ENVISAGE

ENhance **VI**rtual learning **S**paces using **A**pplied **G**aming in **E**ducation

H2020-ICT-24-2016

D7.3 - Data Management Plan

Dissemination level:	Public (PU)
Contractual date of delivery:	Month 6, 8/6/2017
Actual date of delivery:	Month 6, 31/3/2017
Workpackage:	WP7– Management
Task:	T7.2 – Technical and quality management
Type:	ORDP: Open Research Data Pilot
Approval Status:	Final
Version:	1.0
Number of pages:	24
Filename:	D7.3_Data management plan_versionFinal.pdf

Abstract: This deliverable is the Data Management Plan document, which describes the various types of data in ENVISAGE, the procedures followed to collect them and the measures that will be taken in order to ensure that no confidential information will be leaked. Also, the storage, archiving and preservation plan of the data is also sketched, along with our plan to comply with the Open Data Initiative.

Co-funded by the European Union

Copyright

History

Version	Date	Reason	Revised by
V0.1 (alpha)	22/05/2017	Table of Contents – Checked for completeness by the consortium	Giannis Chantas
V0.2 (beta)	2/06/2017	Ready for internal review version	Fabian Hadiji, Marc Müller
V1.0 (final)	15/06/2017	Ready for submission to EC	Giannis Chantas

Author list

Organization	Name	Contact Information
CERTH	Giannis Chantas	gchantas@iti.gr
CERTH	Spiros Nikolopoulos	nikolopo@iti.gr
CERTH	Ioannis Kompatsiaris	ikom@iti.gr
GIO	Marc Müller	marc@goedle.io

Executive Summary

This document is the Data Management Plan that describes the types of data that will be collected in the ENVISAGE activities and how they will be managed and shared throughout these activities. More specifically, the purpose of this document is to: a) specify the data that will be collected during the activities of ENVISAGE, b) investigate the best practices and guidelines for sharing the project outcomes and facilitating open access to research data, while ensuring compliance with the established ethical and privacy rules, and c) define how the data collected in the project will be made available to third parties in contexts such as scientific scrutiny, peer review and use for research purposes.

Abbreviations and Acronyms

2D/3D	**Two/Three dimensional**
API	**Application Program Interface**
DB	**Data Base**
DMP	**Data Management Plan**
DoA	**Description of Actions**

Table of Contents

List of Tables

1 Introduction

During the lifetime of ENVISAGE, data of different nature will be generated and collected. These data are user-related, which means that they contain sensitive information, and thus a clear plan is required on how they are to be managed, i.e., stored, accessed, protected against unauthorized or improper use, etc. Hence, the main goals of ENVISAGE's Data Management Plan (DMP) are to:

1. Outline the types of data that have already been generated at the current stage of the project or foreseen to be generated at later stages, including the context and procedures of this generation, as well as the degree of privacy and confidentiality of the data.
2. Outline the protocols that will be followed to assess the generated/collected data with respect to their sensitivity.
3. Outline the data acquisition plan for the whole duration of the project.
4. Outline the measures and tools that are foreseen for the adequate management of the data from the ethical and security points of view.
5. Outline the guidelines that will be followed in the project with respect to the Open Data initiatives.

In accordance with the guideline on data management in Horizon 2020 [1], the following aspects are discussed in the DMP: a) dataset reference and name, b) dataset description, c) standards and metadata, d) data sharing and archiving and preservation (including storage and backup).

The remainder of the deliverable is structured as follows. Section 2 describes the date and the settings in which data is and will be generated. In Section 3, the data collection processes are outlined. In Section 4, the data are handled according to the different categories regarding confidentiality. In this respect, there are data which are confidential and need special protection, data which are not confidential and can be shared, as well as data which depend on the informed consent of the participant. Section 5 outlines the ethical and confidentiality considerations with respect to the ENVISAGE data and Section 6 summarizes the deliverable.

