



# **Intelligent Verification/Validation for XR Based Systems**

**Research and Innovation Action**

Grant agreement no.: 856716

## **D6.2 – Data Management Plan**

**iv4XR – WP6 – D6.2**

**Version 1.3**

**March 2020**



Project Reference	EU H2020-ICT-2018-3 - 856716
Due Date	31/03/2020
Actual Date	31/03/2020
Document Author/s	Fitsum Kifetew (FBK), Marta Couto (INESC)
Version	1.3
Dissemination level	Public
Status	Final

This project has received funding from the European Union's Horizon 2020 Research and innovation programme under grant agreement No 856716



<b>Document Version Control</b>			
<b>Version</b>	<b>Date</b>	<b>Change Made (and if appropriate reason for change)</b>	<b>Initials of Commentator(s) or Author(s)</b>
1.0	20/01/2020	Initial document structure and contents	FK
1.0	6/03/2020	Draft of section 6	MC
1.0	9/03/2020	Data summary and first complete draft	FK
1.1	13/03/2020	Draft for internal review	FK
1.2	25/03/2020	Addressed comments from internal review	FK
1.2	25/03/2020	Updated section on Ethics based on internal review	MC
1.2	30/03/2020	Draft for submission	FK
1.3	31/03/2020	Final check and minor changes in the format	RP

<b>Document Quality Control</b>			
<b>Version QA</b>	<b>Date</b>	<b>Comments (and if appropriate reason for change)</b>	<b>Initials of QA Person</b>
1.1	16/03/2020	Review and minor suggestions	IS
1.1	16/03/2020	Review and minor suggestions	JC
1.1	17/03/2020	Review and minor suggestions	ML
1.1	20/03/2020	Review and comments on the Ethics section	FB
1.1	24/03/2020	Review and minor corrections	AS

<b>Document Authors and Quality Assurance Checks</b>		
<b>Author Initials</b>	<b>Name of Author</b>	<b>Institution</b>
FK	Fitsum Kifetew	FBK
MC	Marta Couto	INESC-ID

FB	Filipa Borrego	INESC-ID
AS	Angelo Susi	FBK
IS	Ian Saunter	GWE
JC	Jeremy Cooke	GWE
ML	Manuel Lopes	INESC-ID
RP	Rui Prada	INESC-ID

**TABLE OF CONTENTS**

<b>1 EXECUTIVE SUMMARY</b>	1
<b>2 DATA SUMMARY</b>	2
2.1 Data types and formats	2
2.2 Origin of the data	3
2.3 Expected data size	4
2.4 Purpose and utility	4
<b>3 FAIR DATA</b>	4
3.1 Making data findable	4
3.2 Making data openly accessible	5
3.3 Making data interoperable	5
3.4 Increase data re-use	6
<b>4 ALLOCATION OF RESOURCES</b>	6
<b>5 DATA SECURITY</b>	6
<b>6 ETHICAL ASPECTS</b>	6
<b>7 CONCLUSIONS</b>	7
<b>8 REFERENCES</b>	8
<b>APPENDIX A</b>	9

## **1 EXECUTIVE SUMMARY**

This report details the data management plan (DMP) adopted by the iv4XR project. The DMP describes what types of data are produced and/or used within the project and how they are treated, stored, processed, disseminated, and maintained, within the project as well as with the general public, during the lifespan of the project and beyond. This report describes the current state and initial plan with respect to data in the project. It will be continually updated as data management needs evolve during the project's lifetime.

The current report and its content are structured following the guideline and recommendation in the Data Management H2020 online manual [1].

## 2 DATA SUMMARY

In this section we present a summary of the various data types that are expected to be generated/collected and managed by the project during its lifespan and beyond. We also describe why the data is collected, who will use it, and to whom it could also be useful.

### 2.1 Data types and formats

The information regarding data produced and maintained in the project is collected by means of questionnaires that we have prepared for this purpose. Questions in the questionnaire are included in [Appendix A](#) of this report.

The questionnaire has three parts: Part I about whether or not the partner intends to handle personal data, Part II about the nature of each data type to be maintained by the partner, and Part III about storage and security-related aspects. Each partner in the project consortium has completed the questionnaire with respect to their data collection needs as well as the levels of dissemination they expose their data to. The data managed in the project includes both personal data as well as non-personal. Based on the responses to the questionnaires, we have determined the following characteristics of the data to be maintained by the project.

