and preserved (including after the end of the project).

This version of Deliverable 5.2 consists a preliminary DMP for the service and pilot data to be collected during the piloting of the training schemes that will be developed within the project, following the regulations of the Pilot action on Open Access to Research Data of Horizon 2020 [1].

In view of the next version of D5.2 (M24), a service data template will be circulated to all Partners pilot sites and training courses developers) in order to collect information and descriptions of data and metadata types to be collected for each one of the training schemes.

2 Data processes

In order to identify and define the data management procedures, which the SKILLFUL project will follow based on the Guidelines on Data Management in Horizon 2020 document. This version of the deliverable is a preliminary Data Management Plan encompassing the primary aspects to be addressed within the project with two following updates (M24 & M36). The first update will include refined information and descriptions of data and metadata types to be collected for each one of the training schemes with agreed naming conventions, dataset structures and standards to be applied (if any). The second and final update (M36) will contain analytic descriptions of dataset structures with refined restrictions and embargos (if any) for parts/segments of data not only for schemes but for the evaluation indicators per pilot site.

2.1 Data types

Before moving on with describing the initial data management process of SKILLFUL, first, the definitions of data and metadata are provided, as they are used in this project.

> ***"Data** is a set of values of qualitative or quantitative variables; restated, pieces of data are individual pieces of information. Data is measured, collected and reported, and analysed, whereupon it can be visualized using graphs or images or other analysis tools. Data as a general concept refers to the fact that some existing information or knowledge is represented or coded in some form suitable for better usage or processing."[2]*

Data types can be raw, reduced, categorised, analysed, surrogated and in many different forms. They often can be largely categorised in primary clusters of levels of details (e.g. primary, secondary, etc.). Data are categorized in *qualitative* and *quantitative*, and such categorization derives from statistics. Both data types can be *subjective* and *objective*. Neither is exclusive or inclusive. The collection of qualitative data focusses on descriptions like participant's properties, characteristics, features, etc. Quantitative data come from the world of measuring amounts, numbers, sizes, etc. Subjectivity refers to the quality of certain information (i.e. opinion) of the participants referring to experiences, feelings, beliefs, desires, perspectives, etc. Objectivity comes from the attempt of scientists to measure aspects of the natural world without any involvement of emotions, personal biases, and a priori commitments.

The following examples show the possible combinations of qualities [3]:

- **Quantitative/Objective (Q/O):** "The chip speed of my computer is 2 GHz."
- **Quantitative/Subjective (Q/S):** "On a scale of 1-10, my computer scores 7 in terms of its ease of use."
- **Qualitative/Objective (QL/O):** "Yes, I own a computer."
- **Qualitative/Subjective (QL/S):** "I think computers are too expensive."

Occasionally, service/tools might be related to more than one data type collection. Apart from data, metadata will be collected to define the characteristics and in many cases to facilitate processing, storing, and, finally, understanding the data collected during the pilots. Metadata definitions range from quality descriptions of datasets when they are used by analysts who did not participate in the data collection, thus, it is important for them to understand as much as possible about the related processes and procedures to aggregation of data to something different (e.g. values over a threshold might be translated to impairment and volume of impairment). The following definition encompasses more than one use of the term.

*"**Metadata** is data that describes other data. Meta is a prefix that in most information technology usages means "an underlying definition or description." Metadata summarizes basic information about data, which can make finding and working with particular instances of data easier." [4]*

A template will be created and circulated to pilot sites to be completed with information and descriptions about the data and metadata that will be collected during their pilots, as during the current phase of the project the final Pilots form and content are still being formulated. Thus, the types and characteristics of the data and metadata will alter and will be enriched during the development of the training schemes (within WP3) and of the character of the courses that will be evaluated by the SKILLFUL Pilots. The type of data and metadata that will be defined will lead to the final formulation of the proper data management strategy, which will be also formulated to a document signed by all the Pilot sites before the implementation of the Pilots.