2 ENVISAGE context and data description

The main goal of ENVISAGE is to provide the necessary educational technological tools to educators so they can design, create, curate and improvement of virtual labs more efficiently and faster by using of data analytics, i.e., sophisticated, machine learning algorithms for the analysis of data gathered by learners using these labs. Data of this type are divided in two categories, the shallow (aggregated data coming from simple analysis) and deep (data produced by user behaviour modelling) analytics. Also, other types of data are also planned to be gathered, such as the user related information (e.g., name, address, school), data coming from the evaluation of various components that belong in the ENVISAGE context, i.e., the educational tools provided ENVISAGE and the students' skills and performance. Thus, the objectives of data gathering in ENVISAGE is to: a) develop the user profiling and modelling, b) develop and assess the shallow and deep data analysis, and c) evaluate the developed virtual labs during pilot actions and d) assess the efficiency of the overall methodology of transferring game data analytics to learning. To achieve the above, ENVISAGE should:

- Find the curator/teacher needs and define the requirements for virtual labs enhanced with a data analytics facility;
- Track, record, store and provide access to the data showing the user activity when taking lessons using the virtual labs.
- Analyse the tracked data in order to understand and model the user behaviour by categorizing the learners into different profiles.
- Based on user profiling and behaviour modelling, as mentioned above, predict the user future behaviour when using virtual lab.
- Assess the degree of success of this transfer by pilot actions where teachers will use the virtual labs and in this way allow the testing of the overall ENVISAGE methodology.

All technologies to be developed in order to achieve the above goals rely heavily on data collected by tracking, as well as on their efficient management (i.e., storing and accessing). We can distinguish between three general types of data. First, data will be collected by the tracking of the learners/students when using virtual labs, while they are challenged to achieve specific goals in a game-like manner. Second, there will be processed data, i.e., data produced by "shallow" analysis, i.e., application of simple statistical metrics, such as the average and standard deviation, and data produced by "deep" analysis, i.e., data produced by further analysing the former type of data, e.g., future user behaviour prediction. This also means data of this type will include sensitive information, such as the user identity, school, place, ethnicity, etc. Finally, during the pilot activities, data will be collected from the learners in order to measure the impact of the ENVISAGE data analytics technology to the overall process of designing and, then, enhancing using this methodology. Thus, data gathering will be used in a continuous procedure, followed to design and improve the labs in in terms of their effect on achieving the educational objectives set by the educators. Building a product out of the ENVISAGE project, requires a market research. This will be achieved through interviews and surveys with teachers, schools and education software provider. In the following, we describe the three general types of data, as well as their subtypes.

2.1 User data

The users are typically students undertaking the tasks assigned to them by teachers using the virtual labs. Students can be of different ages, classes, schools, city and country. The user data that we want to collect and store are coming from the user related information mentioned previously that varies per student. Of course, this type of data is sensitive, thus, as a first measure to ensure safety, we do not keep the name and surname of the student and other sensitive information that may allow someone to deduce the user identity. Moreover, we plan to acquire, store and access the data to them through a safe communication channel, i.e., encryption based on secret keys known only by the project participants.

2.2 Telemetry data (events generated by the tracking points)

Users of virtual labs are tracked in order to collect data produced by their behaviour, action, choices, achievements, etc. Data of this type is collected though telemetry, i.e., remote tracking of the user when undertaking the virtual lab activities. This type of data will be stored in servers dedicated to this end. For now, access to the tracking data is possible through the transfer of a file, where all user activities tracked for a specific day are stored. Later, a communication channel between the two will be established under a predefined protocol so as to provide open and convenient access.

2.3 Evaluation data

Different entities participating in the project activities will be evaluated. These entities are the students (evaluated by the teachers) and the virtual labs (evaluated by the students and teachers).

2.3.1 Performance scores for students

Performance scores of students have the purpose to track the learning their progress, and, thus, they are strongly related to the learning objectives of the pilot trials. Students are expected, during and after the completion of their activities, to achieve a significant increase of the skills and a higher level of expertise in specific scientific fields. Thus, teachers are to be assign performance scores to the students in order to measure the impact of the ENVISAGE tools applied during the educational pilots and testing within the project activities. It is evident that this type of data is sensitive and must be kept confidential.