Table 1: Personal data

<b>Nature</b>	<b>Type/format</b>	<b>Origin</b>	<b>Expected size</b>
Audio/video recordings	Binary (Audio/video)	Collected from participants	< 10GB
User info (age, sex, ..)	Text	Collected from participants	< 1GB
Usage logs	Text	Collected from participants	< 1GB
Questionnaires	Text	Collected from participants	< 1GB
Consent forms	Text	Collected from participants	< 1GB

Table 2: Non-personal data

<b>Nature</b>	<b>Type/format</b>	<b>Origin</b>	<b>Expected size</b>
Feedback from system usage	Text	Collected from internal users	< 2GB
Anonymized data from experiments	Text/spreadsheet	Collected from experiment participants	< 2GB
Data from various sensors	Binary, Text(JSON,XML)	Collected from sensors in engineering projects	< 100GB
System usage logs, screenshots, reports	Binary, Text (JSON, HTML), Image (png)	Generated during the project when testing systems	TBD
Models of system under test	Binary, Text (JSON)	Generated during the project when testing systems	< 10GB
Tactical terrain	Proprietary format	Produced for the project using an internal tool	< 1GB
2D/3D terrain models	Standard 2D/3D format (TBD)	Produced from raw data using internal tools	TBD
WSDB	Proprietary format	Produced for the project using an internal tool	< 1GB
Scenarios	Proprietary format	Produced for the project using an internal tool	< 1GB
Source code	Text (Various programming languages)	Produced in the project	< 2GB

Please note that Table 1 and Table 2 only present high-level aggregate summaries of what has been collected through the questionnaires. The questionnaires contain more details about the nature of data as well as storage and privacy levels (see [Appendix A](#)). Such details are abstracted away when reporting Table 1 and Table 2 as they contain sensitive information that could not be made public (as the current report has public visibility). The responses to the questionnaires are, however, kept internally within the consortium and they serve as a baseline for the implementation of the data management plan described in this current report.

## 2.2 Origin of the data

As detailed in Table 1, personal data will, in general, be collected from participants in user studies that will be conducted in the context of experiments for evaluating research output. The participants could be both internal (testing) staff as well as external subjects participating in experiments. In both cases, appropriate consent forms will be filled out by the participants and data will be handled in compliance with laws and norms regarding the treatment of personal data.



Furthermore, aspects related to ethics will be handled according to the ethics policy adopted by the project, as described below in [Section 6](#).

On the other hand, non-personal data, as presented in Table 2, is mostly produced in the project (or prior) by applying internal tools owned by the respective partners. Some are collected from different sensors mounted on engineering infrastructures and/or eventually elaborated further using internal tools.

### 2.3 Expected data size

The actual sizes of each data type will become more accurate as the project progresses. Estimated expected sizes are indicated in Table 1 and Table 2 for each data type. In some cases it is difficult at this moment to provide an estimate, hence left as TBD in Table 2. Overall, the size of the raw data is not expected to be extraordinarily large. Considering eventual processing of the data which could result in new forms of data, the overall size is expected to be in the order of hundreds of gigabytes.

### 2.4 Purpose and utility

The main purpose for which data is collected, processed and maintained in the project is for the assessment and evaluation of the solutions developed as a result of the research conducted in the project. The data collected will be mainly used by the consortium members to evaluate and test their solutions. Furthermore, the data (possibly after being processed) could be made available to the research community and the general public at large. Such exposure of data will be handled according to the level of visibility granted by the owner as well as related copyright and intellectual property restrictions. Release of data regarding participants (personal data) will be handled according to the consent given by the persons themselves, as well as after being anonymized and, whenever applicable, in aggregate form.

## 3 FAIR DATA

iv4XR will do its best to adhere to FAIR principles (Findable, Accessible, Interoperable, and Reusable) for the data managed in the project. Research output will be shared openly, except in cases where there could be embargos or other limitations. Datasets will also be shared openly, respecting the specific privacy and copyright restrictions on them from the respective owners. Personal data will not be shared publicly and will be handled according to ethics and regulatory guidelines, as detailed in [Section 6](#).