This template will aim to collect data and information about the following:

- **Data (reference and name):** the type of data, the name and the reference they will use during collection.
- **Data description:** Brief description of data; what they are and what they represent.
- **Metadata:** To define any data to be created in order to describe and provide information about the raw data to be collected.
- **Standards:** Internal (and not only) standards they apply for collecting and processing data (if applicable) and reference/compliance to existing standards **(**i.e. reference to known standards that they comply when collecting, processing, and storing data (they are meant to be universal and known standards besides internal procedures).
- **Privacy, confidentiality restrictions:** restrictions that may govern the data collection, processing and sharing related to both local, national and international data privacy and confidentiality legislation and guidelines. These dimensions are very closely related to the SKILLFUL ethics policy (see D7.2: "Ethics and Privacy Protection Manual") which includes and manages any related activities. The Ethics Board will be responsible for overviewing any data sharing protocols and procedures based on data-specific restrictions, if such apply.
- **Archiving and preservation:** techniques and procedures to categorise and storing and backing-up data.

It will be an open-field in the template for each service responsible to customise and add information based on the specificities of each Pilot, existing at this stage.

2.2 Evaluation data

Moreover, a large part of data that has not been defined yet are the data that will be developed for evaluation purposes and will be specified in the evaluation framework as well as the experimental plans. The following tables provide an overview of the evaluation procedures (Table 1) and the evalualtion criteria categories (Table 2). These are not exhaustive lists and mainly depict the clusters of data the projects aims to collect. The lists will be updated and refined (i.e. including data information, restrictions placed by data owners as soon as the "Pilot Plans" are ready (D5.1, M16).

As indicated from the Table 1, the feedback for the evaluation will be obtained by three sources/ groups:

- ✓ Trainers
- ✓ Trainees
- ✓ Observers (through stakeholders walkthroughs)

Table 1: Initial list of Evaluation Procedures

Evaluation Procedures	Thresholds
Interest of participants (No of trainees)	Less than 50% of class → 0
	Less than full class → 1
	Cover the full class → 2
Trainees evaluation (use of evaluation forms)	Overall satisfaction (across several categories) > 70%
Trainers evaluation (use of evaluation forms)	Overall satisfaction (across several categories) > 70%
Observers rating (use of evaluation forms)	Overall satisfaction (across several categories) > 70%
Success rate in exams	> 50%
Group discussions with trainers, trainees' representatives, Pilot leaders and 3-5 stakeholders' representatives	> 70%

From the processes listed below and especially by the group discussion and the stakeholders' walkthroughs (observers' ratings), the following feedback is expected to be collected:

- Comments for further improvements;
- Comments for missing content;
- Further needs for trainees/ trainers competences;
- Needs in training methodology and equipment.

Table 2 below, includes the evaluation indicators that will be used for the evaluation of the SKILLFUL Pilots.

Table 2: Indicative list of Evaluation Indicators (EIs)

Indicative list of Evaluation Indicators (EIs) to be collected during the pilots
- Usefulness for the job (perceived by trainees and stakeholders) - Inclusiveness of curriculum (perceived by all groups) - Training methodologies/ tool adequacy (perceived by trainees and trainers) - Duration and clustering adequacy (perceived by trainees and trainers) - User acceptance (perceived by trainees and trainers) - Usability ratings (perceived by all groups) - Assessment of learning outcome (perceived by all groups) - Potential for accreditation of courses, modules, programmes (perceived by trainers and stakeholders) - Multi lingual aspect for teaching and learning/training (perceived by all groups) - Qualification of trainers (perceived by trainees)

Sharing these type of data, altogether will foster innovation in training products and processes and in parallel facilitate the innovation chain. Ultimately this will produce benefits for the training schemes final users – the transportation professionals.

2.3 Data storage and back up

Data collected by the pilots have to be securely stored and regularly backed-up. Sometimes multiple copies should be made, especially for large datasets that need to be stored in large capacity external hard drives. A separate checklist has been prepared and should be used by all training providers. Any data that will be stored as a result of regular checks and tests performed by the training schemes responsibles wish to perform regular checks and tests and want to be able to create a database:

Checklist:

- ✓ How will the raw data be stored and backed up during the research?
- ✓ How will the processed data be stored and backed up during the research?

- ✓ Which storage medium will you use for your storage and backup strategy? Network storage; personal storage media (CDs, DVDs, USBs, portable hard drives); cloud storage?
- ✓ Are backups made with sufficient frequency so that you can restore in the event of data loss?
- ✓ Are the data backed up at different locations?

Each site and training scheme responsible should ensure that the research data are regularly backed-up and they are stored in secure and safe location. There is a common "rule-of-thumb" to only store data that you actually need in three different copies. It is advised that copies can be stored in both local and remote storage units/locations.