2.3.2 Student opinion's from questionnaires or interviews

Not only will students be evaluated, but also students will evaluate the ENVISAGE outcomes they use, i.e., the virtual labs, in terms of their user-friendliness and experience, game-like characteristics, educational efficiency and their capability to immerse a user to the gamified educational activities. This is useful information that we want to take from students, since it can validate our claim about the added value of the virtual labs, when used in place of conventional ones.

2.3.3 Teacher opinion's from questionnaires or interviews

Teacher's opinion is also very useful for the evaluation of the ENVISAGE tools. Thus, teachers will evaluate the ENVISAGE outcomes, i.e., virtual labs and the data analytics provided by them, in terms of their educational added value, how they boost and help learning and whether the lab activities are realistic and scientifically sound. This is information of utmost importance, since it can measure the added value of the ENVISAGE outcomes and general concept and context.

2.3.4 Market research data from questionnaires or interviews

To make a product out of the ENVISAGE project, it is necessary to know our target groups and how these target groups are currently using learning analytics. This implies to gain knowledge over the whole learning analytics market, the way how the mass of teachers is currently working with software, which software is available, which kind of learning analytics is already used and last but not least possible connecting factors to education software provider. The education system is different for every nation in the EU. It could also be different within a single nation. In Germany, for example, the curriculum is different for every state. This includes the usage of education software and learning analytics. Therefore, we need a more or less general approach of adding learning analytics to software, so that it can be used independently in any curriculum. This requires finding out which software teachers are currently using, how the distribution/ sales process runs, and who finally pays for a learning analytics system. To fulfil the prerequisites information, we decided to get valid data in addressing surveys' directly to the user groups who are participating in such a system.

2.4 Data Analytics

2.4.1 Shallow analytics: aggregation and visualization

Stored tracked data are going to be used aggregated and, then, act as feed to visualization tools. In order to facilitate accurate and meaningful visualization, data will be aggregated using metrics that provide a single value by taking as input large numbers of data, e.g., average, standard deviation, or parametric representation of the data. This type of data is not as sensitive as the user data, since one cannot deduce personal information from them. However, the data sensitivity still persists even in a lesser degree, since the aggregation can be performed in a class or school level. Thus, special care must be taken in order to avoid leak of information regarding the performance of a specific class or school. This type of information will be stored and kept confidential in the circles of the consortium.

2.4.2 Deep analytics: user behaviour modelling

User behavior is modelled through the analysis of past user behavior in conjunction with personal traits, such as age and school class. User behavior modelling has the purpose to predict the future behavior of a student. Thus, data of this type is in essence the model itself and its parameter values. As it is evident, we consider user models a sensitive data in order to protect students by information leak that might expose their behavior and profile.

2.5 Virtual lab data

2.5.1 3D templates and assets

Virtual labs will be produced by the ENVISAGE authoring tool. Technically, apart from software, a virtual lab consists of assets, i.e., digital creations such as 3D graphics and images used as texture. Moreover, the authoring tool will provide 3D templates of virtual labs with the purpose to be used as the base on which one can design and create many virtual labs by specifying its details (e.g., graphics, functionalities and scenes).

2.5.2 Learning-game design documents

There will be documents that will describe in detail the functionalities of the virtual labs. These documents will act as “manuals” that can help the teachers understand better the general concept of the game, the functioning of its components and how the user can interact with them. Moreover, special care will be taken so as to describe the tracked user activities and what the data analytics functionalities provide.

3 Data collection

3.1 Collection of user data

User data can be collected when using the labs. More specifically, each user will be asked to enter for user id that will be associated with his/her personal information. The association between the id and the personal information will be made by the responsible teacher who has the authority and responsibility to supervise the student when the latter uses the labs. The personal information will consist (at least) of the following fields:

- Name and Surname
- School
- Class
- Age
- Country
- Male/Female

These fields will be entered by the teacher who relates them with a unique user id, or by the student corresponding to this user id, where there will be always a teacher who validates the authenticity of these fields. This also means that each virtual lab will prompt the user before using the lab to enter the user id, so as the tracking data are associated with a specific user. Thus, collection of user data is a procedure done once for all for every student.