### 3.1 Making data findable

Public data shared by the project will be published on Zenodo<sup>1</sup>, unless there are restrictions, or it is otherwise impossible to do so. In this latter case, alternative openly accessible infrastructure will be resorted to.

---

<sup>1</sup> <https://zenodo.org/>

To make data findable, they will be assigned Digital Object Identifier (DOI) and they will be appropriately versioned. Furthermore, they will be tagged with relevant generic metadata as provided by the Zenodo platform. We will evaluate if there will be a need to adopt specific metadata standards and if so which ones. Specific, project wide, naming conventions to be adopted will be defined.

Research output in the form of scientific publications will also be indexed by platforms typically used by the research community, such as Google Scholar<sup>2</sup>. The project website will also fully document and make accessible all research output produced in the project.

### 3.2 Making data openly accessible

As described above, public data will be shared on Zenodo, or equivalent platforms when this is not possible, and it will be indexed by OpenAIRE<sup>3</sup>. This guarantees easy public access and discoverability.

Of course, not all data managed in the project will be made publicly accessible. This includes mainly data from companies which is only intended for internal use within the project consortium. Whenever appropriate, according to the level of permission granted by the respective data owners, processed/aggregated results based on experiments performed using the data could be released in various scientific publications which will be openly released.

In general, for publicly released data it should always be possible to use open source tools for accessing and manipulating the data. Most of the data will be of textual nature with different formats, such as JSON (JavaScript Object Notation) or comma separated values (CSV) which could be easily manipulated using open source tools, such as, LibreOffice<sup>4</sup>. When binary data is released, specific non-open source tools may be required. In any case, appropriate guidelines will be provided on which tools could be used to access and manipulate the data.

Whenever appropriate, precise license information will be provided for each data publicly released.

Source code produced in the project will be hosted in the public project GitHub repository (<https://github.com/iv4xr-project>) together with appropriate license information.

### 3.3 Making data interoperable

Most of the data in the project which could be openly shared is textual in nature with commonly used formats (e.g., JSON, CSV), hence their usage by other researchers should be quite straightforward and open source tools could be used for the purpose.

If, during the course of the project, there arises the need to use customized data formats, appropriate sharing mechanisms will be defined by the project.

---

<sup>2</sup> <https://scholar.google.it/>

<sup>3</sup> <https://www.openaire.eu/>

<sup>4</sup> <https://www.libreoffice.org/>

### 3.4 Increase data re-use

iv4XR will adopt, to the limit it is possible, the most permissive licence. However, depending on the venues where some of the research output will be published, there could be some limitations on making publicly available the research output. In such cases, the public dissemination will be done in accordance with the embargo restrictions involved.

iv4XR plans to make all publicly shared artefacts permanently available also after the end of the project, unless there are specific limitations that otherwise forbid it. For this, independent and well-established repositories will be used to host the artefacts (e.g., Zenodo, GitHub).

## 4 ALLOCATION OF RESOURCES

iv4XR adopts freely available infrastructures for storage and dissemination of artefacts. Internally, Google Drive (freely available for research and academic institutions) is used for storage of artefacts within the project. For publicly shared data as well, free services (e.g. Zenodo, GitHub) will be used. Furthermore, the project website (<https://iv4xr-project.eu/>) contains information released by the project as well as indexing publicly shared artefacts (e.g., publications). The cost of the website hosting for the lifetime of the project is covered by the project. After the end of the project, the domain name will be made available for a period of time yet to be defined, while the content of the website will be transferred to an institutional server hosted at one of the project partners. Details will be defined as the project progresses.

Issues related to the data management in the project are handled collectively by the team composing Work Package 6 (WP6) in which all partners are represented.

## 5 DATA SECURITY

For data stored and managed internally within the project, security is ensured by the respective owners of the data as well as by the project itself. Most of the data is stored by the respective owner institution in private servers with regular backups. Project wide data is stored in the project shared Google Drive with access granted only to consortium members.

For publicly shared data, the repositories used (Zenodo, GitHub) ensure proper security and backup mechanisms. Further, all publicly shared data will also be internally stored in the project shared Google Drive, hence creating redundancy in case of data loss or corruption.