The following data storage options can be used:
External hard drives/USB sticks: will be used in long-trials (WP5) and local evaluations. They will serve as backups and intermediate storage units before transferring data to a permanent/long-term storage place.
Personal computers and laptops: Similarly they will mainly serve as a short-term options and for transferring data after the evaluation sessions to a selected storage place.
Network/fileservers: large data sets will be stored and they will serve as the long-term storage solution. Regular backups will ensure data are not lost or corrupted.
Cloud storage: only aggregated, anonymised and confidential data will be stored on the project cloud storage, depending on the level of agreement between partners who have access to these data. In general, data will be stored that the individual cannot be identified by the shared information and data.

2.4 File naming

File naming depends largely on the training service and the datasets to be derived by this service and/or connected with this service. They have to be **consistent** and **descriptive**.

Partners using this file naming rationale will find it easier to work (and share) the correct version of data and accompanying metadata files. The following file naming offers a consistent naming of the files in order to make it easier to identify, locate and retrieve the data files.

This file and folding naming system will be used for all data and metadata files.
Project acronym: SKILLFUL
Tranining course/ scheme related (in terms of relevant Activity, as described in the DoA)
Researcher name/initials: Matina Loukea (i.e. ML)
Pilot identifier: e.g. VTT for VTT Pilot site
Date or range of pilot: 191017 or 191017-301217
Type of data: UF for users feedback
User group: PO for Port Operators
Version number of file: Only singular number are acceptable (1, 2, 3)
Three letter file extension for application specific files (e.g. xls)

Each data folder will include a regularly updated README.txt in the directory to explain the codes, abbreviations used and, in general, the coding practices and naming conventions used.

Based on the example used above, an efficient naming convention within the SKILLFUL project looks like that:

SKILLFUL_3.4_ML_CERTH_19102917_UF_PO.xls

3 Data sharing and access

At each pilot site a nominated person will be responsible for overseeing that data are safe and secure.

3.1 Data access

One person will have access to full datasets (e.g. higher authorisation level) and the rest of the data team will have medium or lower level of authorisation. Data will be stored in secure areas (physical, network, cloud-based). Higher level of authorisation is granted only for sensitive and personal data. Data to be shared for analysis will not include any personal or identification data. These data, of course, cannot be shared with external databases for further re-use.

Data collection, storing, accessing, and sharing abides to the international legislation (Data Protection Directive 95/46/EC "on the protection of individuals with regard to the processing of personal data and on the free movement of such data" and EU general data protection regulation 2016/679 (GDPR), which will take effect in May 25 2018 and will in which will be more emphasis on at the next version of this deliverable) and guidelines. Different levels of authorisation will exist also for remotely accessing data. High level access to data will not be possible outside the work premises, as they are defined at each pilot site.

Use of cloud store data will be available for medium and lower level of access. Not all individuals will have the same access privileges in order to avoid data corruption, loss and damage. Dataset owners will have full access (read, write, update, delete), however, individuals who want to use/reuse the dataset will be able to read and download but not make any changes or modifications to the specific dataset. The main restrictions with regards to confidentiality are the following:

- Name and personal data
- Raw video and audio recordings
- Data concerning the performance of professionals to activities related to their job/ position

These data are identified based on the initial data pools set by partners responsible for training schemes/courses/modules. Other data restrictions might arise during the development of the training courses.

3.2 Data ownership

Any data gathered during the lifetime of the project are the ownership of the beneficiary or the beneficiaries (joint ownership) that produce them according to subsection 3, Art. 26 of the signed Grant Agreement (723989-SKILLFUL). The beneficiaries have the intellectual property rights of the data they collect and re-use of data is defined by the limitation they might set in how they will make data available. This means partners decide if they make data open-access (no additional restrictions on access to data or publications) or there is an embargo period, whereby permission for accessing the data is given after a certain period of time. As datasets have not been formed yet and training schemes/ courses/ modules are to be developed, therefore this information will be available in an updated version of this deliverable.

3.3 Data Preservation and Archiving

Data will be preserved in each Pilot site's database after the end of the project (for a period of two years) only for complete datasets that partners have agreed to share them with other researchers (if any). Partners will decide, and will be reported in the next version of this deliverable, for how long they would like to retain the data.