3.2 Telemetry data (events generated by the tracking points)

Data collection will occur by tracking students that use the labs in schools or similar educational contexts (e.g., summer schools, workshops). There, teachers will coordinate the process of students undertaking exercises with specific virtual labs and accomplish the goals, related to a specific scientific field to which the lab is dedicated, set to them by the teachers. Also, in cases the presence of a teacher is not possible, it will possible for students to undertake the same exercises through remote access (i.e., Internet), since the virtual labs will be web based. This means personal information will be stored along with the tracked data, which will be later used for user analysis, i.e., user profiling and behaviour modelling. Special care will be taken so as to the student to provide only necessary information which must also be communicated with safe means. The datasets will be collected during these activities and stored in a predefined format and will be able to be accessed through a safe communication channel. Table 1 shows examples of the information hold in the data coming from user tracking.

Also, at a later stage, the tracked data by the users participating in the pilot trials will be wrapped up in a research dataset. The value of this dataset will reside on the protocol followed during the pilot trials, as well as the variety of the target groups participating in these trials (e.g., school class). This dataset will become available through our website. Lastly, it is worth to mention that personal information and details that can lead to the identification of the student will not be published.

Table 1: Tracking data information given as example

Data Source	Data Type	Example
Tracking: Detection of specific events upon their occurrence	Event ID, type and information	User clicked a button, or changed a parameter that controls specific functionalities of the virtual lab
user_id (string)	A unique user identifier which is unique per virtual lab	
ts (integer):	The Unix timestamp when the event was triggered.	25100
event (string):	The name of the tracked event.	view.instructions, view.configuration, increase.parameter, decrease.parameter
event_id (string)	The event identifier which specifies the event, e.g., a lesson identifier	<identifier_instruction_tab>, <identifier_configuration>, <identifier_speed>
event_value (string)	The event value specifies the value of an event, e.g., the value of a controller that is adjusted.	<value_speed_adjustment>, <value_increase_trigger>
group_id	An unique identifier for a group, class, category, or company. A user can have more than one "group_id".	A1, C2, etc.

3.3 Evaluation data

During the pilot trials the following data will be collected:

3.3.1 Performance scores for students

Performance scores will be assigned to students by teachers, in order to assess their progress during while using specific virtual lab(s) in the context of the pilot trials. These data will be collected by asking teachers to deliver the scores in a predefined format. Then all the data will be aggregated to a common archive designed specifically to this end.

3.3.2 Student opinion's from questionnaires or interviews

Questionnaires will be given to students, in order to evaluate specific virtual lab(s) that they used in the context of the pilot trials. Then all the data will be aggregated to a common archive designed specifically to this end.

3.3.3 Teacher opinion's from questionnaires or interviews

Questionnaires will be given to students, in order to evaluate specific virtual lab(s) that they used in the context of the pilot trials. Then all the data will be aggregated to a common archive designed specifically to this end.

3.3.4 Market research from questionnaires or interviews

There will be two kinds of information sources, on the one hand, direct outreaches via e-mail, phone and events with teachers and software provider to create the surveys' with the right questions. On the other hand results of the questionnaires that will be sent as surveys to teachers.

The outreach part also helps us to check if a potential target group is willing to use an ENVISAGE like product. The results of the first outreaches will be used to create a questionnaire, which will collect data about:

- Current usage of learning software
- Frequency of using learning software
- Subjects where learning software is used
- Kind of schools where learning software is used

The results of the questionnaire will be stored in a Google Spreadsheet and will be available as *.csv file, to provide access to the data set. This will allow the creation of histograms and distributions, for an overview of the market and a usage status quo for education software.