## 6 ETHICAL ASPECTS

Iv4XR constituted an ethics advisory board (EAB) and a protocol to ensure compliance with ethical standards. The EAB collects information about the studies taking place and ensures, not only that they follow Iv4XR protocol, but also that the competent Ethics committees evaluate and approve all studies involving humans. The project's *D7.1 – Ethics* presents collected information regarding all data protection officers from the different institutions, as well as the ethics committees available. All studies involving humans need to be approved by an ethics committee

ensuring compliance with European guidelines but also national laws applicable. The same requirement also establishes templates for informed consent and debriefing of human participants.

iv4XR will process personal data. Ethical concerns may arise from processing direct identifiers, from human participants in experiments. We will not share any data containing direct identifiers, and only anonymized data will be shared. In compliance with the data minimization principle, we will only process direct identifiers when it is not possible to obtain the information without them. An exception to this relates to sharing parts of video/audio recordings of the participants, to illustrate the experiments in conferences and scientific articles. We will ask for specific consent for doing this, however, we will delete the majority of videos/audios from participants after processing, in compliance with European guidelines. The datasets that will be shared publicly will be fully anonymized, thus not compromising the confidentiality of the participants.

All processing of personal data will be in full compliance with European guidelines and national laws applicable. For specific details on the provisions of the GDPR (<https://gdpr-info.eu>), compliance with ethical standards, consent forms, and debriefing templates, please see deliverable D7.1 – Ethics [2].

## **7 CONCLUSIONS**

The current report documents the data management plan adopted within the iv4XR project. The report outlines the initial plan of the project with respect to the data needs foreseeable by the consortium members at the current time. Eventual changes with respect to data needs will be incorporated into the report in subsequent versions as the project progresses.

## 8 REFERENCES

[1] Data Management H2020 Online Manual

[https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management\\_en.htm](https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm)

[2] Marta Couto, Pedro Fernandes, Rui Prada, D7.1 - Ethics (Confidential), iv4XR Project Deliverable, v1.7, November 2019

## APPENDIX A

### Internal questionnaire for gathering information about data managed by project partners

Part I Questions on personal data											Answer
1 Do you plan to collect personal data? [Yes/no]											
1.1 If yes, does it include also biometric data? please describe											
If yes, are there any ethical/legal restrictions on data sharing?											
1.2 please describe											
If yes, indicate the nature of the personal data											
1.3 [Text, Audio/video, Usage log, Other] please describe											
Part II											
For each dataset you plan to manage (collect, own) in the project, insert a line below (for both personal and non-personal data)											
2 You can insert additional rows if those below are not enough.											
Dataset (name)	Description [give a description of the dataset]	Format of dataset [PDF, Docx, Txt, Mpeg AV, Other] please describe	What is the origin of the data? [Collected from participants, Produced using tools, Re-used from previous projects/internal, Other] please describe	Who is the owner of the data? [The user to whom the data belong, The organization (you), Other] please describe	Are there any IPR (Intellectual Property Rights) involved? [Yes, No] please describe if yes	What is the expected size of the data you plan to collect? [Less than XX GB, Between XX & YY GB, Between YY & ZZ GB, Above ZZ GB, Not able to say] these options are templates, you can use a different measure suitable for your dataset	To whom will the data be useful? [Free text] please describe	For what purpose do you collect the data? [Research validation, Product testing, Product development, Product marketing/advertising, Other] please describe	What is the duration of the data? [Only during the lifespan of the project, During and beyond the project, Until date XXXX] please describe	What is the visibility level for the data? [Confidential, Public (anonymized), Public (full), Licensed, specify type of license, Other]	
III											
3 Data storage and backup. For some of the questions (e.g., storage), you may need to specify answers for each data type you indicated in the "Part II" above.											
Dataset name (if applicable) from those listed in Part II	What is your mechanism of data storage? [Dedicated /internal storage servers, Google drive, Other external]	Do you perform backups? If so please indicate the nature of the backup and its frequency.	Are there any security measures in place to protect sensitive data? If so what are they? please describe	Do you adopt any standards related to data management, such as metadata standards, naming standards, locator standards, etc? please describe	Is there a person/body responsible for data management in your organization? [Yes/no] please indicate if answer varies according to dataset	For public data (documents, source code, ...) which infrastructure do you currently use? [Github, Bitbucket, Zenodo, Google drive, Other] please describe	For internal storage of data at project level, do you have any issue storing your data in the project's shared Google Drive? [No, I prefer to keep data in my internal]				
3.1											