Decisive factors are different per Pilot and partner as these applications. Any related costs for archiving and preserving data -especially for long periods of time- they will be checked for their justification and then incurred during the lifetime of the project. However, as each validated training

course may create multiple datasets-and in some occasions- datasets may be connected because training schemes will be connected, such costs will be discussed per pilot and dataset in the second revision of this deliverable (M24).

4 SKILLFUL data privacy policy

Participants' personal data will be used in strictly confidential terms and will be published only as statistics (anonymously).

All pilot data will be anonymised. Only one person per site (relevant Ethical issues responsible) will have access to the relation between test participants' code and identity, in order to administer the tests. One month after the pilots end, this reference will be deleted, thus safeguarding full anonymisation of results.

The stored data will only refer to users' age, gender and nationality (no other identifier will be kept). Nevertheless, stored date relate only to users' activities related to their specific position and job, not to a person's beliefs or political or sexual preferences. Moreover, it is very important to also mention that any data related to the performance of each pilot participant/ user to their job/ position duties ("incidental findings") it is not part of the SKILLFUL research and thus will not be taken into account and no relevant information will be disclosed to any 3rd party; including the trainees' colleagues and management. The only performance element that may be taken into account will be to measure performance before and after the pilots' implementation and hence the implementation of the training programs. But even in this case any relevant information will remain anonymous and will only be used as statistical data.

Dynamic data will be limited to storing on pre-defined categories, (i.e. stored events will not include political events, or religious places visited), not all types of each category.

The following data will not be stored:

- Name, address, telephone, fax, e-mail, photo, etc. of the user (any direct or indirect link to user ID).
- User location (retrieved every time dynamically by the system, but not stored).
- Any other preferences/actions by the users, except the ones motioned explicitly above.
- To whom they communicates, their frequent contacts, etc.

4.1 During pilots

During the SKILLFUL Pilot tests:

1. Transport professionals participating in the trials will give their names, address, and contact phone, together with age, gender, nationality to a single person in each pilot site, to be stored in a protected local database (to contact them and arrange for the tests). The contact person will issue a single Participant ID for each of them. This person will not participate in the evaluation and will not know how each user behaved.
2. The names, address and contact phone will be kept in the database only for the duration of each trial (short term trials-up to 1 week, long term trials- up to 1 month). Such data will not be communicated to any other partner or even person in each pilot site. Once the test ends, they will be deleted.
3. Each month the anonymised data will be re-sorted randomly, to mix participants order.
4. Since personal data will be deleted, no follow-up studies with the same people will be feasible.

The partners of the consortium agree and declare that personal data will be used in strictly confidential terms and will be published as statistics (anonymously).

5 Conclusion and next steps

The next step will be to define the data management repository's technical, content, and quality specifications for storing and communicating the datasets to be created during the pilots and apply any ethical and other restrictions that data owners have decided to do so.

Another two versions of this deliverables will be issued during the lifetime of the project (M24 and M36) with detailed descriptions of the data types that will be retrieved within the SKILFUL Pilots (through the template will be created and circulated to pilot sites to be completed with information and descriptions about the data that will be collected during their pilots) and details about any relevant technical specifications and protocols (M24), while the final version (M36) will also contain the evaluation of the whole piloting procedure for future configuration and use.

This deliverable will act as a reference document. Any ethical considerations, especially about data protection, privacy and security will be foremostly discussed with the partner acting as the data owner and with the members of the SKILLFUL Ethics Board.

The overall plan related to data management to be followed during the lifecycle of the project will be further refined with consideration for both pilot planning and respective indicators (D5.1), while it is also in accordance to the SKILLFUL Ethics related policy (D7.2).

References

1. EUROPEAN COMMISSION Directorate-General for Research & Innovation (2016). Guidelines on FAIR Data Management in Horizon 2020. Retrieved from: http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf
2. Data definition. Retrieved 18th September 2017 from: https://en.wikipedia.org/wiki/Data
3. Definitions of objective and subjective data. Retrieved 18th September 2017 from: http://www.userfocus.co.uk/articles/datathink.html
4. National Information Standards Organization; Rebecca Guenther; Jaqueline Radebaugh (2004). Understanding Metadata. Bethesda, MD: NISO Press. ISBN 1-880124-62-9.
5. 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)*
6. Gemou, M., Vitzilaiou, P., Loukea, M. (2015). Data Management Plan (Deliverable 7.4) – IN LIFE EU Research Project (http://www.inlife-project.eu/).