The first version of the survey is accessible under [7]. This survey has been sent to about 200 teachers in Cologne, Germany. We have a response rate of 9 teachers that completed the survey. This follows to a future outreach of 5 to 10 times more teacher, to get a first evidence of the usage of education software. In the second iteration, the whole process will be repeated with the focus on learning analytics in education software. We will start outreaching to teachers again, create a questionnaire and then start talking to education software provider if they want to make use of learning analytics. This questionnaire will collect data about:

- Usage of statistics to improve learning
- Tools that help to create learning analytics
- For which king of subject learning analytics is interesting
- Are they planning to use learning analytics

Both questionnaire data will be publicly available, while teacher call-protocols and interviews with software providers are only summarized, because of the lag of comparability.

3.4 Produced data by data analytics

Taking as input the tracked data, data analytics algorithms will produce data that will be abstractions of the former, such as average values, meaningful visualizations of them, user models that can predict their behaviour, etc. These are divided in two subtypes and their collection is described next.

3.4.1 Aggregated data and visualization

Aggregated data will be generated using the data coming from tracking. Aggregation will be performed by algorithms run in servers, dedicated to this end. After the aggregation, the results will be stored in a server and they will accessible via safe means. Particularly, the server will be hosted in Amazon Web Services that provide technologies for high traffic, storage, security and maintenance at low cost. The raw (tracking) is stored in S3, while the aggregated, and the augmented data will be in a database based on DynamoDB technology. Access is needed by and will be granted to the visualization tools that a teacher can use to monitor the activities undertaken in the labs in a meaningful and abstract manner. Thus, collection of this type of data is a continuous and automatic procedure.

3.4.2 User behaviour modelling

User behaviour models data will be generated using the data coming from tracking data and user information. The data of the models are in essence the software implementing the models and the parameter values of them. User models will be created by machine learning algorithms run in servers dedicated to this end. After their creation, the results will be stored in a specific database and they will accessible via safe means. Access is needed by the deep analytics tools that a teacher can use to predict the future behavior of students and take actions according to the prediction. More specifically, an API server will be used for sending the data to the Deep Analytics Server for extracting meaningful information for the educators or to the game authoring tool for the visualization of game analytics. Also our implementation allows metric based data to be removed from the server securely.

3.5 Virtual lab data

3.5.1 2D/3D assets

When creating a new virtual lab, the user of the authoring tool imports 2D images and 3D graphics (i.e., the assets), which comprise the visual part of the lab that the user/student interacts with. Thus, the designer must create such assets, which can be done using external graphics editing tools, and import them to the lab via the authoring tool. The user is free to use any graphics editing and design tools that finds convenient to this end. These assets, in order to be used not only in the lab for which they are designed but also for future labs, must be collected and stored appropriately so as to be easy to manage them and provide access to them in order to be used by other lab creators and designers.

3.5.2 Virtual lab templates

The virtual lab templates, which are created to speed up the lab creation process, are designed and created using the Unity3D editors on which the authoring tool is based. More specifically, one must use both a template and a set of assets in order to create a new lab. The templates will be stored with the assets and we will provide access to them to the users of the authoring tool.

3.5.3 Learning and game design documents

The game design and learning documents will be collected by the authoring tool users that create a game/lab while the learning documents will be collected by the teachers who employ the lab to their educational and teaching activities and use these documents to give tasks to the students when using the labs. All these will be stored in the same place along with the other virtual lab data and be accessed by the learners/students.

4 Data management and sharing

Almost all data collected in ENVISAGE, which we have to manage, is generated within the activities of the project. More specifically, the data will be generated by the tracking and analytics algorithms, as well as, during the various evaluation procedures foreseen to take place in the project activities. A subset of the aforementioned data is user-related and thus requires a clear plan on how they are to be managed (i.e., stored, accessed, protected against unauthorized or improper use, etc.).

4.1 Guidelines for generated data

We plan to collect data through the pilot trials that implement and test the ENVISAGE tools in the context of various case studies. More specifically, during these pilot trials, students will be asked to use the virtual labs and undertake specific tasks assigned to them by teachers. Also, teachers will be asked to design virtual labs using our authoring tool. The pilot trials of ENVISAGE are planned to take place in Athens, Greece, and specifically at the premises of EA. Thus, data of all types described in Section 2 and collected via the procedure mentioned in Section 3. In handling these data, we will make sure to comply with national and EU legislation, as well as follow the best practice for ethics regarding user privacy, confidentiality and consent. In Section 5 of the DoA [4], detailed guidelines are provided for the activities of: a) data collection, b) storage and transmission, c) retention, d) treating sensitive data, and e) term of usage. Thus, in managing ENVISAGE's data we will make sure to comply with all guidelines specified in the DoA [4]. This means that the design and development of novel interfaces (i.e., virtual labs) as well as the logging of the user's activity and learning progress, require careful deliberation of the ethical implications that may arise.

4.2 Categories of data based on their confidentiality level

We can categorize the collected data based on their confidentiality level in three categories; i) open data, which can be shared openly, ii) protected data, which can be shared but the participants have to provide their consent, and iii) confidential data, which cannot be shared outside the project.

Table 2 categorizes the data generated in ENVISAGE to one of the aforementioned categories. We can see that apart from the sensitive data that has to do with the participation learners/students, all other types of data can be made public, provided that the necessary consents are obtained from the participants.

Table 2: ENVISAGE data categorized based on their confidentiality level.

Topic	Objective	Data Type	Source	Category
User Data	Virtual labs user personalization and profiling	Text	Students personal information	Confidential
Telemetry data	Track student	Structured	Tracking	Protected

		activities in order to apply data analytics	(database) and text	module of virtual labs	
Evaluation data	Student evaluation	Evaluate the students within ENVISAGE pilot trials	Numbers (scores)	Scores assigned by teachers to students	Confidential
	ENVISAGE outcomes evaluation	(protected) Evaluation of educational tools by students and teachers. (open) Evaluation of the learning analytics and education software market.	Numbers (scores) and text	Answers to questionnaires	Open/Protected
Data Analytics	Shallow	Data aggregation & visualization	Visual plots	Shallow data analytics algorithms	Protected
	Deep	User behaviour modelling	Structured text	Deep data analytics algorithms	Protected
Virtual lab data		Provide the functioning components to virtual labs	2D/3D graphics	ENVISAGE authoring tool and graphics editors	Protected

4.3 Data cataloguing

Below we provide details on how we intent to document a dataset before making it public, so as to facilitate convenient identification and access.

4.3.1 Dataset reference and name

A unique identifier will be given to each dataset. At this stage of the project, individual data sets have not been formally identified. Once this is done we will be able to develop

knowledge of the target datasets required and so can commence populating the data catalogue.

4.3.2 Dataset description

For each dataset in the catalogue there will be a description outlining: the nature and scale of the data, to whom it could be useful, and whether it underpins a scientific publication. Information on the existence (or otherwise) of similar data and the possibilities for integration and reuse would also be included.

4.3.3 Metadata standards for tracking data

For the tracking data we plan to use a simple text format. The aim of the format is to combine and integrate all tracking data into a repository that provides easy access. In future, we plan to use metadata standards that best fit to our data management plan and activities.

4.4 Data sharing

Depending on the confidentiality level of the data different sharing approaches will be employed.

4.4.1 Private sharing

This approach consists in setting up FTP servers by the partners acting as data controllers (see DoA – Section 5 [4]), so as to share within the consortium all data generated in the project. Tracking data will be stored in a server at the premises of GIO, as explained in Section 2.2 . Shallow and deep analytics data will be stored together in a second server. Lastly, virtual lab evaluation data and will be stored in a server lying and the facilities of CERTH. The FTP servers will be used to host all different types of data (i.e. open, protected and confidential) and will apply strict accessibility rules, such as anonymization, password protection, and transmission in a 128-bit encrypted form through a secure communication channel (SSL). The task of archiving the data on the servers will be carried out by the partner that collected the data (i.e., the data controller), who is also responsible for obtaining the informed consent of the participants from his data generation procedure, if needed. The Private sharing approach will be put in place to facilitate the smooth communication of the generated datasets, while ensuring that only the members of the consortium can access them.

4.4.2 Controlled sharing

This approach consists in sharing the dataset through the project web-site and allowing externals to download them, after requesting to provide their personal and affiliation details and to agree with the terms of a data usage agreement. In this case, we will include the types of data belonging to open and protected categories, whereas for the latter we will only share the part of data where informed consents have been obtained. This means that one can have free access to open data and access after consent by the data owners in the case of the protected data. More specifically, based on the cataloguing information described in Section 4.3 we will generate the appropriate web-forms (under ENVISAGE's web-site) to describe and link the datasets. All technical details for accessing and processing the data will

be also included as part of these web-forms. Subsequently, we will seek opportunities to list our datasets in sites that serve as aggregators of data analytics datasets, such as [2] **Error! Reference source not found.**. This will boost the visibility of the generated datasets and ensure their widespread use.

4.5 Archiving and preservation

As already mentioned the datasets (in their raw form) will reside on the institutions of the project participants acting as data controllers. These institutions typically offer services like redundancy, back-up and migration that are consider sufficient for the preservation of the generated data. In addition, these institutions usually maintain a large digital library that is used to index and archive the content generated by their activities. Finally, it is important to mention that the services offered by these institutions are free of charge and may sustain for many year after the end of the project.

4.6 Open Data initiative

Regarding the Open Data initiative, ENVISAGE aims to take part of it by:

a) Specifying the data that will be collected through the learning analytics module in terms of the learners' progress within the pilots;
b) Investigating the best practices and guidelines for working with Open Data and take into account the ones released by the: i) Open Data Foundation and ii) Open Knowledge Foundation [6];
c) Defining how the data collected in the project will be made available to third parties in a scientific context and use for research purposes.

5 Ethical and confidentiality considerations

ENVISAGE will give specific attention to any ethical issues that will arise and will address them in a professional way following very closely established EU regulations and corresponding national laws about user privacy, confidentiality and consent. During the participation of users in the pilots of the project, their activities will be logged by the learning analytics framework. In this respect and based on the above, the objective of ethics consideration in ENVISAGE is twofold: a), it intends to ensure that the identification of the user takes place without raising any ethical issues and without breeching any confidentiality, and b) second, it will evaluate the ethical impact of the project outcomes on the end users.

Thus, the adopted ethical practices, which are described in more detail in the DoA (Section 5), are related to the ethical assessment of (i) the impact of the data tracking infrastructure on the end users, (ii) the impact of the virtual lab interfaces on the end users and (iii) the design process itself, with the intention of validating the approaches taken to address the education-related issues that the project is dealing with.

6 Conclusions

In this deliverable, we have presented the data that will be generated in the context of ENVISAGE including data from tracking, shallow and deep analytics and evaluation. Our data management plan is built upon analysing the generated data with respect to their confidentiality level and employing a different sharing approach depending on this level. More specifically, three confidentiality levels are envisaged (i.e., open, protected and confidential) and two sharing practices are described (i.e., private and controlled sharing). In this respect, confidential data and data rated as not shareable according to ethical considerations will not be shared. Finally, the DMP strictly builds on ensuring the necessary informed consents, as well as respecting the sphere of privacy of each participant. This document will be further developed, as we gain more information about the specific user requirements for the authoring tool and the usage scenarios (i.e., virtual labs) implemented in ENVISAGE.

7 References

[1] 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

[2] https://zenodo.org/

[3] http://www.scientix.eu/

[4] Description of Actions - Readable form (requires authentication): http://mklab.iti.gr/envisage/lib/exe/fetch.php?media=partb_envisage_sec1-3.pdf

[5] http://www.ich.org/products/guidelines/efficacy/efficacy-single/article/good-clinical-practice.html

[6] https://okfn.org

[7] https://goo.gl/forms/9yxyYki16S79oLh53